ABSTRACT

Title of Document:	SPIN: LEXICAL SEMANTICS, TRANSITIVITY, AND THE IDENTIFICATION OF IMPLICIT SENTIMENT
	Stephan Charles Greene Doctor of Philosophy, 2007
Directed By:	Professor Philip Resnik, Department of Linguistics and Institute for Advanced Computer Studies

Current interest in automatic *sentiment analysis* is motivated by a variety of information requirements. The vast majority of work in sentiment analysis has been specifically targeted at detecting subjective statements and mining opinions. This dissertation focuses on a different but related problem that to date has received relatively little attention in NLP research: detecting *implicit sentiment*, or spin, in text. This text classification task is distinguished from other sentiment analysis work in that there is no assumption that the documents to be classified with respect to sentiment are necessarily overt expressions of opinion. They rather are documents that might reveal a *perspective*. This dissertation describes a novel approach to the identification of implicit sentiment, motivated by ideas drawn from the literature on lexical semantics and argument structure, supported and refined through psycholinguistic experimentation. A relationship predictive of sentiment is established for components of meaning that are thought to be drivers of verbal argument selection and linking and to be arbiters of what is foregrounded or backgrounded in discourse. In computational experiments employing targeted lexical selection for verbs and nouns, a set of features reflective of these components of meaning is extracted for the terms. As observable proxies for the underlying semantic components, these features are exploited using machine learning methods for text classification with respect to perspective. After initial experimentation with manually selected lexical resources, the method is generalized to require no manual selection or hand tuning of any kind. The robustness of this linguistically motivated method is demonstrated by successfully applying it to three distinct text domains under a number of different experimental conditions, obtaining the best classification accuracies yet reported for several sentiment classification tasks. A novel graph-based classifier combination method is introduced which further improves classification accuracy by integrating statistical classifiers with models of inter-document relationships.

SPIN: LEXICAL SEMANTICS, TRANSITIVITY, AND THE IDENTIFICATION OF IMPLICIT SENTIMENT

by

Stephan Charles Greene

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, College Park, in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
2007

Advisory Committee:
Professor Philip Resnik, Chair
Dr. Donald Hindle
Professor Jeffrey Lidz
Professor V.S. Subrahmanian
Professor Amy Weinberg

Dedication

This dissertation is dedicated to

Sabrina Parker Greene and Owen Gates Greene

Hey, guys—I'm done!

Acknowledgements

The completion of this dissertation, at times a remote and preposterous notion, would not have come about without the substantial support of many generous individuals.

My advisor, Philip Resnik, has given his time and talent in ways that simply cannot be adequately acknowledged. Philip has provided constant, immeasurable intellectual challenge and inspiration. His energy and enthusiasm are renowned and I benefited greatly from them time and again. I thank him also for always believing in the work. At a practical level, in any situation, large or small, Philip was able to identify a realistic set of parameters that in each case proved essential to engineering this endeavor successfully. Buying a house around the corner from me was, of course, the ultimate expression of his dedication to my work.

Co-advisor Don Hindle has been the project's sage, a sounding board for ideas, who always responded with insight, wisdom, and practical advice. Don dedicated substantial amounts of his time to this project but was remarkably efficient and concise in saying just the right thing at the right time. On top of countless discussions about the work itself, his instruction to "visualize the dissertation" provided perhaps the three most powerful words spoken to me in the course of this effort.

I thank my committee members, Professors Amy Weinberg, Jeff Lidz, and VS Subrahmanian for their insightful questions, their support, and the contribution of their valuable time. I owe the Department of Linguistics at the University of Maryland, along with the Computer Science department, a great deal of gratitude for their extended support. Maryland was a good fit in balancing my life as a working professional, a family guy, and a student. I particularly thank those professors and others who have taught or worked with me over these years that have been the most rewarding educational experience of my life: Norbert Hornstein, David Lightfoot, Juan Uriagereka, Andrea Zukowski, Colin Phillips, Paul Pietroski, Laura Benua, Linda Lombardi, Mark Arnold, Bonnie Dorr, Mari Broman Olsen, and Rebecca Hwa.

In my professional life I have had the great fortune to work with many very talented and supportive people. I thank George Krupka in particular for his help in making this achievement possible. Thanks also go to Elena Spivak, Lorraine Bryan, Tony Davis, Cheinan Marks, and Anand Kumar

For generously sharing of their work (and sometimes, data), I thank Wei-Hao Lin, Theresa Wilson, Janyce Wiebe, Lillian Lee, Matt Thomas, Bo Pang, Ed Kako and Talke MacFarland.

I thank Becky Bishop Resnik for frequently alerting me to related work that I would not have discovered otherwise. I thank Chip Denman for crucial help with statistical analyses. Thanks to Frank Keller and the WebExp development group at the

University of Edinburgh for sharing and assisting with their system. Thanks to Mark Damrongsri for Web administration.

I thank my parents, my in-laws, the rest of my family, and my friends for giving me endless encouragement and support. My kids, Sabrina and Owen, have been rooting for me like nobody else, despite the intrusion that this work has been. I thank them for their tolerance, and their smiles.

My wife, Linda Parker Gates, is an essential partner to whatever success I am able to claim. She inspires me and envisions things for me that I could never imagine alone. And she has encouraged, cajoled, and hounded me, as necessary, to bring this work to fruition. Thanks for seeing this through with me.

Table of Contents

List of Tables

List of Figures

List of Equations

1 Identifying Implicit Sentiment

1.1 Introduction and Overview

The automatic detection of sentiment in text is generating great interest and has become an important and active area of research in natural language processing. Interest in this problem is motivated by a variety of information requirements. For example, for certain information seekers, search by keyword or classification by topic is insufficient because their specific information needs require a more refined level of classification. They require topical information that is filtered by the attitude or sentiment toward the topic of interest. The ability to deliver information that is characterized in this manner will help reduce the costs of information seeking and provide more valuable resources to analysts and decision makers who routinely confront an ever increasing mountain of information.

As specific examples, consider the following applications of automated sentiment analysis:

- Media Studies and Political Economy. Analysis of sentiment and opinion is central to many studies in journalism, media studies, and political economy (Gentzkow and Shapiro 2006a, 2006b; Groseclose and Milyo, 2005; Quinn et al. 2006). Automated filtering or classification of information with respect to sentiment can greatly facilitate this kind of research.
- Email classification and routing for prioritization. In a customer service setting, electronic messages are typically classified by topic for the purpose of routing the messages to appropriately trained customer service representatives. It would be valuable to companies to enhance topic-based routing by detecting the presence of negative sentiments (anger, frustration) in a customer's message. Some preliminary work has been done in this specific area (Durbin, et. al. 2003). The potential benefits include improved customer retention and the gathering of information about the nature of previous contacts with customer service.
- Forum and Blog Search. Online forums and blogs are an increasingly widespread medium for the exchange of experiences, opinions, and complaints. Keyword search of forums can help find messages generally related to subject areas of interest. It would be far more useful to be able to search for or filter messages based on their positive or negative sentiment toward a particular subject.
- Business Intelligence.[1] Classifying forum, blog, and email data by attitude or sentiment can serve business and marketing intelligence purposes. What is the 'word on the street' with respect to a certain product, service, or enterprise as a whole? What are some key quotes or phrases that vividly illustrate these

[1] Commercialization of such information services is well underway. See, for example, http://www.nielsenbuzzmetrics.com, http://www.biz360.com, and http://bazaarvoice.com/.

opinions? How do these opinions spread online and what influence do they have?

- Open Source Intelligence. E.g., what points of view are evident in the Arab press about potential new sanctions against Iran over its nuclear program?
- Citizen-Government Communications. Governmental institutions in the United States are working to solicit citizen input via the Internet, with such goals as electronic rulemaking (*eRulemaking*) for the "electronic collection, distribution, synthesis, and analysis of public commentary in the regulatory rulemaking process," which may "[alter] the citizen government relationship" (Shulman and Schlosberg, 2002; Thomas, Pang, and Lee, 2006).

With these kinds of applications in mind, there are several specific threads of research in *sentiment analysis*, as the problem is generally referred to. Sentiment analysis has been investigated at both the document and sub-document level. Opinion mining in particular seeks to distinguish objective vs. subjective statements, operating at the level of a clause, sentence, or passage. Document-level sentiment classification seeks to determine if a document indicates a positive opinion or shows support for, or a negative opinion or opposition to, the topic of discussion. Movie and product reviews in particular have received significant attention for this task.

Within the broader realm of sentiment analysis, this dissertation focuses on the automatic identification of *implicit sentiment*, at the document level. This is a text classification task that I distinguish from other work in sentiment analysis in two ways. First, I do not assume that the documents to be classified with respect to their sentiment are necessarily overt expressions of opinion. They rather are documents that might reveal a *perspective*. As described by Lin et al. (2006), this means the identification of the *point-of-view* from which the document is produced. Second, because of the distinction between implicit and explicit sentiment, I do not attempt to distinguish subjective from objective statements or exploit the distinction in building text classifiers.

Many different approaches to sentiment analysis and sentiment classification have been reported in the literature and they are reviewed in Section 1.3. In this dissertation, I develop a novel approach to the identification of implicit sentiment. The approach is motivated by ideas drawn from the literature on lexical semantics and argument structure, which I support and refine through psycholinguistic experimentation. I employ targeted lexical selection for verbs and nouns and extract semantically motivated features for the selected terms, including argument structure. I then exploit these features with machine learning methods for large scale text classification with respect to perspective. While I begin with some manually selected lexical resources, I generalize the method to require no manual selection or hand tuning of any kind. I show this linguistically motivated method to provide satisfying results across several text domains, and obtain the best classification accuracies yet reported for several classification tasks.

My text classification method rests on a several foundational ideas. In this work I take seriously the idea that differences in linguistic form always indicate at least some difference in meaning (Bolinger, 1968). This is particularly relevant to identifying *implicit* sentiment. The hypothesis is that speakers employ specific constructions in a manner that exploits these (sometimes subtle) differences in meaning in a way that reveals, intentionally or not, aspects of the speakers' point-of-view. When referring to 'differences in linguistic form' this is meant to include everything from classic diathesis alternations such as the dative alternation, causative-inchoative alternation and the active-passive alternation, to differences in the nominal forms for discourse referents (such as the use of proper nouns or *–er* nominalizations) and on down to specific lexemes. In this dissertation I exploit reflections of these cues as features for text classification tasks.

Another foundation of my approach is theoretical work that has articulated a decompositional semantics for transitive clauses (Dowty 1991, Hopper and Thompson 1980). I take the semantic components of transitivity as outlined by Dowty (1991) and Hopper and Thompson (1980) as principles underlying observed differences in meaning among forms. These principles have been proposed in the contexts of argument linking and in an analysis of the nature of transitivity in grammar and discourse, and thus have an established role at the syntax-lexical semantics interface. Both syntactic form and lexical entries are believed to reflect the operation of these principles. Chapter 2 provides additional background on these ideas and presents evidence from psycholinguistic experiments that support and motivate the use of linguistically informed features in text classification for sentiment.

Finally, I consider the fact that the level of semantic analysis required to understand how the semantic components of transitivity are actually reflected in any given clause is well beyond the current capabilities of natural language processing systems. However, I subscribe to the notion, as articulated by Gamon (2004a), that consistent use of a language processing system will at least be consistent and systematic in its errors. In that context, machine learning techniques can effectively identify and exploit features amidst the noise. The key, therefore, is to identify features that are observable reflexes of the semantic components that can be practically extracted by current NLP techniques, even if noisily so. I will thus refer to the linguistically motivated features I use as Observable Proxies for Underlying Semantics (OPUS).[2]

1.2 Defining the Task

Amid the flurry of recent NLP research in sentiment analysis, a variety of related terms are invoked: attitude, opinion, affect, emotion, subjectivity, bias, slant, spin, perspective, and point-of-view. Unlike the objects of interest in tasks such as named entity extraction, information extraction, and fact-oriented question answering, defining precisely what is meant by these terms or how they differ among themselves

[2] Not to be confused with the OPUS software system described in Burstein (1979).

is not easily accomplished. Psychologists and social scientists studying these matters try to be more precise in their definitions, though there, as well, consistency and precision can be difficult to establish (Ekman, 1994; Oskamp 1991). For example, Oskamp (1991) illustrates various levels of definition for *attitude*, which appear to at least try to establish some compatibility with each other (see Table 1).

Level	**Definition**
Comprehensive	An attitude is a mental or neural state of readiness, organized through experience, exerting a directive or dynamic influence upon the individual's response to all objects and situations with which it is related. (Allport, 1935)
Simple	Attitudes are likes and dislikes. (Bem, 1970)
Emphasis on Evaluation	Attitude is a psychological tendency that is expressed by evaluating a particular entity with some degree of favor or disfavor. (Eagly & Chaiken,1993)
Emphasis on Learning and Consistency	An attitude is a learned predisposition to respond in a consistently favorable or unfavorable manner with respect to a given object. (Fishbein & Ajzen, 1975)

Table 1 – Various definitions for *attitude* [(from Oskamp (1991)]

Yet despite Oskamp’s careful treatment of the term *attitude*, he and others in that community tend to use the term *opinion* almost interchangeably. Within the NLP community, there is considerably less concern about formal and precise definitions for these concepts. However, several more or less commonly accepted task definitions have emerged within NLP research.

Under the main heading of ‘affect analysis’, several subtasks have been the target of recent research. A kind of ‘general’ affect analysis task is to determine the overall affective or emotional content of text, either with respect to multiple affect categories (Subasic and Huettener, 2000; Read, 2004) or in selecting one of several possible affect categories as primary (Mishne 2005 – therein called ‘mood’). There is also the task of classifying texts at the document or message level as a whole with respect to its attitude, positive or negative. This task has been investigated in the context of knowing in advance that the text in question purports to indicate some polar attitude (Pang et al. 2002; Pang and Lee 2004; Read 2005) as well as in the harder but more interesting context of extracting attitude (Grefenstette 2004) or affect (Subasic and Huettener 2000) in more general expository text. More fine grained applications determine opinion with respect to specific entities, and provide both a polar assessment of the opinion (positive or negative) as well as a measure of its intensity (Cesarano et al. 2006, 2007). There is also the task of distinguishing subjective from objective text passages (Bethard et al. 2004; Riloff and Wiebe, 2003; Wiebe et al. 1999; among others), otherwise known as the identification of ‘opinion’ clauses. While seemingly highly related, I view the identification of opinion clauses as a task distinct from my focus here. Lin et al. (2006) distinguish the task of identifying the *perspective* or *point-of-view* from which a document is written. The NLP tasks investigated in this dissertation are most closely aligned with Lin's task definition (and indeed, I experiment with Lin's corpus data and classification tasks in Chapter 4). Central to my work is the notion that clauses that are not obviously subjective will be useful in determining the perspective of texts which may or may not be overtly evaluative.

This brings us to the related notions of *bias*, *slant*, and *spin*. Again, precise definitions are difficult to establish. But at the least, these terms imply a *hidden* and *deliberate* attempt to frame entities and events in a particular way so as to serve some agenda. The notion that such attempts are hidden captures that fact that accusations of bias and slant are made in situations where an expectation of neutrality is operative. The notions of perspective or point-of-view as I use them here can be considered related to bias, slant, and spin. Perspective is distinguished from these, however, because it can also be assumed of situations when neutrality is not specifically expected. At the same time, perspective is distinguished from opinion or subjectivity in that one can exhibit perspective without necessarily expressing an opinion or making subjective statements.

A few anecdotal examples will help illustrate some of these points. The first example comes from the Washington Post (Blaine, 2006), in an article on the controversy over the listing of killer whales in Puget Sound as an endangered species. The article notes that supporters of the listing generally referred to the animals as 'orcas', while opponents generally referred to them as 'killer whales'. Here, lexical choice is employed as a 'spin' tactic.

Another example comes from a study by the media watchdog group HonestReporting.com. They conducted an informal study of Reuters newswire service headlines related to the Israeli-Palestinian conflict.[3] An example comparison they make is as follows:

- "Israeli Tank Kills 3 Militants in Gaza – Witnesses"

  Israel named as perpetrator; Palestinians ("Militants") named as victims; described in active voice.

- "Israeli Girl Killed, Fueling Cycle of Violence"

  Palestinian not named as perpetrator; killing described in passive voice.

HonestReporting.com accused Reuters of systematic bias in favor of the Palestinians and against Israel. They claimed, based on the headline data they collected and studied, that Reuters routinely phrased such headlines in a way that portrayed Israelis as aggressors and Palestinians as victims. As a journalistic organization, Reuters is particularly vulnerable to the accusation of bias. Note that the example headlines

[3] See http://www.honestreporting.com/articles/critiques/Study_Reuters_Headlines.asp.

shown above are not specifically subjective or opinion expressing statements. If a person was simply describing such events to a friend, a systematic pattern of expression like this would be an example of perspective.[4]

Consider another example, taken from Gentzkow and Shapiro (2006b). They provide the opening portion from three different news reports of the same event, a battle involving American troops in the Iraqi city of Samarra on December 2, 2003.

> Fox News:
>
> In one of the deadliest reported firefights in Iraq since the fall of Saddam Hussein's regime, US forces killed at least 54 Iraqis and captured eight others while fending off simultaneous convoy ambushes Sunday in the northern city of Samarra.
>
> The New York Times:
>
> American commanders vowed Monday that the killing of as many as 54 insurgents in this central Iraqi town would serve as a lesson to those fighting the United States, but Iraqis disputed the death toll and said anger against America would only rise.
>
> The English-language Web site of satellite network Al Jazeera (AlJazeera.net):
>
> The US military has vowed to continue aggressive tactics after saying it killed 54 Iraqis following an ambush, but commanders admitted they had no proof to back up their claims. The only corpses at Samarra's hospital were those of civilians, including two elderly Iranian visitors and a child.

Gentzkow and Shapiro (2006b) comment as follows: "All the accounts are based on the same set of underlying facts. Yet by selective omission, choice of words, and varying credibility ascribed to the primary source, each conveys a radically different impression of what actually happened. The choice to slant information in this way is what we will mean … by media bias." By using the term *choice*, Gentzkow and Shapiro imply that these journalistic organizations have deliberately framed these events in a particular manner, and hence the accusation of bias. I will remain agnostic with respect to that claim, but this example is highly illustrative. The three event descriptions in this example are, as in the Reuters example, not obviously subjective or opinion-bearing statements. And again, outside a journalistic context, one might attribute the differences to variations in the perspective of the authors.

These examples illustrate the kind of implicit sentiment that I target in my psycholinguistic experiments and computational classification tasks. My use of

[4] I discuss the HonestReporting.com example in greater detail in Chapter 2.

linguistically motivated features in some sense formalizes, in an engineered setting, heretofore descriptive, informal and/or anecdotal beliefs about how writers reveal their implicit sentiments: that speakers and writers *spin* their language in a way favorable to their position, not only in opinion-expressing text but also in expressions that are not specifically opinionated.

1.3 Related Work

I review here some of the most significant and related work in computational sentiment analysis. I first focus on lexically oriented methods, and then consider methods using additional feature types such as structural features, all described in the context of the particular tasks to which they are applied.

Significant success has been achieved in sentiment analysis using lexically based approaches, often targeting adjectives and adverbs. Much work has been done to automatically acquire opinion word resources (Gamon and Aue, 2005; Hatzivassiloglou and McKeown 1997; Baroni and Vegnaduzzo 2004; Takamura et al., 2005; Turney and Littman, 2003). Turney and Littman's pointwise mutual information method has been applied successfully in a number of sentiment analysis applications, including opinion retrieval from corpora of email and blogs (Oard et al. 2006; Wu et. al. 2006) and affect classification in short story fiction (Read 2004).

Other lexically based approaches to sentiment analysis have used lexicons manually crafted specifically for sentiment or affect analysis (Durbin et al 2003; Nigam and Hurst, 2004, 2006). Cesarano et al. (2006, 2007) semi-automatically developed an opinion-expressing word bank by applying a flexible word scoring function based on document-level human annotations. The annotations included harshness scores that reflect the degree to which documents were considered positive or negative. In their OASYS opinion analysis system, the word bank scores feed into several document scoring functions that both qualitatively assess opinion with respect to specific topics, as well as generate a quantitative measure of that opinion. (Cesarano et al., 2006, 2007; Benamara et al., 2007).[5] Unlike many other research systems, OASYS runs continuously and has been applied to a growing inventory of millions of documents.

Subasic and Heuttner's (2000) affect analysis system relies on a fully manually developed word list and distinguishes affect as a set of emotional categories that does not address opinion types directly. Subasic and Huettner describe an affect word lexicon of about 4000 terms, of various parts of speech. Each word is manually annotated with an affect category and a centrality and intensity value with respect to that category. The lexicon was shown to be useful for determining the overall affect content of documents, though the affect categories are numerous and sometimes quite subtle in their distinctions (e.g. repulsion, aversion). Their affect measurement system provides only gross overall scores for each affect category, and does not provide information about specific entities or events to which the affect is attributed,

[5] See http://oasys.umiacs.umd.edu/oasys/ for more information and to try the system directly.

in contrast to the OASYS system, which assesses opinion directly with respect to specific topics.

Grefenstette et al. (2004) used a modified form of Subasic and Huettner's lexicon to build an opinion mining application. They removed the centrality and intensity values for each lexical entry, but tagged each entry with respect to whether the word had positive or negative connotations, similar to much other work, e.g. the General Inquirer Dictionary (Stone 1997). Extracting news stories about a particular person, they calculated the simple ratio of positive to negative affect words in a text window around the mention of the person. This approach was tested by extracting relevant text from sites with known slants, and it was shown to be a fairly effective, gross measure of attitude toward a person. This work is most notable for its attempt to detect slant or bias in ostensibly objective text.

Following Klavans and Kan (1998), Chesley et al. (2006) use proprietary verb class definitions and Wiktionary[6]-derived adjective information for sentiment classification of blogs.

Building on lexically-driven work, a number of researchers have examined syntactic and other more linguistically-informed approaches to sentiment analysis. In a series of studies, Wilson and colleagues have built opinion extraction systems trained on corpora with detailed, manual annotation to extract lexical and syntactic features for the tasks of extracting opinion clauses of varying strengths, and recognizing the contextual polarity of expressions (Wilson et al., 2004; Wilson et al., 2005; Wilson et al. 2006). Similarly, Bethard et al. (2004) use a small corpus manually annotated with propositional information to experiment with their newly defined task of propositional opinion detection.

Riloff and Wiebe (2003) describe a bootstrapping method that starts with a small set of lexical seed terms that are known to have been successfully used for subjectivity detection. They build high-precision subjectivity classifiers from these terms which are then used to bootstrap a set of linguistically richer structural clues for subjectivity detection using extraction pattern methods.

Yi et al. (2003) build a sentiment lexicon and couple that with a part-of-speech based sentiment pattern database and successfully apply their system to the product review domain. Similarly, in applying sentiment classification to the noisy domain of customer feedback data, Gamon (2004b) successfully uses part-of-speech trigrams but also improves results a bit further using deeper linguistic features such as semantic relation trigrams, constituent structure patterns, and tense information.

Building on adjective-driven approaches, Whitelaw et al. (2005) use semi-automated methods to build a database of appraisal groups, phrase-level items built from a lexicon of adjectives and their modifiers, resulting in a set of expressions such as

[6] See http://en.wiktionary.org/wiki/.

“very good” and “not terribly funny”. They successfully apply this method to the movie review domain.

Among prior authors, Gamon’s (2004b) research is perhaps closest to the work described here, in that he uses some features based on a sentence’s logical form, generated using a proprietary system. However, his features are templatic in nature in that they do not couple specific lexical entries with their logical form. Hearst (1992) and Mulder et al. (2004) describe systems that make use of argument structure features coupled with lexical information, but neither provides implementation details or experimental results. To my knowledge, no previous research in sentiment analysis draws an explicit and empirically supported connection between theoretically motivated work in lexical semantics and reader’s perception of sentiment.

1.4 Dissertation Roadmap

The dissertation proceeds as follows:

Chapter 2 provides psycholinguistic evidence for a predictive connection between the underlying semantic properties of transitive clauses and readers' perceptions of the author's implicit sentiment. This result lays the groundwork for the exploitation of features based on these semantic properties for the classification of text with respect to implicit sentiment. The results in this chapter also contribute support for the psychological reality of the semantic properties proposed in the lexical semantics literature.

Chapter 3 presents my method for extracting OPUS features, as defined in Section 1.1. I then introduce the semantic field of kill verbs and discuss my motivation for selecting them for investigation into sentiment classification. This leads to a discussion of the development of a corpus of documents related to the death penalty. I first demonstrate the value of OPUS features for document-level implicit sentiment classification using this corpus. I then demonstrate improvements over baseline features when using a manually developed but well-motivated list of target terms related to killing. Finally, I successfully extend the method to employ a list of target terms that is fully automatically derived. The results of this chapter confirm the hypothesis that classification of text for implicit sentiment can be improved with the use of features motivated by the results in Chapter 2.

Chapter 4 extends the methods of Chapter 3 and applies them to two additional domains using corpora publicly available to the research community: one focused on the Israeli-Palestinian conflict and the other comprising U.S. Congressional floor debates. In these experiments, I extend the use of OPUS features for classification, and introduce a novel classifier combination method. I obtain the best classification accuracies yet reported for these corpora. By successfully extending the method of Chapter 3 to additional domains, I provide strong evidence for the generality of the method.

Chapter 5 provides discussion and conclusions, and outlines a number of areas for future work.

2 Connecting Lexical Semantics to Perceived Sentiment: Psycholinguistic Evidence

2.1 Introduction

This chapter introduces and reviews ideas from the literature on lexical and constructional semantics, framing these ideas in terms of how they underlie my focus on linguistic features for sentiment classification. The primary goal of this chapter is to provide evidence for a predictive connection between underlying semantic properties of transitive clauses and readers' perceptions of the author's implicit sentiment. In the computational work reported in Chapters 3 and 4, I exploit features that are observable proxies for these underlying semantic properties in experiments on text classification with respect to implicit sentiment. In establishing this connection between underlying semantic properties and sentiment, I provide a foundation, grounded in linguistic theory and supported by psycholinguistic evidence, for the linguistic features I experiment with computationally. Moreover, I provide some evidence for why they work.

As a vehicle for investigating the connection between semantic properties and implicit sentiment, I experiment with data modeled in part on the Reuters headline data collected and studied by the media watchdog organization HonestReporting.com, mentioned in Chapter 1. That organization conducted a study that claimed to have found bias, against Israel and in favor of the Palestinians, by the ostensibly neutral Reuters news wire service. This study utilized a small corpus of Reuters headlines describing acts of violence, primarily killings, collected over a one-month period. Their analysis hinged upon factors such as the presence or absence of volitional agents and use of passive voice in the headlines. To quote one example from their analysis, they present the following comparison of these two headlines:

> "Israel Kills Three Militants; Gaza Deal Seen Close"
>
> Israel named as perpetrator; Palestinians ("Militants") named as victims; described in active voice.

> "Bus Blows Up in Central Jerusalem"
>
> Palestinian not named as perpetrator; Israelis not named as victims; described in passive voice (sic).

The analysis is flawed in several respects and is rightly criticized in a blog entry by Pullum (2004). However, it is interesting that the study of the Reuters data, while informal and flawed, focused on elements of language use such as named volitional agents, present or omitted objects, lexical choice of nominalization, and (an inaccurate analysis of) active vs. passive voice. Related elements of language use such as number of participants, volition, agency, kinesis, affectedness and telicity have been studied at length in the scientific literature in a wide variety of settings and frameworks. They have served as the underpinnings for an explanation of garden path effects in reduced relative clauses (Filip et al., 2001), for a seminal theory of thematic roles and argument selection (Dowty, 1991) and in a comprehensive analysis of the fundamental role of transitivity in grammar and discourse (Hopper and Thompson, 1980).

The work of Dowty (1991) is particularly relevant in this regard. Dowty's theory of "thematic proto-roles" is based on the premise that the surface expression of (verbs') arguments in linguistic expressions is closely connected to properties of those arguments and of the event. For example, if the referent of an argument is volitional and causal with respect to the event communicated by the verb, properties traditionally associated with the thematic role of agent, then it is more likely to surface in subject position. If the referent of the argument undergoes a change of state and is causally affected by another participant in the event, properties traditionally associated with a patient thematic role, then it is more likely to surface as an object. Similarly, Hopper and Thompson (1980) describe *semantic* transitivity as a complex of gradient properties that connect conceptual and semantic features to specific morphosyntactic reflexes in eventive clauses across a wide variety of languages. They demonstrate, for example, a correlation between syntactic transitivity and a high degree of agency (being human or otherwise autonomous and exercising independent causation over another entity). Moreover, they show that clauses exhibiting high degrees of any of the gradient properties they distinguish tend to be the clauses that are foregrounded in discourse. I thus predict that the expression of sentiment is connected in part with how particular entities are profiled in this manner, which can be revealed in text by the grammatical relations in which they appear.

(1)"Israeli Troops Shoot Dead Palestinian in W. Bank"

(2) "Israeli Girl Killed, Fueling Cycle of Violence"

Examples (1) and (2) illustrate the potential connections among semantic properties, surface form, and sentiment. Both (1) and (2) use highly causative verbs, but (1) contains a volitional agent, an overt result expression, and an overt, highly affected object; in contrast, (2) omits the overt agent, leaving no argument for which to infer volition.

The experiments in this chapter formally investigate the question of whether these underlying semantic properties and their observable forms are predictive of the implicit sentiment expressed by those forms. The chapter proceeds as follows. In

Section 2.2, I discuss in greater detail the semantic properties introduced by Dowty (1991) and Hopper and Thompson (1980) that I link to the perception of implicit sentiment, and I review additional related literature. In Section 2.3, I describe the framework which guides the experimental designs within the chapter. I then preview the main results of the chapter with respect to this framework. In the heart of the chapter, Section 2.4, I report on the psycholinguistic investigation of the connection between semantic properties and perception of sentiment.

2.2 Background: Lexical Semantics and the Syntax and Semantics of Transitivity

Much of the research in lexical semantics concerns how the semantic properties of particular verbs, or classes of verbs, are linked to the syntactic phenomena associated with those verbs. Dowty (1991) and Hopper and Thompson (1980) propose theories that could be roughly described as 'decompositional' in the sense that they factor out specific elements of semantics that appear to trigger particular syntactic reflexes. Dowty (1991) focuses on the elimination of the traditional thematic roles of agent and patient, which are typically linked to grammatical roles such as subject and object. He describes a theory in which these thematic roles are replaced by prototype roles ('proto-agent' and 'proto-patient'), each typically characterized by a set of specific attributes. His argument linking theory is essentially that whichever NP in a clause has the most proto-agent properties is linked as the subject, and whichever NP in a clause has the most proto-patient properties is linked as the object.

Hopper and Thompson (1980) focus on the core linguistic notion of Transitivity from a clausal perspective.[7] They distinguish a set of semantic components constituting Transitivity, and show how each component can have a 'high' or 'low' value with respect to its contribution to Transitivity. Thus each semantic component works to indicate the degree to which some 'transfer', 'change', or 'effect' takes place in an event represented by a Transitive linguistic encoding.[8] Their central claim is the Transitivity Hypothesis, which essentially says that if a clause exhibits a high Transitivity value for any of the components, then other components exhibited elsewhere in the clause will also be high in their Transitivity values. The converse is implicit in their hypothesis. Importantly, they show that clauses exhibiting high Transitivity are mostly the ones that encode the foregrounded or profiled events within a narrative, which will potentially be reflective of the kind of variation in event profiling discussed in Section 1.2.

Table 2 shows one possible mapping between Dowty's proto-role properties and Hopper and Thompson's components of Transitivity.

[7] From this point forward, following Hopper and Thompson (1980) I will use 'Transitivity' with a capital 'T' to distinguish the idea of *semantic* transitivity from that of *syntactic* transitivity, which refers more specifically to the structural presence of an object of the verb.

[8] I use the term 'encoding' in fairly standard fashion to indicate a particular surface realization, from among the set of possible realizations.

Dowty Proto-Role Property	**Hopper and Thompson Component**
Volitional involvement in the event or state	*Volition*
Sentience (and/or Perception)	
Cause event or change of state in object	*Agency*
Movement relative to the position of another participant	*Kinesis*
Exists independently of the event	
Undergoes change of state, Causally effected by another participant	*Affectedness of Object*
Does not exist independently of the event	
Stationary relative to movement of another participant	*Kinesis*
	Object individuation
(Exists independently of the event?)	*Subject-object individuation*
(Incremental Theme?)	*Punctuality*
	Aspect (Telicity)

Table 2. A Mapping Between Dowty (1991) and Hopper and Thompson (1980)

In a similar vein, Pustejovsky (1991) proposes a ‘decompositional’ approach, one that is centered within lexical entries themselves. Pustejovsky’s framework decomposes lexical entries into four substructures. His ‘Qualia’ structure appeals to the notions of object creation and individuation, among others. Levin (1993), in a by now classic study of English verbs, organizes English verbs into classes according to diathesis alternations as a means of isolating syntactically relevant elements of verb meaning. In doing so, she appeals to the notions of kinesis, aspect, change of state, and agency as factors relevant for classifying verbs by their syntactic behavior.

More recent work has begun to establish the psychological reality of such lexical semantic theories. Kako (2006a) experimentally established the psychological reality of Dowty’s (1991) proto roles hypothesis. In Kako’s experiments, subjects reliably attributed Dowty’s proto-agent properties to syntactic subjects, and reliably attributed Dowty’s proto-patient properties to syntactic objects, even when nonsense words were used. Wright (2001) found that distinctions between internally and externally caused change of state verbs could be explained in terms of concepts similar to those distinguished by Hopper and Thompson (1980). Wright found that internally caused change of state verbs were more marginally grammatically acceptable in transitive constructions than were externally caused change of state verbs. In contrast, the more prototypically transitive externally caused change of state verbs were found to be much more acceptable in transitive constructions. Wright appeals to the notions of agency, volition, and control as central to an explanation of the syntactic distribution and selectional restrictions of internally and externally caused change of state verbs.

Additionally, research has begun to examine the semantics of syntactic frames themselves (Fisher, Gleitman and Gleitman, 1991; Goldberg 1995; McKoon and Ratliff 2003, 2007). Stefanowitsch and Gries (2003) propose associations between meanings and syntactic frames for a variety of different constructions. Particularly interesting are several constructions they claim inherently carry negative sentiment:[9]

- The [*X think nothing of* V_{gerund}] construction – This idiomatic expression is primarily associated with contextually undesirable or risky actions, e.g. *In their present mood people would* ***think nothing of*** *mortgaging themselves for years ahead in order to acquire some trifling luxury*.

- The [into-] causative construction: [S_{agent} V $O_{patient/agent}$ into A_{gerund}] – This construction is frequently associated with verbs of force, coercion, and trickery, e.g. *He* ***tricked*** *me* ***into employing*** *him.*

- The [*cause X*] construction – This construction predominantly indicates a negative view of an event, e.g. *I am sorry to have* ***caused*** *you some inconvenience.*

Kako (2006b) showed that syntactic frames carry meaning irrespective of particular lexical entries. Lemmens (1998) attempts to synthesize both lexical meaning and constructional meaning into a theory of argument structure. Lemmens attempts to more naturally allow for novel usages, integrating an explanation for coercion effects or sense extensions. And in this flexibility, it is akin to Pustejovsky (1991). Corpus-driven examples are crucial in this kind of work, and Lemmens (1998) in particular uses a large set of corpus examples, almost exclusively, as the data to be explained. The trend in this work is toward a view that 'argument structure' is not a static property of verbs, but rather should be thought of as the expression of the dynamic interplay among lexical and structural factors, both of which carry meaning. The dynamic interplay is driven by the modulation of meaning at the core of creative language use and language change. The investigation reported here does not specifically commit to or further this line of research, but the interplay between underlying semantic properties and observable structural features remains an important theme throughout. My approach to text classification will use corpus-based features that will possibly reflect novel usages and will use machine learning to discriminate usage trends for verbs and nouns, putting actual usage at center stage.

In order to begin to establish a connection between event encoding choice and implicit sentiment, several reasonably well established facts point to useful starting points. Hopper and Thompson (1980) and Lemmens (1998) focus squarely on the concept of semantic Transitivity. Transitivity is fundamental to language, and causation is fundamental to transitivity. As described by Levin (1993), McKoon and MacFarland (2000, 2002) and others, the source of causation in an event can be

[9] These constructions, however, are generally too rare to be usefully exploited for sentiment analysis.

internal or external to the object of the caused event. Change of state verbs in particular show this distinction. Consider these examples:

(1) The wind eroded the beach.
(2) The boy broke the vase.

(3) The beach eroded.
(4) The vase broke.

Both *erode* and *break* exhibit the causative-inchoative alternation, as shown in (1)-(4). The verb *erode* is considered an internally caused change of state verb because erosion is an event that comes about due to the inherent properties of the argument that changes state, *the beach*. The verb *break*, however, is an externally caused change of state verb because there necessarily must be some external agent that brings about the event denoted by *break* even when that agent is unexpressed, as in (4) (Levin and Rappaport Hovav, 1995). McKoon and MacFarland (2000, 2002) posit that externally caused change of state verbs carry a lexical semantic event template containing two subevents, as in (5), and internally caused change of state verbs carry a lexical semantic event template with only one subevent, as in (6):

(5) (y CAUSE (BECOME (x <STATE>)))

(6) (BECOME (x <STATE>)

McKoon and MacFarland attribute the ability of an internally caused change of state verb to appear in a transitive frame, as in (1), to a "content" portion of the verb's lexical semantic structure. They distinguish the notion of "inherent participant," in which verbs seem to lexically specify (through "content") a restricted set of possible causers. Thus for literal uses of verbs like *erode* as in (1), there is a selectional restriction to entities like *the wind*. Further, McKoon and MacFarland report psycholinguistic evidence of increased processing times for externally caused change of state verbs over internally caused change of state verbs on a variety of tasks, regardless of the (in-)transitivity of the construction in which they appear. They attribute this distinction in processing times to the increased processing load of the two subevent template.

Lemmens (1998) characterizes the difference between internal and external causation in a distinct but analogous manner. He separates the two into a paradigmatic opposition between transitivity and ergativity, terms he uses in the following circumscribed manner. External causation is captured within the transitive paradigm, while internal causation is captured within the ergative paradigm. The transitive paradigm profiles properties of the agent, and the ergative paradigm profiles properties of the patient. Analogous to McKoon and MacFarland's notion of the "content" portion of a verb licensing an "inherent participant," Lemmens introduces the notion of the possibility of external instigation of the internally-caused event. This characterizes the subject in a sentence like (1) as a co-participant, along with the object, in the change of state event. In this sense, the notion that the event is possible due to some properties of the object (internal causation) is retained. In contrast, external causation remains necessary in sentences like (4).

By whatever means, speakers utilize these fundamental facts about language. Lexemes and constructions are chosen by speakers to elicit a particular construal within comprehenders.[10] That is to say, speakers take an interest in having listeners interpret their language in their preferred and intended way. As mentioned earlier in this section, Kako (2006a) has begun to show that specific semantic components of Transitivity are reliably attributed to subject position in two participant clauses. Extracting lexical-positional features thus should reflect these facts as they play out in real text, such as which event participants are more likely to be, or have been portrayed as, volitional agents. Because the prototypically causative verbs of killing are the objects of previous study and are likely to be good exemplars of the semantic components of Transitivity, the experiments discussed in this chapter investigate this verb class and the construal of transitive clauses projected from its members. The results of these experiments provide evidence for the psychological reality of a decompositional view of Transitivity, and establish the connection between the semantic components of Transitivity and the perception of sentiment.

2.3 Investigative Approach

Levin (1993) works to explain the behavior of verbs by attempting to isolate and identify the components of meaning that appear to trigger that behavior. The focus in that work is syntactic behavior, diathesis alternations in particular. Additionally, by design her research program attempts to push to the limit the notion that such behavior can be explained in terms of verb meaning alone. Levin works to explain syntactic reflexes through the preliminary grouping of verbs according to shared alternations and common semantic elements. Levin (1993) established the groundwork for additional research into determining how elements of meaning are syntactically reflected, and in what ways. Subsequent work has begun to suggest possible modifications to some of the preliminary Levin classes, and to reconsider the characterization of certain elements of meaning. Within Levin's preliminary verb classes, such limits are reflected in the need to cross-classify verbs in unprincipled ways, and in the fact that ultimately, many of the verb classes show little coherence with respect to the syntactic behavior of their verbs. Many exceptions must be noted for such classes. In an example discussed in Lemmens (1998), Levin defines a POISON class of verbs that includes *asphyxiate, crucify, drown, electrocute, strangle, shoot, suffocate*, and more. Some of these verbs allow an inchoative usage, while others do not:

7(a) The witch poisoned Snow White. (Levin 1993)
7(b) *Snow White poisoned. (Levin 1993)

[10] I use *construal* in a relatively non-technical sense to simply indicate the interpretation a comprehender attributes to a particular linguistic signal.

8(a) Somebody drowned Esther Williams. (Lemmens 1998)
8(b) Esther Williams drowned. (Lemmens 1998)

Levin (1993) thus cross-classifies verbs like *drown* in the SUFFOCATE class, described as verbs “related to the disruption of breathing.” In addition to being somewhat arbitrary, this verb class begins to approach the kind of specificity that the program of generalization over verb classes was intended to eliminate.

Moreover, while such cross-classification does begin to address the syntactic alternation issue, it does not highlight the important semantic distinctions at work. Verbs like *poison*, *crucify*, and *electrocute* embody strong elements of an instrument or means, and often then by extension an agent distinct from the object, aligning with external causation. Verbs like *drown* and *asphyxiate*, on the other hand, profile the attributes of the object undergoing the event, aligning with internal causation. This generalization over event participant profiles could be a more natural explanation for the differences in syntactic behavior shown in (7) and (8), and Lemmens uses the agent- or patient-profiling tendencies of the kill verbs as the primary criterion for classifying them.

Viewing these differences in syntactic behavior from the perspective of the event participant profiles projected from a verb and the way these profiles are amplified by particular encodings prompts the present investigation of the semantic components of Transitivity. Specifically, subject-object individuation, affectedness of object, volition, agency, and movement relative to other participants all seem to be directly related to the question of whether the lexical semantics of a verb profiles the agent or the patient. The presence or absence of specific participants in any given encoding could further modulate the degree to which an encoding exhibits each of these semantic components.

In Section 2.4, my results confirm the hypothesis that the form of an event encoding affects the perceived sentiment regarding participants in the event. I also show that experimental participants were sensitive to the element of internal causation, or patient-profiling, in ergative verbs. Participants rated these verbs as exhibiting more sympathetic sentiment than transitive verbs within my experimental paradigm. Using multiple linear regression models, I establish that the degree to which specific semantic components of Transitivity are exhibited within a clause predicts the implicit sentiment attributed to the clause. A principal components analysis further supports the connection between the semantic components and implicit sentiment, and provides additional evidence for Dowty's distinction between Proto-Agent and Proto-Patient properties.

Taken together, the results in this chapter establish that a set of underlying components of meaning, motivated by the lexical semantics literature, can be used as the basis for statistical classifier models that predict author sentiment on the collective basis of sentences that are not necessarily overtly subjective or evaluative.

2.4 Psycholinguistic Investigation: Linking Transitivity Components to Sentiment Construal

In this section, I present a formal investigation of the hypothesis that semantic components of Transitivity in sentences predict the construal of sentiment in those sentences.

The specific focus I take in investigating links to the interpretation of sentiment is motivated by two studies in particular. The first study is the scholarly work of Lemmens (1998), which as discussed in Chapter 1, studies verbs of killing in depth. Lemmens notes that the choice of this semantic field was motivated by the generally held belief that verbs of killing are causative verbs *par excellence* (Lakoff, 1987; Wierzbicka, 1980). Moreover, especially in literal usage, verbs of killing connote a strong element of physical force, considered fundamental to transitivity (Hopper and Thompson 1980). The second motivating study, discussed in Section 2.1, is the informal work that the media watchdog HonestReporting.com conducted with respect to the Reuters news wire service headlines about violent acts in the Israeli-Palestinian conflict. I thus investigate the link between semantic components of Transitivity and implicit sentiment using experimental materials fashioned as newswire headlines about acts of killing.

I designed and executed two experiments to carry out the investigative program just described. In Experiment 1 I asked participants to rate transitive sentences with respect to the semantic components of Transitivity using an established software suite for web-based data collection.[11] I then analyzed this data for statistically significant differences in component ratings with respect to verb class and construction type, in the vein of the results reported by Kako (2006a,2006b) and Wright(2001). The sentence data from Experiment 1 was used again in Experiment 2, an experiment designed specifically to measure sentiment in headline-like sentences related to acts of violence, modeled in part on the Reuters data. In the analysis of data in Experiment 2, I used the data from Experiment 1 to build statistical models in which semantic component ratings are the predictors for the measurements of sentiment in Experiment 2.

[11] Both experiments were conducted using a customized version of WebExp (Keller, et al. 1998). WebExp is a collection of Java classes that implement functionality to conduct psycholinguistic experiments over the Web. It provides a number of important features, such as non-intrusive participant authentication, response timing, and management of stimulus randomization. Additionally, the Java-based approach prevents participants from returning to previous items to review or change responses, and disallows missing responses. And importantly, data collected using WebExp has been shown to be comparable to data collected in a traditional laboratory setting (Keller and Alexopoulou 2001).

Because WebExp supported only two basic experimental paradigms, sentence completion and magnitude estimation, I developed a custom user interface in Java that I overlaid over the core WebExp components that provided client-server communication, response timing, stimulus data management and randomization, and participant authentication.

2.4.1 Experiment 1

This experiment presented transitive sentences with verbs of killing and asked participants to rate them with respect to a variety of semantic components. This experiment was conducted using WebExp software.

2.4.1.1 Stimuli and Procedure

The primary stimuli were 24 sentences with verbs of killing. The sentences are listed in *Appendix 1 – Experimental Sentences for Experiment 1*. The 24 sentences use 11 verbs of killing. There are six verb instances for each of two verb classes: the 'transitive' class, which tends toward the externally caused end of the spectrum, and the 'ergative' class, which tends toward the internally caused end.[12] Each verb instance is then presented in two formats. The first is a transitive syntactic frame with a human agent as subject, and the second is a nominalization of the verb as subject and the verb 'kill' as the predicate. The kill verb sentences in Experiment 1 were also used as experimental material in Experiment 2, in a form altered slightly to appear to be newswire headlines (e.g. 'The terrorists slaughtered nine hostages.' here became 'Terrorists slaughter nine hostages' in Experiment 2). Further details on kill verb related materials are provided in the discussion of Experiment 2 in Section 2.4.2. Twenty-four distractor sentences with externally and internally caused change of state verbs, taken from McKoon and MacFarland (2000), were mixed in with the presentation of the kill verb sentences.

Each sentence was presented to participants above a list of questions about that sentence. Responses to each question were given on a 1 to 7 Likert scale. Sentences for each participant were presented in a unique random order, with no more than one sentence from the same block ever presented in a row. The blocks are defined by verb class and/or form, as shown in *Appendix 1 – Experimental Sentences for Experiment 1*. Participants were required to provide a response to every question. If participants attempted to proceed to the next item before completing all responses, a warning was issued and the missing item was highlighted.

The questions asked of each item probed both Dowty's proto-role properties as well as Hopper and Thompson's Transitivity components. A sample stimulus is shown in Figure 1. The semantic components and/or proto-role properties corresponding to each question are shown in Table 3. Drawing on Kako (2006a), the Dowty proto-agent property "Causing an event or change of state in another participant" was factored out into two questions, one related to a change of state, and another to the causing of an event. Both can be considered factors in Hopper and Thompson's Agency component.

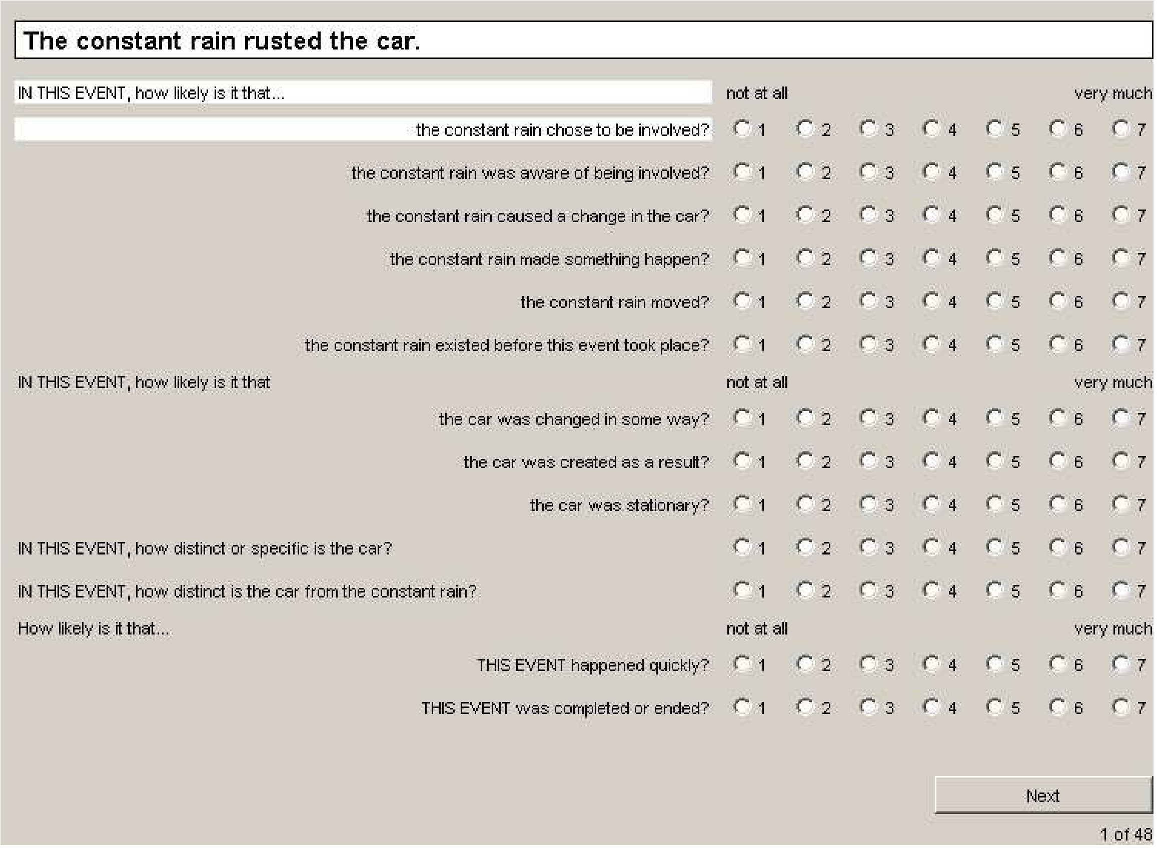
The phrase "IN THIS EVENT" is used repeatedly in order to encourage participants to focus on the particular event described in the sentence, rather than on the entities or events denoted in general. The subject and object phrases are repeated within each

[12] The transitive verb class had five members, and the ergative class had six members. Thus, for the transitive class, the prototype verb 'kill' is used twice in order to balance the number of sentences per class. See section 2.4.2 for further detail on these materials.

question resumption, in order to repeat it for the question at hand, rather than relying on it ‘carrying down’ from the question lead-in.

Question in Experiment 1	**Dowty Proto-Role Property**	**Hopper and Thompson Component**
IN THIS EVENT how likely is it that…		
[subject] chose to be involved?	Volitional involvement in the event or state	*Volition*
[subject] was aware of being involved?	*Sentience (and/or Perception)*	
[subject] caused a change in the [object]?	*Cause change of state in object*	*Agency*
[subject] made something happen?	*Cause event*	*Agency*
[subject] moved?	*Movement relative to the position of another participant*	*Kinesis*
[subject] existed before this event took place?	*Exists independently of the event*	
IN THIS EVENT how likely is it that…		
[object] was changed in some way?	*Undergoes change of state, Causally effected by another participant*	*Affectedness of Object*
[object] was created as a result of?	*Does not exist independently of the event*	
[object] was stationary?	*Stationary relative to movement of another participant*	*Kinesis*
IN THIS EVENT, how distinct or specific is <object>?		*Object individuation*
IN THIS EVENT, how distinct is <object> from <subject>?	*(Exists independently of the event?)*	*Subject-object individuation*
How likely is it that…		
THIS EVENT happened quickly?	*(Incremental Theme?)*	*Punctuality*
THIS EVENT was completed or ended?		*Aspect (Telicity)*

Table 3. Experimental questions and corresponding semantic components and/or proto-role properties



The constant rain rusted the car.

IN THIS EVENT, how likely is it that...	not at all						very much
the constant rain chose to be involved?	○ 1	○ 2	○ 3	○ 4	○ 5	○ 6	○ 7
the constant rain was aware of being involved?	○ 1	○ 2	○ 3	○ 4	○ 5	○ 6	○ 7
the constant rain caused a change in the car?	○ 1	○ 2	○ 3	○ 4	○ 5	○ 6	○ 7
the constant rain made something happen?	○ 1	○ 2	○ 3	○ 4	○ 5	○ 6	○ 7
the constant rain moved?	○ 1	○ 2	○ 3	○ 4	○ 5	○ 6	○ 7
the constant rain existed before this event took place?	○ 1	○ 2	○ 3	○ 4	○ 5	○ 6	○ 7
IN THIS EVENT, how likely is it that	not at all						very much
the car was changed in some way?	○ 1	○ 2	○ 3	○ 4	○ 5	○ 6	○ 7
the car was created as a result?	○ 1	○ 2	○ 3	○ 4	○ 5	○ 6	○ 7
the car was stationary?	○ 1	○ 2	○ 3	○ 4	○ 5	○ 6	○ 7
IN THIS EVENT, how distinct or specific is the car?	○ 1	○ 2	○ 3	○ 4	○ 5	○ 6	○ 7
IN THIS EVENT, how distinct is the car from the constant rain?	○ 1	○ 2	○ 3	○ 4	○ 5	○ 6	○ 7
How likely is it that...	not at all						very much
THIS EVENT happened quickly?	○ 1	○ 2	○ 3	○ 4	○ 5	○ 6	○ 7
THIS EVENT was completed or ended?	○ 1	○ 2	○ 3	○ 4	○ 5	○ 6	○ 7

Figure 1. Sample stimulus item from Experiment 1

2.4.1.2 Participants

A total of 18 participants completed the experiment. All were volunteers and were native speakers of English.

2.4.1.3 Analysis and Results

Linear mixed model ANOVAs were run with kill verb class as the fixed effect, and each semantic component as a dependent variable. Experiment 1 is a repeated measures design in the sense that each participant is measured repeatedly against the different verb classes. Moreover, each participant is measured repeatedly against each specific verb class. The mixed model allows for the participant effect and works to separate the variance internal to each participant from the variance in the dependent variables.

Over the two kill verb classes, transitive and ergative, two components differed significantly; Kinesis (Movement), F(1,430) = 6.471, p = .011, and Punctuality, F(1,430) = 25.617, p < .001. Telicity was nearly significant, F(1,430) = 3.240, p = .073. Subdividing the ergative verbs into two subgroups as suggested by Lemmens (1998), additional significant differences are revealed for Kinesis (Stationary), F(2,429) = 4.066, p = .018, and Telicity, F(2,429) = 3.355, p =.036. Kinesis and Punctuality are most reliably significantly different across verb classes. Along with Telicity, these three components logically make sense as important components for the prototypically transitive kill verbs and their strong association with physical force.

Because the data were perfectly balanced for the appearance of human agents, Volition and Sentience were neutralized in the data. This in effect seems to have transferred some of the differentiation of the components from being sourced in the verbs to being sourced in the syntactic frame in which they were presented. Indeed, analyzing the difference between the means with the stimulus sentence form as the fixed effect supports this view. The means, F ratios, and p values for this analysis are shown in Table 4.

Kill Verb Items	Active Transitive Form	Nominalized Form	Significance
Volition	6.523	1.241	F(1,430) = 3261.111, p < .001
Sentience	6.472	1.269	F(1,430) = 2575.531, p < .001
Cause Change of State	6.903	6.755	n.s.
Cause Event	6.912	6.106	F(1,430) = 30.630, p < .001
Kinesis (Movement)	6.296	2.213	F(1,430) = 712.190, p < .001
Independent Existence	6.639	2.037	F(1,430) = 944.241, p < .001
Causal Effect	6.907	6.875	n.s.
No Independent Existence	1.319	1.176	n.s.
Kinesis (Stationary)	3.551	3.236	F(1,430) = 5.503, p = .019
Object Individuation	5.625	5.454	n.s.
Subject - Object Individuation	6.514	5.75	F(1,430) = 20.096, p < .001
Punctuality	5.176	5.014	n.s.
Telicity	6.273	6.569	F(1,430) = 6.912, p = .009

Table 4. ANOVA results for the semantic components for kill verbs with Form as fixed effect.

When the transitive syntactic frame is held constant but the verb is varied, the data support the conclusion that the presence of an ergative verb in the frame has the effect of reducing Transitivity ratings, and thus such sentences are considered to have less of the ‘effect’ or ‘transfer’ associated with Transitivity. For example, the degree to which a human subject exhibits volitional involvement in the event can be modulated downward by the choice of verb. This difference is the type of distinction I predict to have implications for sentiment detection. Consider each of the kill verb sentence forms, the active transitive (transitive effective) and nominalized form, in isolation. The components for which significant differences were found for each form by the transitive-ergative verb class distinction are shown in Table 5 and Table 6. This data supports Lemmens’s reorganization of the kill verbs primarily with respect to whether they lexically profile the agent or the patient. For the nominalized form, the transitive verb class was rated significantly higher for both Kinesis (Movement), which is movement of the subject, and Independent Existence of the subject. This reflects the fact that the ergative verbs denote actions that are intimately tied to the patient. The activity denoted is rated less likely to involve movement independent from the patient, and it is rated like likely to exist independently of the patient. For the active forms, we see that the transitive verb class is rated significantly higher for Volition and Sentience. All of these sentences had human agents, yet the human agents for the sentences with ergative verbs, which profile properties of the patient, were rated lower on the scale for Volition and Sentience.

A pattern appears to emerge from these results in which patient-profiling, or internally caused changes of state, are considered to be events that are slower to unfold than agent-profiling or externally caused changes of state. Punctuality differs for both forms, with the transitive class rated higher in both cases. This indicates that greater punctuality is attributed to verbs such as *shoot*, *assassinate*, and *poison* over verbs like *choke*, *suffocate*, and *starve*. Telicity was rated significantly higher for the transitive verb class for the active form only. This appears to stem from the fact that of the verbs in each class, all the transitive verbs are lexically telic with respect to the killing event except perhaps for *poison* and *shoot*, whereas all of the ergative verbs were lexically atelic with respect to the killing event except for *drown* and perhaps *suffocate*. The fact that telicity was significantly different for the active form only can be attributed to the fact that all the nominalized sentence forms were unambiguous with respect to the telicity of the killing event (more details about the verbs used and the stimulus data is found in the discussion of Experiment 2 in Section 2.4.2).

Nominalized Form Only	Ergative	Transitive	Significance
Kinesis (Movement)	1.63	2.796	F(1,214)=23.382, p < .001
Independent Existence	1.583	2.491	F(1,214)=12.977, p < .001
Punctuality	4.704	5.324	F(1,214)=9.665, p = .002

Table 5. Significantly different component ratings, nominalized kill verb stimuli only

Active Form Only	Ergative	Transitive	Significance
Volition	6.324	6.722	F(1,214)=9.010, p = .003
Sentience	6.259	6.685	F(1,214)=8.188, p = .005
Punctuality	4.769	5.583	F(1,214)=16.343, p < .001
Telicity	6.046	6.5	F(1,214)=7.213, p = .008

Table 6. Significantly different component ratings, active kill verb stimuli only

Overall, the significant differences for the ratings of the semantic components of transitivity observed in Experiment 1 support the claims of the theoretical work that distinguished the components. The ratings correlate with verb classes and sentence forms in the ways we would expect, and these results add further support to the literature providing evidence that the components are psychologically real.

2.4.2 Experiment 2

Within the broader investigational goal of linking semantic components to implicit sentiment, Experiment 2 was designed with two specific goals in mind. First, I wanted to establish that different surface event encodings for violent acts between opposing sides in a conflict reliably trigger different responses in comprehenders. It seems intuitive that whether an encoding explicitly refers to the perpetrator or not would make a difference in how the encoding portrays the event in the minds of comprehenders. The first goal of Experiment 2 was to establish this fact experimentally. (Recall that the method by which the perpetrator is mentioned in event encodings was at the heart of the claims in the study by HonestReporting.com.)

Second, In addition to showing that the encoding manipulations reliably alter the perceived sentiment toward the event, Experiment 2 shows that perceptions of implicit sentiment can be predicted by the semantic component ratings gathered in Experiment 1.

2.4.2.1 Stimuli and Procedure

Short newspaper-like paragraphs describing conflicts resulting in death were presented to participants. Specific circumstances of each scenario aside, for our purposes here we refer to the main participants in each event as the perpetrator and the victim. Alternate newspaper headlines were shown for the event paragraphs, and experimental subjects were asked to rate them with respect to how sympathetic they perceive them to be to the perpetrator and to the victim. Here sympathy, as a measure of favor or bias, is the particular sentiment examined (cf. the discussion in Section 1.2).

For the paragraph description of each conflict, the following criteria held:

- There was an obvious nominal referent for both the perpetrator and the victim
- It was clear that the victim dies
- In the scenario, the perpetrator was directly responsible for the resulting death, not indirectly (e.g. through negligence)
- Some paragraphs were based an actual news stories, and some were not. In either case:
    - No proper names (persons and places) were used, to avoid any inadvertent emotional reactions or legal issues
    - The description was not completely devoid of emotional impact. We wanted readers to have some emotional basis with which to judge the headlines.

The verbs of killing used were as follows, taken from Lemmens (1998), and based on his suggested reorganization of verbs of killing in Levin (1993) according to their degree of agent- or patient-profiling properties, as discussed in Section 2.3.

- kill
- slaughter
- assassinate
- shoot [action]
- poison [instrument]

- strangle
- smother
- choke
- drown
- suffocate
- starve

The verbs are listed above in two blocks, the first block ‘transitive’ and the second block ‘ergative’. The verbs are ordered from most ‘transitive’ or agent-profiling (kill, …) to most ‘ergative’ or patient-profiling (strangle, …) in the sense of Lemmens (1998). This corresponds to a continuum in which the prototypical meaning of the verb goes from most externally caused to most internally caused.

For each paragraph description, there are three alternate headline types:

- Transitive effective (causative)
- Transitive effective with nominalized agent
- Passive with no ‘by’ phrase

Each verb is used once per headline type, except the verb ‘kill’ which is used twice per headline type. This was done to balance the number of instances for transitive and ergative verbs, as there are six ergative verbs in the list above, but only five transitive ones.

The transitive effective (causative) constructions explicitly refer to the perpetrator and victim as the subject and object, respectively. The second is also a transitive effective, but with a nominalization as the subject. The third is a passive with no ‘by’ phrase included. An example is:

(1) Terrorists slaughter nine hostages
(2) Slaughter kills nine hostages
(3) Nine hostages are slaughtered

These forms were chosen in order to test the effects of explicitly naming both participants, or leaving the perpetrator unnamed by either nominalizing the act or using a passive construct with no perpetrator named. The complete set of headline stimulus items is listed in *Appendix 2 – Experimental Stimulus Headlines*.

As shown in (1)-(3), the nominal referent forms for the participants were used consistently among the items. Nominalized forms, as in (2), always use ‘kill’ as their main verb, as this is the prototype verb in the semantic field of killing. It provides the least information about how the killing was accomplished, and as the prototype, it has the largest and most varied distribution in the field. Conversely, headlines with nominalized perpetrators using the verb ‘kill’ require some other nominalization, so they don’t say ‘Killing kills victim’. For these two cases in the data, an appropriate nominalization drawn from the event description was used (e.g., ‘explosion’):

(1) Terrorists kill eight marketgoers
(2) Explosion kills eight marketgoers
(3) Eight marketgoers are killed

The stem of the chosen nominalization always appeared in the event description in either verbal or nominal form.

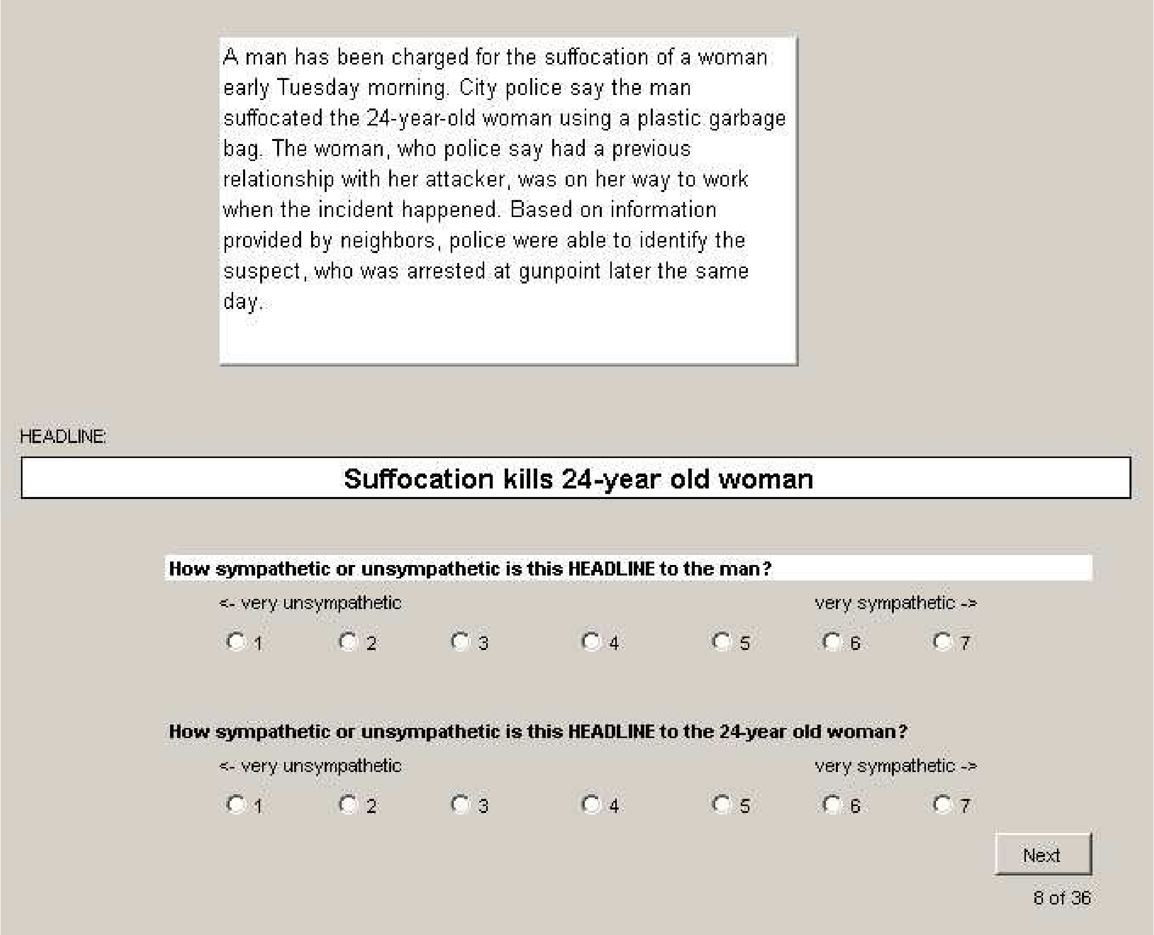
It was important to not make it obvious to participants in the experiment what was being manipulated. Thus distractor data were developed to be interleaved with the stimulus items. Distractor data adhered to the same criteria as described above for verbs of killing. The differences were:

- A conflict was at the core, however:
    - It did not result in death
    - It might involve violence, or it might be more abstract (e.g. litigation)
    - Non-causative verbs were used (e.g. *accuse*, *sue*, *deny*, *pressure*, etc.)
- There were not three alternate headlines, but rather one per distractor event description. Half the distractor headlines ranged a bit more freely in structure than the stimulus data (e.g., a prepositional phrase was added) in order to distract from the more rigid format of the stimuli, while the other half conformed to the format.
- The single headlines for the distractor items were evenly distributed among the three basic headline forms

The headline data were divided into three variants for the experiment, such that each participant saw data from all three conditions (headline forms), but each participant saw exactly one condition for each event description, and each participant saw the same number of headlines for each of the three conditions. Thus, in a sense, this design is between-subjects for specific stimulus items, but within-subjects overall for the three conditions to be tested. For the three conditions corresponding to the three headline types, the stimulus data were block randomized as in Experiment 1 such that no two items in sequence were of the same form (Keller, et al. 1998). There were 24 distractor items, also randomized, providing a two to one ratio of distractors to stimuli.

The headline data for form types (1) and (2) (i.e., not the passive condition) are the headlines that appeared in sentential form as the kill verb stimulus data in Experiment 1, as discussed in Section 2.4.1.1.

Using the WebExp framework, stimuli were presented in sequence, and two questions were asked about each story and headline pair. The first asked how sympathetic the headline was to the perpetrator, and the second asked how sympathetic the headline was to the victim. The question about the perpetrator is the one of interest here, as the manipulations were performed with respect to the perpetrator, and our hypothesis centers on sentiment toward the perpetrator. The question about the victim serves as a distractor. A sample stimulus item presentation from the experiment is shown in Figure 2. The interface emphasized the word HEADLINE in order to remind participants that the question is about the headline relative to the event, not the event itself or the event description.



A man has been charged for the suffocation of a woman early Tuesday morning. City police say the man suffocated the 24-year-old woman using a plastic garbage bag. The woman, who police say had a previous relationship with her attacker, was on her way to work when the incident happened. Based on information provided by neighbors, police were able to identify the suspect, who was arrested at gunpoint later the same day.

HEADLINE:

Suffocation kills 24-year old woman

How sympathetic or unsympathetic is this HEADLINE to the man?

<- very unsympathetic very sympathetic ->

○ 1 ○ 2 ○ 3 ○ 4 ○ 5 ○ 6 ○ 7

How sympathetic or unsympathetic is this HEADLINE to the 24-year old woman?

<- very unsympathetic very sympathetic ->

○ 1 ○ 2 ○ 3 ○ 4 ○ 5 ○ 6 ○ 7

Figure 2. Sample stimulus presentation from Experiment 2

2.4.2.2 Participants

A total of 31 participants completed the experiment. All were volunteers and were native speakers of English.

2.4.2.3 Analysis and Results: Effect of Surface Encoding on Perceived Sentiment

A mixed model ANOVA was run with the headline form as fixed effect. The overall effect of form was significant. The means, F ratio and p value are shown in Table 7. The pairwise differences between forms were tested for significance using Bonferroni correction. The Transitive Effective form was significantly lower in sympathy ratings against both the Nominalized Effective Form (p < .001) and the Passive Form (p < .001). No significant difference was found between the Nominalized Effective Form and the Passive Form.

	Transitive Effective Form	Nominalized Effective Form	Passive Form	Significance
Sympathy for Perpetrator	2.29	3.524	3.444	F(2,369) = 33.902, p < .001

Table 7. ANOVA results for effect of headline form on sympathy for perpetrator.

This result confirms the hypothesis that the form of the event encoding affects sentiment about the subject in the event encoding, and specifically that the most Transitive surface encoding predicts less perceived sympathy for the perpetrator. We also do not reject the null hypothesis that the Nominalized and Passive Forms will not differ with respect to attitude toward the subject.

A mixed model ANOVA was run with verb class as the fixed effect. Here again, verb class had a significant effect on attitude. The means, F ratio and P value are shown in Table 8.

	Ergative Verb Class	Transitive Verb Class	Significance
Sympathy for Perpetrator	3.258	2.914	F(1,370) = 5.430, p = .020

Table 8. ANOVA results for effect of verb class on sympathy for perpetrator.

I interpret this result to mean that participants were sensitive to the element of internal causation, or patient-profiling, in the ergative verbs. In the Nominalized and Passive Forms, the ergative verbs render the encodings ambiguous with respect to the very existence of a volitional or sentient agent. Thus on the whole these verbs are rated as more sympathetic toward the perpetrator than the transitive verbs. This result also confirms that participants were successful in evaluating the headlines directly, rather than basing their responses on the conflict vignettes.

2.4.2.4 Analysis and Results: Effect of Underlying Semantic Components on Perceived Sentiment

Having established that the different encoding forms had a significant effect on the sympathy ratings, the next step in the analysis was to investigate the central question of this chapter: Can the sympathy ratings gathered in Experiment 2 be predicted by the semantic component ratings gathered in Experiment 1? I developed several multiple linear regression models to address this question.

It was necessary to use two distinct participant pools for Experiments 1 and 2 because each experiment revealed information about the manipulations in the other which would be likely to alter the outcome if known to the participants. Because participant data could not be correlated for participants between the experiments, the regression models were run over the mean values of each observation in the experimental data.

The items of comparison in the regression are the 24 stimulus sentences that bridged both experiments. These are the 12 transitive effective kill verb sentences, and the 12 nominalized kill verb sentences. Note that the passive sentences were used only within Experiment 2 because many of the questions in Experiment 1 were inapplicable to passive sentences.

The general rule of thumb for multiple regression models is that between five and ten observed items are necessary for each independent variable in the model. With 24 observed items, we must limit ourselves here to three or four independent variables to serve as predictors of the dependent variable, which is the rating of sympathy for the perpetrator. Experiment 1 gathered data on 13 semantic components and we also have the fixed effects of verb class and individual verb, all as potential predictors. However, it is not predicted that all semantic components will be active simultaneously. Hopper and Thompson (1980) and Dowty (1991) explicitly describe their lists of attributes as a set of possible items, of which only some or perhaps even just one will be exhibited within any particular clause. Because the clauses and verbs used in the data were generally uniform, we would expect some consistency with respect to potential predictor variables. The key to this analysis, then, is to find the right set of predictors.

I first produced scatter plots for each semantic component predictor variable, and the dependent variable. These are shown in *Appendix 3 – Scatter Plots*. Based on which variables appeared to have the strongest relationships with the dependent variable, I ran several regression models with three to four predictors in an exploratory manner. While significant models with R^2 values in the range of .6 to .8 could be found rather easily, often not all predictor variables were significant within the models. In addition, considering the possible set of predictors from a logical perspective, it seemed that a model using Volition, Verb, and Telicity made sense. Volition is clearly important to the data in this study. Verb seemed likely to remain important because despite the semantic factoring, each distinct verb still contributes unique content. As one participant remarked, "'Strangle' just always says something nasty." Telicity seemed important because it reflects both lexical and grammatical factors. And indeed, these three variables produce a strongly predictive model as summarized in Figure 3.

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Durbin-Watson
1	.880(a)	.775	.741	.42195	1.896

a Predictors: (Constant), VERBITEM, VOLITION, TELICITY
b Dependent Variable: SYMPATHY

ANOVA

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	12.260	3	4.087	22.953	.000(a)
	Residual	3.561	20	.178		
	Total	15.821	23			

a Predictors: (Constant), VERBITEM, VOLITION, TELICITY
b Dependent Variable: SYMPATHY

Coefficients

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	.517	1.686		.307	.762
	VOLITION	-.200	.036	-.653	-5.594	.000
	TELICITY	.589	.257	.271	2.290	.033
	VERBITEM	-.105	.029	-.392	-3.606	.002

a Dependent Variable: SYMPATHY

Figure 3. Regression model with Volition, Telicity, and Verb as predictors

This model accounts for 77.5% of the variation in the sympathy ratings. The model itself is significant with $p < .001$. Each of the predictor variables is significant within the model. The adjusted R^2 value is often used in cases where the number of items is small relative to the number of predictor variables. It is reasonable to assume that is the case with the amount of data in this study, but even with the adjusted value, the model accounts for 74.1% of the variance in the dependent variable.

Because the scatterplots for Telicity and Punctuality are so similar, I ran another model substituting Punctuality for Telicity as a predictor variable. This model is summarized in Figure 4. This model is virtually identical to the previous model. It has a slightly higher R^2 value, and slightly lower p values for the individual predictors.

With Punctuality and Telicity apparently serving nearly identically as predictors, I sought to explore a bit further. Because Kinesis (Movement) had the most nearly identical behavior to Volition (aside from Sentience) based on its scatterplot and bivariate R^2 value, I next ran a stepwise regression model with Volition, Punctuality, Verb, and Kinesis (Movement) as predictors. The stepwise regression procedure is an algorithm that can converge on the optimal set of predictors. Interestingly, it settled on a model with Kinesis (Movement), Punctuality, and Verb as the best set of predictors, swapping out Volition in favor of Kinesis. The model is summarized in

Figure 5. This model accounts for around 80% of the variation in the dependent variable, with all predictors significant in the model.

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Durbin-Watson
1	.883(a)	.780	.747	.41750	1.933

a Predictors: (Constant), PUNCTUAL, VERBITEM, VOLITION
b Dependent Variable: SYMPATHY

ANOVA

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	12.334	3	4.111	23.588	.000(a)
	Residual	3.486	20	.174		
	Total	15.821	23			

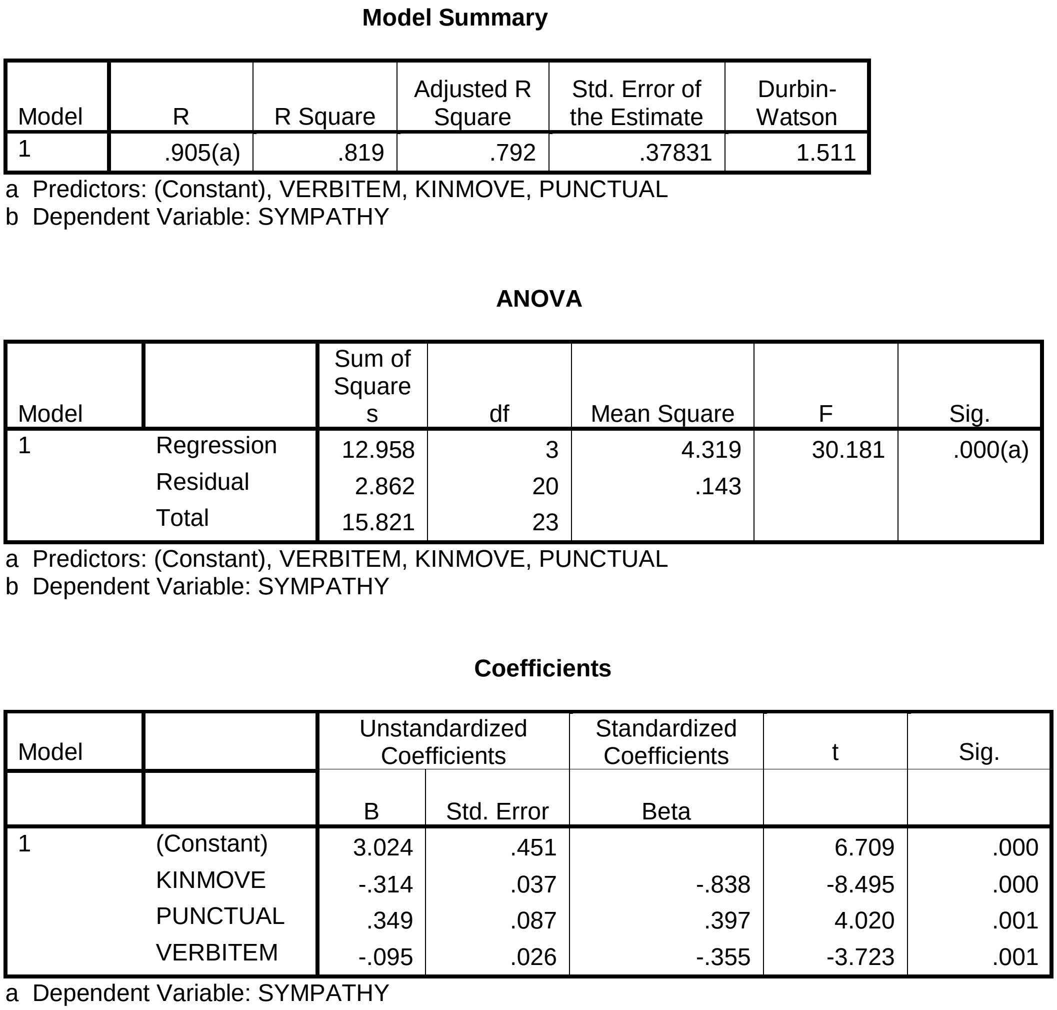
a Predictors: (Constant), PUNCTUAL, VERBITEM, VOLITION
b Dependent Variable: SYMPATHY

Coefficients

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	3.264	.501		6.510	.000
	VOLITION	-.241	.032	-.787	-7.462	.000
	VERBITEM	-.096	.028	-.357	-3.393	.003
	PUNCTUAL	.223	.093	.254	2.405	.026

a Dependent Variable: SYMPATHY

Figure 4. Regression model with Volition, Punctuality, and Verb as predictors



Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Durbin-Watson
1	.905(a)	.819	.792	.37831	1.511

a Predictors: (Constant), VERBITEM, KINMOVE, PUNCTUAL
b Dependent Variable: SYMPATHY

ANOVA

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	12.958	3	4.319	30.181	.000(a)
	Residual	2.862	20	.143		
	Total	15.821	23			

a Predictors: (Constant), VERBITEM, KINMOVE, PUNCTUAL
b Dependent Variable: SYMPATHY

Coefficients

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	3.024	.451		6.709	.000
	KINMOVE	-.314	.037	-.838	-8.495	.000
	PUNCTUAL	.349	.087	.397	4.020	.001
	VERBITEM	-.095	.026	-.355	-3.723	.001

a Dependent Variable: SYMPATHY

Figure 5. Regression model with Kinesis (Movement), Punctuality, and Verb as predictors

2.4.2.5 Analysis and Results: Empirical Correspondence with Lexical Semantic Analysis

Having explored these various models and having observed what appears to be the clustering of predictor variables, I ran a principal components analysis (PCA) to determine if in fact such clustering is operative, and if so, to what extent the clustering is consistent with the perspective of the lexical semantics literature that identified these variables, as discussed in Sections 2.2 and 2.3.

Two component analyses were run over all 13 semantic component variables along with the verb variable. First, before the analyses, the KMO measure and Bartlett’s test were run to ensure that the data is amenable to a principal components analysis. Results are shown in Table 9. Both tests determine whether the strengths of the correlations between the variables are large enough to proceed with a PCA. A KMO value closer to 1.0 indicates that a PCA is admissible for the data. For our variables here, the KMO value is 0.711, which is reasonably high. Also, the Bartlett’s test results, with $p < .001$, indicate that we can reject the hypothesis that the variables are not correlated enough to perform a PCA.

KMO and Bartlett's Test

Kaiser-Meyer-Olkin Measure of Sampling Adequacy.		.711
Bartlett's Test of Sphericity	Approx. Chi-Square	351.769
	df	91
	Sig.	.000

Table 9. KMO and Bartlett's Test for the set of all Predictor variables

Using both the Scree Plot method and the Kaiser criterion of retaining all factor components in a PCA for which the Eigenvalue is greater than 1.0 results in a PCA in which we retain four components. This factorization accounts for about 83% of the total variance among the variables. As the factor loadings in the component matrix (Table 10) and rotated component matrix (Table 11) make clear, the predictor variables cluster reasonably well along logical groupings, though the groupings are altered somewhat by the rotation.[13] The component matrix shows that component 1 comprises mostly the agent-oriented predictors. Component 2 comprises mostly the patient-oriented predictors. Component 3 joins Punctuality and Telicity into what might be usefully called the temporal predictors. Lastly, component 4 captures the verb itself.

The rotated component matrix moves Kinesis (Stationary) from component 1 to component 2, and No Independent Existence appears to shift from component 1 to component 4. In the former case, this shifts a patient-oriented variable to component 2 with the other patient-oriented variables. The case with No Independent Existence is less straightforward. Its factor loading value suggested it should stay in component 1, yet the statistical software sorted it as if it were in component 4. In any case, this variable did not exhibit significant differences over sentence form or kill verb class, and its factor loading values are quite low.

[13] The rotation of the component matrix attempts to make the components as orthogonal as possible.

Component Matrix

	Component			
	1	2	3	4
Cause Event	**.930**	-.031	.146	-.066
Sentience	**.925**	-.307	.133	.002
Volition	**.920**	-.326	.127	.001
Subject - Object Individuation	**.916**	.091	-.011	-.071
Independent Existence	**.903**	-.222	.107	.033
Kinesis (Movement)	**.871**	-.328	.274	-.022
Kinesis (Stationary)	**.639**	.360	-.471	.190
No Independent Existence	**.350**	-.274	-.264	.317
Causal Effect	.438	**.790**	-.088	-.264
Object Individuation	.453	**.695**	.007	.124
Cause Change of State	.579	**.618**	-.128	-.210
Punctuality	.056	.128	**.901**	-.061
Telicity	-.390	.450	**.673**	.140
Verb	.131	.244	.103	**.906**

Extraction Method: Principal Component Analysis.
a 4 components extracted.

Table 10. Factor loadings for each variable, by component (4).

Rotated Component Matrix

	Component			
	1	2	3	4
Volition	**.973**	.092	-.112	.034
Sentience	**.971**	.109	-.101	.038
Kinesis (Movement)	**.968**	.045	.037	.013
Independent Existence	**.912**	.173	-.102	.079
Cause Event	**.873**	.359	-.009	.016
Subject - Object Individuation	**.773**	.490	-.122	.022
Causal Effect	.062	**.934**	.085	-.101
Cause Change of State	.242	**.845**	-.026	-.066
Object Individuation	.127	**.775**	.116	.270
Kinesis (Stationary)	.298	**.647**	-.462	.274
Punctuality	.231	-.027	**.884**	.010
Telicity	-.353	.067	**.809**	.210
No Independent Existence	**.341**	-.087	-.403	**.285**
Verb	.023	.097	.072	**.945**

Extraction Method: Principal Component Analysis. Rotation Method: Varimax with Kaiser Normalization.
a Rotation converged in 5 iterations.

Table 11. Rotated factor loadings for each variable, by component (4)

Because the multiple regression models of the Experiment 2 data used three predictor variables, a PCA that retained only three components was also run. In this case, the three retained components accounted for about 75% of the total variance. The component matrix is shown in Table 12 and the rotated component matrix is shown in Table 13. Again we see the same general factorization into agent-, patient-, and temporally oriented clusters. Interestingly, the Verb variable is placed into component 2, along with the mostly patient-oriented variables. This could be reflective of the well-known fact from word sense disambiguation research that the object of a verb generally shares more information with the verb sense than does the subject (Olsen and Resnik 1997; Resnik 1997).

Component Matrix

	Component		
	1	2	3
Cause Event	**.930**	-.031	.146
Sentience	**.925**	-.307	.133
Volition	**.920**	-.326	.127
Subject - Object Individuation	**.916**	.091	-.011
Independent Existence	**.903**	-.222	.107
Kinesis (Movement)	**.871**	-.328	.274
Kinesis (Stationary)	**.639**	.360	-.471
No Independent Existence	**.350**	-.274	-.264
Causal Effect	.438	**.790**	-.088
Object Individuation	.453	**.695**	.007
Cause Change of State	.579	**.618**	-.128
Verb	.131	**.244**	.103
Punctuality	.056	.128	**.901**
Telicity	-.390	.450	**.673**

Extraction Method: Principal Component Analysis.
a 3 components extracted.

Table 12. Factor loadings for each variable, by component (3)

Rotated Component Matrix

	Component		
	1	2	3
Volition	**.972**	.103	-.110
Sentience	**.971**	.121	-.100
Kinesis (Movement)	**.967**	.058	.038
Independent Existence	**.911**	.190	-.101
Cause Event	**.868**	.365	-.021
Subject - Object Individuation	**.767**	.490	-.139
Causal Effect	.046	**.906**	.036
Cause Change of State	.229	**.823**	-.067
Object Individuation	.121	**.815**	.102
Kinesis (Stationary)	.298	**.673**	-.468
Verb	.044	**.259**	.136
Punctuality	.227	.004	**.883**
Telicity	-.353	.123	**.817**
No Independent Existence	**.351**	-.046	-.376

Extraction Method: Principal Component Analysis. Rotation Method: Varimax with Kaiser Normalization.

a Rotation converged in 4 iterations.

Table 13. Rotated factor loadings for each variable, by component (3)

The PCA results support the claim that the multiple regression models with three predictor variables defined above have validity as models of the correlation between the semantic component ratings and the sympathy (i.e. sentiment) rating. The factorings shown by the PCA results support the particular choices of predictor variables used. Moreover, the variables identified by the PCA show a close correspondence to linguistically motivated groupings of semantic properties, particularly Dowty's distinction between Proto-Agent and Proto-Patient properties.

2.5 Summary and Discussion

This chapter confirms the hypothesis that manipulation of event encodings in specific ways yields specific effects on the sentiment perceived by the readers of those encodings. Different encodings were shown to exhibit varying degrees of the semantic components of Transitivity. Furthermore, the regression models demonstrate the predictive power that these matters of degree have for the implicit sentiment attributed to those encodings.

These results formally establish that we can evaluate implicit sentiment on the basis of surface encodings, and that such evaluation can focus on observable features such as named agents, present or omitted objects, lexical choice of verb and nominalization, and voice. The underlying semantic components of Transitivity, with the role they play in the profiling of events and event participants, have been shown to track with these surface features. For example, I have shown that comprehenders are sensitive to the element of internal causation, or patient-profiling, in the ergative verbs, and that the semantic components of Transitivity are uniformly reduced in degree for such verbs.

In Chapters 3 and 4, I apply these results by building statistical classifier models that identify and exploit such surface features for the practical task of document level text classification with respect to (implicit) sentiment and perspective.

Beyond establishing the connection to perceived sentiment, the results in this chapter join those of Kako (2006a, 2006b), McKoon and MacFarland (2000, 2002) and Filip et al. (2001) in highlighting the value of exploring semantic components empirically. As discussed at the outset of this chapter, the behavior of verbs can be viewed as dependent not only on a verb’s meaning, but also on the construction in which it appears and the entities over which it predicates. Each of these elements can modulate the other in complex ways. What I have begun to examine in this chapter is how commonly identified semantic elements might characterize verb behavior. Dowty's theory demonstrates that certain semantic elements are associated with certain argument positions. McKoon and MacFarland (2000, 2002) and Wright (2001) produced distributional information that showed statistically significant differences in the types of entities used as arguments for ostensibly similar verbs. Hopper and Thompson's framework, while geared more toward clauses on the whole, shows how the common semantic components can vary in degree. I have tried to show here, in a preliminary way, that the kind of distributional information revealed by McKoon and MacFarland correlates with variations in the degree to which the common semantic elements are exhibited.

I have also begun to unify recurrent themes in work in lexical semantics and argument structure under a common psycholinguistic investigation. Linking theories in argument structure research (Dowty 1991, Grimshaw 1990), verb classifications in lexical semantics research (Levin 1993), and research on the generative lexicon (Pustejovsky 1991) and transitivity in grammar and discourse (Hopper and Thompson 1980) all make reference, in various guises and forms, to the semantic components measured here. The results reported here support the hypothesis that the semantic components are at work in sentences in ways that vary reliably along verb class lines.

While diathesis alternations represent a fascinating and important phenomenon, additional aspects of verb distributions are worthy of attention. Levin (1993) suggests that a syntactic alternate sometimes is accompanied by a change in meaning. Others have argued that a syntactic alternation always implies a change in meaning (Bolinger, 1968; Lemmens 1998). The change in meaning need not rise to the level of a change in word sense for the verb, though it can. But even when the verb sense remains constant, a syntactic alternation places a different construal on an event. The results in this chapter also suggest that different construals correlate with different values for the semantic components of Transitivity.

3 Linguistically Informed Identification of Implicit Sentiment: The Death Penalty Debate

3.1 Introduction and Overview

In Chapter 2, I established evidence for the hypothesis that in combination with its main verb, the degrees to which a clause exhibits (some subset) of the underlying semantic components of Transitivity comprise a strongly predictive model of perceived sentiment. This model was evaluated according to a standard criterion for psycholinguistic models: the semantic component ratings accounted for a large percentage of the variance in sentiment ratings. In this chapter, I experiment with the task of text classification for implicit sentiment. The challenge in this work is to render observable representations of the underlying (and hence unobservable) semantic components, and to use these observable representations to build classifier models. This effort will be evaluated according to a standard criterion for models in computational linguistics: they must make good predictions on previously unseen data.

This chapter is structured as follows. In Section 3.2 I discuss my approach to rendering observable representations of the underlying semantic components. I define such representations with respect to usages of terms identified as particularly relevant to a domain. I provide details on their definition and the method for instantiating them as document features (Observable Proxies for Underlying Semantics, or OPUS features). In Section 3.3 I introduce the Death Penalty Corpus (hereafter referred to as the DP Corpus). I summarize its motivation and preparation, and describe its dimensions and characteristics. In Section 3.4 I report on text classification experiments using OPUS features for verbs of killing, where the classification task is to determine if a document is written from the pro-death penalty perspective or the anti-death penalty perspective. I report baseline results and obtain improvements using OPUS features. Generalizing the approach, in Section 3.5 I describe a corpus-based technique for automatically extracting verbs for the death penalty domain. In Section 3.6 I report on further experiments with sentiment classification using OPUS features for the automatically identified verb set, where I obtain additional improvements in classification accuracy.

3.2 Defining OPUS Features

In order to exploit the predictive power that the underlying semantic components of Transitivity have for implicit sentiment, I use observable grammatical relations, drawn from the usages of terms determined to be relevant to a domain, as proxies for them. The current state of the art in dependency parsing allows us to extract reasonably accurate grammatical relations in unrestricted text. The relations, in conjunction with their targeted lexical content, comprise as proxies for the kinds of Transitivity phenomena that were shown to be active and predictive through the experimental collection of linguistic intuition data in Chapter 2.

As discussed in Chapter 2, verbs of killing have been a frequent target for in-depth linguistic analysis and are considered to be a set of prototypically causative verbs (Lemmens, 1998; Levin 1993). Chapter 2 investigated the family of kill verbs from a psycholinguistic perspective, and they were found to be good exemplars of the set of decompositional properties of transitivity identified by Hopper and Thompson (1980). Verbs are killing are prominent in a number of domains, for example the domain of violent crime reporting as seen in the experimental materials in Chapter 2. Additional examples would include wartime journalism, abortion, and the death penalty debate. As these properties have been shown to be linked to the realization of predicate argument structure, in this chapter I investigate the use of OPUS features instantiated over verbs of killing for the purpose of document level sentiment classification.

3.2.1 Feature Engineering

Producing OPUS features for terms requires syntactic as well as some semantic analysis. After evaluating several different parsers, I chose the Stanford Parser (Klein and Manning, 2003) for extracting document features based on grammatical relations in sentences.[14] Although nothing in principle restricts my approach to this parser, important practical factors in this decision were as follows:

- The parser performs at about the state of the art, and is reasonably fast and quite robust against degenerate input (Klein and Manning, 2003).
- The parser provides a medium-sized set of grammatical relations motivated by practical concern for real-world applications (de Marneffe et al., 2006). The set of relations is discussed in more detail later in this section.
- The parser is written in Java and is thus platform independent. It also provides an extensive Application Programmer Interface (API) enabling custom Java programs to parse corpora and extract information.

The Stanford Parser includes both an unlexicalized probabilistic context free grammar (PCFG) component, trained on the Penn Treebank (Marcus, 1993), as well as a lexicalized dependency grammar component (Klein and Manning, 2002). I use the PCFG component alone due to its speed and sufficient accuracy (Klein and Manning, 2003).

The parser's primary output is Penn Treebank-style constituent parses. As discussed in detail in de Marneffe et al. (2006), an integrated module within the parser extracts lexical dependencies between individual words, on the basis of the constituent parses. These are typed dependencies which indicate the grammatical relation between the words, for example `subject`, `object`, and `indirect object`. These typed dependencies are primarily generated using a set of manually developed pattern expressions in a constituent tree-oriented pattern-matching language called `tregex`

[14] See http://nlp.stanford.edu/software/lex-parser.shtml for more information and to download the parser.

(Levy and Andrew, 2006), quite similar to `tgrep2`.[15,16] The patterns are matched over the constituent phrase structure parses produced by the core parser, and the grammatical relations, or typed dependencies, are indicated by the matched patterns. The pattern-based grammatical relations module is tightly integrated with the constituent parser, and as such the system acts in effect like a batch mode version of the interactive Linguist's Search Engine (Resnik and Elkiss, 2005), with the built-in patterns operating as a library of established, typed queries.

The set of grammatical relations available from the Stanford Parser is based in part on those described in (Carroll et al., 1999) and (King et al., 2003). The relations are arranged hierarchically from general to specific, with the most specific relation chosen wherever possible. *Appendix 4 – Grammatical Relations of the Stanford Parser* shows the grammatical relations in their hierarchical context, along with descriptions and examples of each.

3.2.2 Document Feature Extraction

To extract OPUS features for a document, I generate a constituent parse for each of its sentences, from which the grammatical relations are then extracted. In general, each grammatical relation triple involving a relevant term can give rise to two features, one for relation-head and the other for relation-modifier. I describe specific features in the context of the relevant experiments.

The subject and object are the primary grammatical relations that should be expected to reflect usage patterns for what kinds of entities are being encoded as the agents and patients of actions. These relations then serve as proxies for the degrees to which the particular verbs and their arguments exhibit, for example, the components of Volition, Sentience, Kinesis, Punctuality, and Telicity, for which we saw significant differences with respect to subjects in Chapter 2. NP-internal relations can then propagate that information to the additional parts of compound nouns (in the nn-relation). Adverbial and adjectival modifiers typically represent intensifications or other modifications to those values. The negation modifier tends to invert those values.

In addition to the grammatical relations the parser provides, in this chapter I experimented with two additional features extracted based on a sentence's constituent tree and the built-in grammatical relations: transitive usages with overt NP subject and NP object (the TRANS feature), and sentences with no overt NP object (the NOOBJECT feature). For example, consider the example sentence in Figure 6, shown along with part of its constituent parse tree, a subset of its grammatical relations, and features yielded by those relations. The clause in bold would generate an instance of the feature TRANS-murder. The sentence in Figure 10 is an example of a sentence that would generate an instance of the feature NOOBJECT-kill.

[15] See http://tedlab.mit.edu/~dr/Tgrep2/ for more information.

[16] Some relations are identified through post-processing with procedural code that examines the constituents and pattern-based relations.

Sentence:

Life Without Parole does not eliminate the risk that **the prisoner will murder a guard, a visitor, or another inmate**.

Constituent Parse (excerpt):

```
  (S
    (NP (DT the) (NN prisoner))
    (VP (MD will)
      (VP (VB murder)
        (NP
          (NP (DT a) (NN guard))
          (, ,)
```

Grammatical relations (excerpt):

```
nsubj(murder, prisoner)
aux(murder, will)
dobj(murder, guard)
```

OPUS features (excerpt):

```
TRANS-murder
murder-nsubj
nsubj-prisoner
murder-aux
aux-will
murder-dobj
dobj-guard
```

Figure 6 – An example sentence with an instance of the TRANS-murder feature

At the same time, we should never ignore the risks of allowing the inmate to **kill again.**

OPUS features (excerpt):

```
NOOBJECT-kill
```

Figure 7 - An example sentence with an instance of the NOOBJECT-kill feature.

The TRANS-<verb> feature extracts canonical, syntactically transitive usages of relevant verbs where there is an NP subject and an NP object. Instances of this feature for verbs can stand as a proxy for the cluster of Transitivity component values that it carries, which will vary in the manner shown in Chapter 2. The NOOBJECT-<verb> feature, as in the example in Figure 7, often captures a habitual reading. This feature carries, in this example, reduced Object Individuation, and reduced Punctuality. The habitual reading invokes Dowty's proto-patient role of Incremental Theme, which I suggested in Table 2 might be mapped to Hopper and Thompson's Punctuality component.

The OPUS features in Figure 6 capture the fact that an event takes place with *prisoner* as the subject, representing a volitional action. Because OPUS features are defined with respect to targeted terms, the features reflect this volitional action with respect to events treated as particularly salient in the domain under consideration. Similarly, the syntactic transitivity exhibited by the `TRANS-murder` and `murder-dobj` features can correlate with high degrees of Volition, Agency, Kinesis, and Affectedness of the Object for the targeted *murder* event.

Although there are undoubtedly other ways to extract meaningful features that reflect underlying semantic properties, OPUS features extracted in this manner are intuitively plausible, avoid the data sparseness issues that would accompany the use of full relation triples, and, as I will shown in Sections 3.4 and 3.6, yield positive results in practical experimentation.

3.3 The DP Corpus

A natural domain in which to explore verbs of killing is the death penalty debate. The death penalty issue is a considerably polarized one that triggers passionate argumentation from both sides. As both the crimes and the punishments in this domain involve acts of killing, I predicted that the documents related to this debate would exhibit extensive use of kill verbs. I built a corpus of documents discussing the death penalty and found this prediction to be confirmed.

3.3.1 Corpus Preparation

I developed the DP Corpus to provide textual material representing the viewpoints of both the pro-death penalty and anti-death penalty movements. The DP Corpus consists of documents downloaded from web sites where it can be clearly determined that the sites identify themselves as being associated with one of these two polar sentiments, and that there are materials available in support of their position. I collected documents from five pro-death penalty sites and three anti-death penalty sites.

My working assumption was that it would be possible to create ground truth “pro” and “anti” (con) labels at the document level without requiring extensive human annotation, by associating document sentiment with the pro- or anti-death penalty sentiment identified by the document’s web site of origin. Based on partial human annotation, this assumption proved to be correct. *Appendix 5 – Development of the Death Penalty Corpus* provides extensive details on the preparation and annotation of the DP Corpus.

3.3.2 Characteristics of the DP Corpus

The DP Corpus consists of documents downloaded from five pro-death penalty sites and three anti-death penalty sites. Table 14 identifies the sites (with shorthand labels that will be used to refer to them), the number of documents in the corpus from each, and a brief description of the documents from each site.

Web Site (Shorthand label)	**Number of Documents in the DP Corpus**	**Description**
PRO:		
www.prodeathpenalty.com (PRO1)	117	A variety of document types including descriptions of particular death penalty cases, op-eds, journalistic pieces, fact sheets and briefs on particular cases.
www.clarkprosecutor.org (PRO2)	437	A few broad discussions of the role of the death penalty in criminal justice, and many documents describing the crimes and punishments in individual capital cases.
www.yesdeathpenalty.com (PRO3)	26	Primarily somewhat scholarly articles exploring specific arguments in support of the death penalty. Also contains a number of scientific and op-ed style pieces.
www.thenewamerican.com (PRO4)	7	Magazine op-ed style pieces.
www.dpinfo.com (PRO5)	9	Essay and op-ed style pieces.
	596 Total	
CON:		
www.deathpenaltyinfo.org (CON1)	319	Reports and fact sheets related to the death penalty. This site specifically claims to not offer opinions.
www.nodeathpenalty.org (CON2)	212	Journalism-like reports about the death penalty and particular death penalty cases.
www.deathpenalty.org (CON3)	65	Fact sheets and editorials on the death penalty.
	596 Total	

Table 14 – The DP Corpus

3.4 PRO/CON Sentiment Classification Experiments with the DP Corpus: The Kill Verbs

3.4.1 Document Features

I define a feature vector for each document in the corpus containing the following features:

- *Features based on kill verb grammatical relations.* The set of grammatical relations found from the constituent parse of each sentence in the document is extracted. For each relation in which a kill verb form appears as either the governor or dependent term, I extract a binary feature identified with the union of the verb and relation. All relations were eligible to be extracted in features. See Figure 6 for an example, which shows three relations for the verb *murder*.
- *Features based on kill verb nominalizations.* Grammatical relations in which instances of kill verb nominalizations occurred were extracted. Any relation to a kill verb was filtered out, as these were already accounted for with the kill verb features. Binary features for the nominalization were grouped together according to whether the nominalization was the governor or dependent in the relation, and thus were not articulated to the level of the individual relations as the verb features were (e.g. from "*The merciless killer felt no remorse,*" the features `nsubj-killer` and `amod-killer` would be group as two instances of the same feature).
- *Unigrams.* Unigram frequency features were included for the kill verbs and their nominalizations only.
- *Two Additional OPUS features*, as described in section 3.2.2: TRANS-<verb> and NOOBJECT-<verb>

Features were included for all occurrences of 14 kill verbs and their nominalizations. The verbs were: *kill, slaughter, assassinate, shoot, poison, strangle, smother, choke, drown, suffocate, starve, murder, execute, stab*. The nominalizations were: *killer, killing, slaughterer, slaughter, assassin, assassination, shooter, shooting, poisoner, poisoning, strangler, strangling, smotherer, smothering, choker, choking, drowner, drowning, suffocater, suffocation, starver, starvation, murderer, murder, executioner, execution, stabber, stabbing.*[17] The list of verbs includes all the kill verbs used in the experimental materials in Chapter 2, with the addition of *murder*, *execute,* and *stab*. *Murder* and *stab* are discussed briefly in Lemmens (1998). The verb *execute* can be ambiguous in other contexts, but is quite frequent and is primarily unambiguously used in its *kill* sense in the DP Corpus. This list of verbs and nominalizations generated 1016 distinct features.

[17] The –er nominalizations have been argued to exhibit strong indications of volition and agency (Lemmens 1998). Some of these nominalizations will of course not be attested in the corpus, but I included them for consistency across all the verbs.

3.4.2 Classification Results

Using the feature vectors thus defined, I built classifiers using three different algorithms: Naïve Bayes, C4.5 decision trees, and SVM, all implemented with the WEKA package of machine learning tools (Witten and Frank, 2005). SVM classifiers are my main target, as these typically have shown the best performance in text classification (Joachims, 1998). I included the other algorithms in these initial experiments largely for exploratory purposes. My primary interest is in comparing feature sets, not learning algorithms.

The task is to classify documents as being pro-death penalty (PRO) or anti-death penalty (CON). I conducted baseline classification experiments using only word n-grams as features. In preliminary experiments I tested both unigrams and bigrams, using both word forms and stems. The performance among these did not differ significantly, so I arbitrarily chose stemmed bigrams as the baseline. In order to control, in these experiments, for the difference in the number of features available to the classifier, I extracted the 1016 most frequent stemmed bigrams as the baseline feature set. Performance with these features was high, hovering around 90%. Evaluation was done using 10-fold cross-validation. All three algorithms achieved similar baseline performance (see Table 15).

Feature Set	**Naïve Bayes**	**C4.5**	**SVM**
baseline n-grams	91.69	91.19	93.37
OPUS	89.18	89.43	92.20

Table 15 - Sentiment classification accuracy of the DP Corpus, evaluated for three algorithms using 10-fold cross validation, in percent correct.

Classification performance based on OPUS was also high; again around 90% evaluated using 10-fold cross validation for the same three classification algorithms. Using the corrected resampled t-test as integrated into WEKA’s Experimenter tool, the results for the baseline and OPUS feature sets in Table 15 are not significantly different.

In this initial experiment, the OPUS features did not improve upon baseline, but it is notable that they achieved parity with baseline given that they are generated from a set of only 14 verbs and their nominalizations.

I next investigated a more realistic scenario: one in which test documents are not sampled (thanks to cross validation) from the same data sources as the training data. I segregated documents in the DP Corpus according to their web sites of origin, and conducted a series of experiments in which the training data and the test sets were from distinct sets of web sites. In these experiments, I found that the OPUS features significantly outperformed the baseline features.

Assignment of each site’s documents to training and test sets was driven by the number of documents each represented within the corpus (see Table 14). For the first experiment under this approach, there seemed a natural split of web sites into train and test sets because sites PRO3-PRO5 and CON3 have far fewer documents than the other sites. Thus these sites were set aside as test data, with documents from sites PRO1, PRO2, CON1 and CON2 serving as training data. A small number of documents were eliminated (arbitrarily) in order to balance the training and test sets across the two sentiment classes, resulting in a total of 1062 training documents and 84 test documents. Results for this experiment are shown in Table 16.

Using this web-site specific train-test split, overall classification performance dropped quite a bit, as one might expect when making the test data more distinct from the training data. But the OPUS feature set performed significantly better than baseline in all cases. Results are shown in Table 16. Here the t-test is not appropriate, as evaluation was not done using cross-validation. Instead, I use the Sign Test, a non-parametric matched pairs test similar to the Wilcoxon Matched-Pairs Signed-Ranks Test.[18] The baseline and OPUS classifiers are matched in pairs by document with respect to the error of their classification prediction for each document.[19]

Train-Test Split	Feature Set	Naïve Bayes	C4.5	SVM
PRO1+PRO2+ CON1+CON2 train PRO3+PRO5+ CON3 test	baseline n-grams	48.81	50.00	55.95
	OPUS	66.67	60.71	66.67
		$n_+ = 21$ $n_- = 6$, $p < 0.006$	$n_+ = 10$ $n_- = 1$ $p < 0.01$	$n_+ = 10$ $n_- = 1$ $p < 0.01$

Table 16 - Sentiment classification accuracy for the DP Corpus, evaluated for three algorithms using initial Web-site specific train-test splits, in percent correct.

I next performed a series of similar experiments where train and test splits were defined exhaustively using a two-by-two matrix of the four web sites with the greatest number of documents as seen in Table 14 (PRO1, PRO2, CON1, CON2) – in effect, a site-wise cross validation. The results were similar to those for the experiment summarized in Table 16 with a single web-site specific train-test split.

Table 17 shows the weighted average percent correct score as an aggregate of the four-fold test sets, where the weighting took into account the different numbers of documents in the test sets. I tested the significance of the weighted average percent correct by aggregating the matched pairs of the test items as input to the Sign Test. The results are shown in Table 17. Again, the classifiers using OPUS features score significantly higher than the baseline features. Table 18 breaks out the results by individual fold of the four-fold cross-validation.

[18] I used the implementation provided at http://www.fon.hum.uva.nl/Service/Statistics/Sign_Test.html. Note that the Sign Test is considered an insensitive test.

[19] Going forward, all significance tests reported will use the Sign Test in this manner unless specifically noted otherwise.

Feature Set	Naïve Bayes	C4.5	SVM
baseline n-grams	68.63	71.63	68.37
OPUS	82.42	77.84	82.09
	$n_+ = 295$ $n_- = 84$ $p < 0.0001$	$n_+ = 303$ $n_- = 208$ $p < 0.0001$	$n_+ = 362$ $n_- = 152$ $p < 0.0001$

Table 17 - Classification accuracy for weighted average percent correct across 4-fold train-test experiments

Train-Test Split	Feature Set	Naïve Bayes	C4.5	SVM
CON2+PRO1 train CON1+PRO2 test	baseline n-grams	57.31	49.29	39.15
	OPUS	92.22	75.24	77.83
		$n_+ = 178$ $n_- = 30$ $p < 0.0001$	$n_+ = 187$ $n_- = 77$ $p < 0.0001$	$n_+ = 234$ $n_- = 70$ $p < 0.0001$
CON1+PRO2 train CON2+PRO1 test	baseline n-grams	46.58	77.35	64.53
	OPUS	60.26	61.54	73.08
		$n_+ = 41$ $n_- = 9$ $p < 0.0001$	$n_+ = 17$ $n_- = 54$ $p < 0.0001$	$n_+ = 35$ $n_- = 15$ $p < 0.006$
CON2+PRO2 train CON1+PRO1 test	baseline n-grams	86.75	79.49	87.18
	OPUS	90.60	91.45	89.32
		n.s.	$n_+ = 30$ $n_- = 2$ $p < 0.0001$	n.s.
CON1+PRO1 train CON2+PRO2 test	baseline n-grams	77.59	81.50	82.29
	OPUS	81.03	80.56	85.58
		$n_+ = 62$ $n_- = 40$ $p < 0.04$	n.s.	n.s.

Table 18 - Sentiment classification accuracy for the DP Corpus, evaluated for three algorithms using two-by-two matrix of web-site specific train-test splits, in percent correct.

I next investigated the question of whether or not it is kill verb usages and the encoding of events they embody that is truly driving the improved classification accuracies obtained. In an experiment designed to address this question, I extracted the same set of OPUS features, but here for the 14 most frequent verbs found in the DP Corpus that were *not* in the list of kill verbs, along with their nominalizations. These verbs were: *say, sentence, appeal, find, tell, convict, take, see, hear, punish, testify, rob, give, deny*. This resulted in a classifier using 1518 distinct features. I compared classification using this feature set against the baseline classifiers, and found that they did not differ significantly from baseline performance, as the *kill* verb-based feature set had. The results are summarized in Table 19. This experiment establishes that it is not simply term frequency or the presence of particular grammatical relations that the kill-verb OPUS models were able to exploit. Rather the specific kinds of event encodings employing the strongly causative kill verbs are the effective features for our sentiment classification task.[20]

Train-Test Split	**Feature Set**	**Naïve Bayes**	**C4.5**	**SVM**
PRO1+PRO2+CON1+CON2 train PRO3+PRO5+CON3 test	baseline n-grams	48.81	50.00	55.95
	OPUS (Frequent *non-kill* verbs)	57.14	51.19	55.95
		n.s.	n.s.	n.s.

Table 19 - Classification accuracy based on frequent non-kill verbs from the DP Corpus, in percent correct.

The experimental results in this section confirmed that the two sides of the death penalty debate tend to use the kill verbs in ways that are different enough to be exploited by machine learning algorithms for sentiment classification with respect to that debate. While several kill verbs are among the most frequent in the corpus, the results in Table 19 confirm that it is not the case that simply looking at frequent verbs is necessarily useful. The results build on the findings in Chapter 2 by showing that the two sides in the debate appear to encode their discussions of killing events in a way that might, for example, attribute more Volition to criminals on the pro-death penalty side, and more Volition to the State on the anti-death penalty side. The DP Corpus contains narratives of violent crimes and of State executions, precisely where one might expect such manipulations to be exhibited. The use of OPUS features for machine learning demonstrates that it is not necessary to accomplish full natural language understanding in order to detect and exploit such patterns of differing usage between the two sides.

3.5 Identifying Domain Relevant Terms

Generalizing sentiment classification with the use of OPUS features requires that it be able to address domains for which we do not necessarily have a well-studied and motivated set of prototypically transitive verbs of physical force. To that end, in this section I apply a technique for automatically extracting sets of terms to be targeted for extraction of OPUS features for use in classification, in the manner that was done for

[20] It is possible that these results might be due in part simply to preferences for using different subsets of verbs for the two sides in the death penalty debate. I address this question in the experiments reported in Section 3.6.

the kill verbs. In Section 3.6 I report on the successful employment of such an automatically extracted set of verbs to our classification task with the DP Corpus.

I will use the description *domain relevant* for terms that are particularly characteristic of a corpus rather than the more common *domain specific*. The latter implies that the terms are technical and/or not found elsewhere to any degree. For example, the kill verbs are often invoked metaphorically and are not *specific* to the death penalty, but of course they are highly *relevant* to it, and thus to the DP Corpus.

Several methods exist for identifying significant collocations within a single corpus, for example the χ^2 test and log-likelihood ratios (Dunning, 1993). However, my concern is with the relevance of individual terms to a particular domain relative to the world at large; therefore I use the relative frequency ratio (Damerau, 1993) to determine what terms are domain relevant in a corpus, comparing against a large reference corpus of general text.[21]

The relative frequency ratio R_{rf} is defined in Equation 1, where

F_{dc} = the frequency of the term in the domain corpus
N_{dc} = the total number of tokens in the domain corpus
F_{rc}= the frequency of the term in the reference corpus
N_{rc} = the total number of tokens in the reference corpus

$$R_{rf} = \frac{\frac{F_{dc}}{N_{dc}}}{\frac{F_{rc}}{N_{rc}}}$$

Equation 1 - The Relative Frequency Ratio[22]

I use the British National Corpus (BNC) as my reference corpus. The BNC "is a 100 million word collection of samples of written and spoken language from a wide range of sources, designed to represent a wide cross-section of current British English, both spoken and written."[23,24] The BNC is thus advantageous because it is both very large as well as representative of text from a wide variety of domains and genres.

[21] Some corpus linguists use the term *monitor corpus* rather than *reference corpus* (Sinclair, 1991)
[22] Note that I have inverted the ratio from the way it is formulated in Manning and Shütze (1999, following Damerau, 1993). In this way, a "higher" number means "more domain relevant."
[23] See http://www.natcorp.ox.ac.uk/ for more information.
[24] I was originally concerned that spelling differences between British and American English would be an issue. However, in fact the BNC contains many instances of American English spellings. Normalizing between the two spellings and summing frequencies across variants would likely be advantageous, but for present purposes using the raw frequencies of the American spellings proved to be sufficient.

Conveniently, a number of researchers have made available term frequency tables derived from the BNC, and I have incorporated these rather than processing the BNC directly (Leech et al. 2001; Kilgariff, 1997; Pedler 2003). For my experiments I have used a table of frequencies for lemmatized, part-of-speech tagged terms for the entire BNC, with no minimum frequency cutoffs in effect.[25] For compatibility, the frequencies for domain corpus terms were also calculated by part-of-speech.

3.6 PRO/CON Sentiment Classification Experiments with the DP Corpus: Automatically Identified Verbs

Using the relative frequency ratio, I extracted all domain-relevant verbs for the DP Corpus from among all verbs that occurred a minimum of 800 times in the DP Corpus. There were 117 verbs identified as such, and these are shown in Figure 8, in order from highest to lowest R_{rf}.[26]

testify, convict, sentence, execute, aggravate, file, strangle, affirm, stab, schedule, rape, rob, violate, overturn, accord, murder, confess, pronounce, plead, shoot, kill, deny, arrest, condemn, commit, fire, witness, request, steal, review, appeal, decline, grant, rule, die, reject, state, impose, conclude, question, charge, beat, drive, attempt, release, admit, refuse, present, recommend, conduct, order, serve, receive, argue, determine, suffer, seek, issue, claim, note, discover, enter, fail, strike, find, identify, result, return, tell, include, indicate, arrive, sign, force, stop, say, pull, support, reveal, live, raise, ask, visit, drop, believe, hear, love, represent, regard, occur, hit, decide, express, involve, prove, stay, walk, consider, write, spend, end, place, fight, plan, face, base, continue, leave, call, hold, watch, allow, try, obtain, cause, begin, set

Figure 8 – Automatically derived domain-relevant verbs for the DP Corpus

Upon qualitative inspection, the set of verbs in Figure 8 appears to be satisfyingly representative of terms that are relevant to the death penalty domain. Note that it includes six of the 14 kill verbs used in the experiments in Section 3.4.[27] At the same time, the set introduces, among others, many verbs which have some senses in which they are transitive verbs of physical force: *rape, rob, steal, beat, strike, force, fight*.

I repeated the experiments summarized in Table 17, using the two-by-two web site based cross validation evaluation scenario, but this time employing OPUS features based on the 117 automatically identified verbs in Figure 8. Note that in these experiments, I used only the verbs, with no nouns or corresponding nominalizations. I compare the results against two baselines. First, extracting OPUS features for the 117

[25] See List 1.1 at http://www.comp.lancs.ac.uk/ucrel/bncfreq/flists.html.

[26] The relative frequency ratio does not admit of a significance test. As a general rule, we can interpret ratio values greater than one as indicating domain relevance, and values less than one as indicating the opposite, with the magnitude of the values indicating a matter of degree for those designations. In Chapter 4 I introduce the use of a threshold value ρ, used to control the number of domain relevant terms used in classification experiments; here all verbs with $R_{rf} > 1$ are included.

[27] The kill verbs not in this list are *slaughter, assassinate, poison, smother, choke, drown, suffocate, starve*.

verbs resulted in 7552 features, so I compare against a baseline of the top 7552 bigrams. I also compare against a baseline of unigram features for the 117 verbs themselves. I report results for SVM classifiers only. Table 20 shows the average pairwise percent correct for each classifier model. In all experiments, the OPUS features perform better than both baselines, significantly so in three cases (I report significance for each baseline relative to OPUS).

Feature Set	SVM
baseline bigrams	71.96[†]
baseline R_{rf} unigrams	84.51[‡]
OPUS	88.10
	$n_+ = 367$, $n_- = 149$, $p < 0.001$[†] $n_+ = 206$, $n_- = 151$, $p < 0.01$[‡]

Table 20 - Classification accuracy for weighted average percent correct across 4-fold train-test experiments

There are several notable facts about the results in Table 20. First, the accuracy achieved for the unigram-based classifiers is quite high, especially considering that these represent only 117 features. This result provides strong support for the idea of focusing on domain relevant terms. OPUS features, representing more detailed usage-driven articulation of those term features, increase accuracy even further. Second, and even more important, is that in three of the four train-test splits, the accuracies obtained for the terms derived automatically using R_{rf} represent improvements over the results obtained using the hand-picked kill verbs and nominalizations. This comparison is illustrated in Table 21.

Train Test Split	Feature Set	SVM
CON1/PRO1 train, CON2/PRO2 test	OPUS – kill verbs	85.58
	OPUS – R_{rf} verbs	89.66
		n.s.
CON2/PRO2 train, CON1/PRO1 test	OPUS – kill verbs	89.32
	OPUS – R_{rf} verbs	87.18
		n.s.
CON1/PRO2 train, CON2/PRO1 test	OPUS – kill verbs	73.08
	OPUS – R_{rf} verbs	78.63
		n.s.
CON2/PRO1 train, CON1/PRO2 test	OPUS – kill verbs	77.83
	OPUS – R_{rf} verbs	91.51
		$n_+ = 93$, $n_- = 35$, $p << 0.001$

Table 21 - Sentiment classification accuracy for the DP Corpus, evaluated for SVM classifiers built with the kill verb set, compared with the automatically derived verb set, using two-by-two matrix of web-site specific train-test splits, in percent correct.

3.7 Summary and Discussion

In this chapter I defined OPUS features and confirmed the hypothesis that OPUS features for kill verbs are highly effective for the classification of sentiment in the death penalty debate. By extracting observable features of event encodings related to killing, the machine learning classifiers were able to detect patterns of usage that were distinct for each side in the death penalty debate. These regularities were discernable with no attempt to target evaluative, opinionated, or subjective language over objective language. This result shows that the underlying semantic properties, as indicated by the results in Chapter 2, were likely to be operative in the language of this debate and that current tools in NLP can usefully exploit them. In section 3.6 I showed that we can obtain further improvement in classification accuracy for our task with a fully automatic end-to-end process completely exclusive of any manual selection or tuning. I obtain the best accuracies using OPUS features for automatically derived terms, in almost all cases beating unigram and bigram baselines as well as OPUS features for manually selected terms. Moreover, OPUS features based on domain-relevant terms outperform unigram features for the terms themselves.

The challenge I take up next is to determine if this method can be successfully employed in other domains. Many of the documents in the DP Corpus are not explicitly opinion pieces, and many contain quite dispassionate language. Nonetheless, an issue such as the death penalty does exhibit some uniformity in the lines of argumentation. In addition to particular capital cases that remain active topics of discussion for long periods of time, topics such as deterrence and recidivism are recurrent in the debate. Thus the possibility that a certain repetitive polarity exists in the debate must be considered. We cannot expect to find similar phenomena in many other domains. Chapter 4 necessarily takes up this question.

4 Linguistically Informed Identification of Implicit Sentiment: Extension to Additional Domains

A major test of my approach to text classification is to see how well it generalizes. Confidence in the approach increases if it can be successfully extended to additional domains. In this chapter, I report on experiments with two additional corpora. The first corpus is a collection of essays and commentaries related to the Israeli-Palestinian conflict, and the second corpus is a collection of United States Congressional floor debate speeches.

In the experiments discussed in this chapter, I modified my approach to exclude any use of unigrams and to apply a more focused use of linguistically informed features only. I also introduce a novel classifier combination method in my work with the Congressional data. For the classification tasks defined for each corpus, I achieve the highest accuracies yet reported.

4.1 Domain Extension 1: Bitter Lemons

4.1.1 The Bitter Lemons Corpus

The web site www.bitterlemons.org is the source of the Bitter Lemons corpus, hereafter referred to as the BL Corpus[28]. In its own words,

> *"Bitterlemons.org is a website that presents Israeli and Palestinian viewpoints on prominent issues of concern. It focuses on the Palestinian-Israeli conflict and peace process. It is produced, edited and partially written by Ghassan Khatib, a Palestinian, and Yossi Alpher, an Israeli. Its goal is to contribute to mutual understanding through the open exchange of ideas. Bitterlemons.org aspires to impact the way Palestinians, Israelis and others worldwide think about the Palestinian-Israeli conflict."*

The BL Corpus has a number of interesting properties. First, its topic area is one of significant interest and considerable controversy, yet the general tenor of the web site is one that eschews an overly shrill or extreme style of writing. This quality of the writing makes the BL Corpus a nearly ideal test bed for exploring the problem of automatically identifying the *point of view* from which the documents are written. In their work with the corpus, Lin et al. are to be credited with distinguishing the task and originating the phrase *identifying perspective*, which captures the idea of

[28] See http://perspective.informedia.cs.cmu.edu/demo/bitterlemons/data for more information and to download the corpus. The BL Corpus was prepared and distributed by Wei-Hao Lin and Theresa Wilson and made publicly available on January 25, 2007. The Bitter Lemons editors, Ghassan Khatib and Yossi Alpher, kindly agreed to make the data available for research purposes and their cooperation is here gratefully acknowledged. The BL Corpus is relatively new and thus to date the published research utilizing the corpus is limited to that of its developers, Wei-Hao Lin and colleagues.

identifying sentiment as point-of-view, a task distinct from distinguishing subjective from objective text, identifying overtly evaluative language or explicit statements of opinion or position, and/or using any of those elements to classify document-level sentiment. For my purposes, I consider this notion essentially the same as what I have described as identifying implicit sentiment.

Additionally, the structure of the Bitter Lemons web site is such that the BL Corpus has a very natural balance to it. The site is organized into weekly editions. Each week, the two editors of the site each write a piece on a designated issue, event, or topic within the greater Israeli-Palestinian conflict. Also in each weekly edition, two guest editors, one from each side, contribute a piece, sometimes in the form of an interview. Thus each week there are two pieces from the Palestinian perspective, and two pieces from the Israeli perspective. The corpus is therefore naturally balanced between the two sides and across the specific subtopics that are discussed. Also, following Lin et al. (2006), I take advantage of the natural split of editor and guest contributions to create training and test sets. Table 22 summarizes the contents of the BL Corpus as described by Lin et al. (2006).

	Palestinian	Israeli
Written by editors	148	149
Written by guests	149	148
Total number of documents	297	297
Average document length	740.4	816.1
Number of sentences	8963	9640

Table 22 - Descriptive Statistics for the BL Corpus (Lin et al. 2006)

Lin and Hauptmann (2006) describe a method for determining if two document *collections* are written from different perspectives. In their experiments, they test their method by comparing the BL Corpus against a corpus widely used in topically-oriented text classification, the Reuters-21578 corpus.[29] Documents in the Reuters corpus are annotated with respect to a list of 135 topics. Lin and Hauptmann’s method involves computing the Kullback-Leibler (KL) divergence (Kullback and Leibler 1951) between the statistical distributions of two corpora being compared, where in their formulation documents in the corpora are represented using bag-of-words feature vectors. They find that corpora written from two different perspectives will reliably have a KL-divergence in a middle range, whereas in contrast, corpora that are on two different topics have a high KL-divergence and corpora written on the same topic or from the same perspective have a low KL-divergence. These results show that, for corpora on the same topic, “the authors of different perspectives write or speak in a similar vocabulary, but with emphasis on different words.” Of particular note for my purposes, their method showed the Palestinian and Israeli halves of the BL Corpus to indeed exhibit differing perspectives.

[29] See http://www.ics.uci.edu/~kdd/databases/reuters21578/reuters21578.html for more information and to download the corpus.

4.1.2 Previous Classification Work Using the BL Corpus

Lin et al. (2006) report on experiments in document- and sentence-level text classification with respect to perspective using the BL corpus. The classification task is to determine if a document is written from a Palestinian perspective or an Israeli perspective. I will directly compare my results with their document-level results. Using unigram frequency feature vectors, they report results for a baseline SVM classifier and two Naïve Bayes classifiers. They used a linear kernel for their SVM classifier and optimized it by finding the best parameters using grid methods. Their first Naïve Bayes classifier (NB-M) uses maximum a posteriori estimation and the second (NB-B) uses full Bayesian inference. Their interest is in part to compare generative models like Naïve Bayes with discriminative models like SVM. I continue to use an SVM classifier, but compare my results to all of Lin et al.'s, as their best results are achieved using Naïve Bayes.

Lin et al. follow an evaluation approach similar to the one I employed for the DP Corpus in Chapter 3. Evaluation is conducted under the following two scenarios:

- Test Scenario 1: The classifier is trained on the guest documents, and the model is tested on the editor documents.
- Test Scenario 2: The classifier is trained on the editor documents, and the model is tested on the guest documents.

The results reported by Lin et al. under this evaluation framework are shown in Table 23. As can be seen there, using simple unigrams, Lin et al. achieve quite high accuracy.[30]

Training Set	Test Set	Classifier Model	Accuracy
Guests	Editors	SVM	88.22
Guests	Editors	NB-M	93.27
Guests	Editors	NB-B	93.46
Editors	Guests	SVM	81.48
Editors	Guests	NB-M	84.85
Editors	Guests	NB-B	85.85

Table 23 - Classification accuracy, in percent, for the classifiers of Lin et al. (2006)

4.1.3 Experiments with the BL Corpus using OPUS Features

As in the earlier experiments with the DP Corpus, I conducted classification experiments with the BL Corpus using OPUS features, motivated by the results of Chapter 2 that showed the impact that lexical semantics and event encoding choice can have on readers' perceptions of sentiment. In these experiments, feature vectors

[30] Lin et al. also report results using 10-fold cross-validation, but I focus here on the more interesting evaluation scenarios just described.

contained only OPUS features driven by automatically extracted lists of domain-relevant verbs and nouns.[31]

I introduced two new variations to the experimental setup. First, rather than always extracting OPUS features for all terms identified as domain relevant by the method described in Section 3.5, I introduced a threshold value ρ, which is used as a cutoff point for such terms. With a particular value of ρ in effect, an experiment only targets terms for which $\log(R_{rf}) \geq \rho$, where R_{rf} is the relative frequency ratio score for a term as defined in Equation 1 of Section 3.5. The higher the value of ρ, the more "relevant" the domain terms must be according to their relative frequency ratio scores. Note, then, that as ρ increases, the sizes of the term lists decrease. Table 24 shows selected values of ρ, the sizes of the corresponding term lists, and the top 30 terms from each list, for nouns and verbs from the BL Corpus.

Part of Speech	$\boldsymbol{\rho}$	**Total number of terms in domain-relevant list**	**Top 30 terms**
Verb	0.5	953	*fulfill, neutralize, pressure, given, favor, disengage, dismantle, maneuver, legitimize, reoccupy, preoccupy, mandate, acquiesce, annex, confiscate, pend, preempt, delegitimize, envision, reelect, democratize, negate, accord, honor, stabilize, reciprocate, practice, rebuff, forego, peacemake*
Verb	1.5	234	same as for ρ = .05
Verb	2.0	115	same as for ρ = .05
Noun	0.5	1823	*roadmap, disengagement, today, media, neighbor, intifada, incitement, favor, behavior, settler, statehood, outpost, terrorism, accordance, redeployment, reoccupation, dynamic, extremist, ceasefire, quo, coexistence, roadblock, latter, escalation, neighborhood, annexation, democratization, ramification, occupation, quartet*
Noun	1.5	478	same as for ρ = .05
Noun	2.0	276	same as for ρ = .05

Table 24 – Sample values of ρ, corresponding term list sizes, and example terms.

[31] In these experiments I used a new version of the Stanford Parser, released in summer 2006 (version 1.5.1, available at http://nlp.stanford.edu/software/StanfordParser-2006-06-11.tar.gz). This version featured some improvements to the grammatical relations output (e.g. improved tregex patterns for some relations). Also, the parser's API changed considerably from the previous release, rendering the Java code of my feature extraction client completely obsolete. Rather than investing heavily in additional client code that might be difficult to support in the future, I decided to use only the grammatical relations output as provided directly by the parser. Thus in all experiments going forward, I did not do the additional feature post-processing that I did earlier for the DP Corpus (e.g. extracting the TRANS feature).

Second, I introduced *filters* on the set of grammatical relations native to the parser's available output that were extracted when populating the document feature vectors. Rather than extracting all relations, I experimented with extracting only subsets of them, targeting for removal small sets of certain relations that (based largely on intuitive judgment) one would suspect to have little value with respect to my hypothesis. That is to say, the relations lack value in that they do not usefully reflect any of the underlying semantic components of transitivity investigated in Chapter 2 that I have hypothesized to be exploitable through grammatical relations. I report results for two filters:

1. DeterminerFilter. This filter extracted all grammatical relations from the parser except two:

   - det (determiner)
   - predet (predeterminer)

2. GeneralFilter: This filter extracted all grammatical relations from the parser except the following:
   - det (determiner)
   - predet (predeterminer)
   - preconj (preconjunct)
   - prt (phrasal verb particle)
   - aux (auxiliary verbs)
   - auxpas (passive auxiliary verbs)
   - cc (coordination)
   - punct (punctuation)
   - complm (complementizer)
   - mark (marker)
   - rel (relative)
   - ref (referent)
   - expl (expletive)

4.1.4 Results

For a broad range of values for *ρ*, I found that the OPUS features consistently beat the best reported results in Lin et al. (2006) for test scenario 1. Figure 9 shows the results with the DeterminerFilter in effect. In the graphs in Figure 9-Figure 14, the left-hand vertical axis indicates values of *ρ*, which are plotted as marks. Classifier accuracies are plotted as lines and are scaled to the right-hand vertical axis. Thus each tick mark along the horizontal axis represents an individual experiment, showing the *ρ* value in effect for verbs, the *ρ* value in effect for nouns, and the corresponding OPUS SVM classifier accuracy.

My results are shown in the data series labeled *OPUS*. Lin's best Naïve Bayes result (NB-B, 93.46%) and SVM result (88.22%) are shown for comparison. I varied *ρ* separately for the noun and verb term lists, and for all combinations, the OPUS feature-based classifiers perform better. The average accuracy across the 70 individual experiments represented in Figure 9 is 95.67%. The best experiment achieved 97.64% accuracy and is the highest reported yet for this task. This represents a 64% reduction in error against NB-B, and an 80% reduction in error against Lin's SVM.

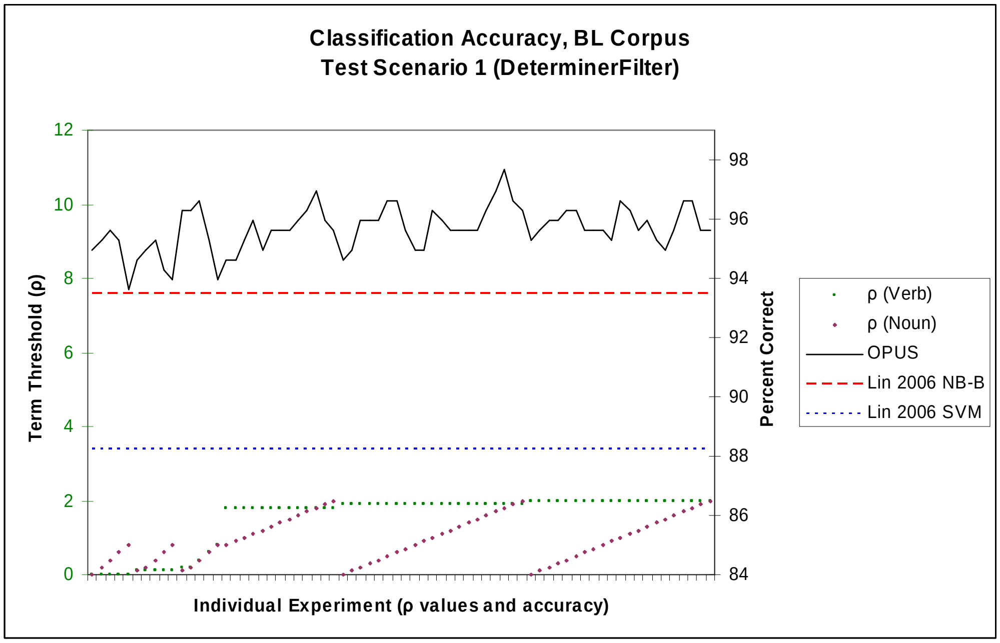


Figure 9 – SVM Classifier Accuracy using OPUS features for the BL Corpus, Test Scenario 1, with the DeterminerFilter applied

It is notable that I obtained these results using all default parameters to WEKA's SVM implementation, whereas Lin's SVM results were obtained with parameters optimized using grid methods (we both use linear kernels). This result holds up under a number of different conditions, which I turn to next.

Figure 10 shows the results of larger a set of experiments with the GeneralFilter applied, again compared to Lin's NB-B and SVM results. The average accuracy across the 423 individual experiments represented in Figure 10 is 95.41%. The best experiment, as in the results with the DeterminerFilter, achieved 97.64% accuracy (though here using different values of *ρ*). While eight of the experiments scored slightly below the accuracy of NB-B (results between 92.93% and 93.27%), overall the results are as robust as those in Figure 9, even though the classifiers built with the GeneralFilter use fewer features. This suggests that some of the features removed by the GeneralFilter were useful for classification, but the gains in accuracy can nonetheless be largely preserved after their removal.[32]

Turning to test scenario 2, the first thing to notice, as seen in Table 23, is that accuracy for all of Lin's models is uniformly lower than for test scenario 1. This is not terribly surprising: it is likely that training a classifier on the more uniform authorship of the editor documents builds a model that generalizes less well to the more diverse authorship of the guest documents (though accuracy is still quite high). Another likely factor is that the editor-authored documents comprise a smaller training set, consisting of 7,899 sentences, while the guest documents have a total of 11,033 sentences, a 28% difference.

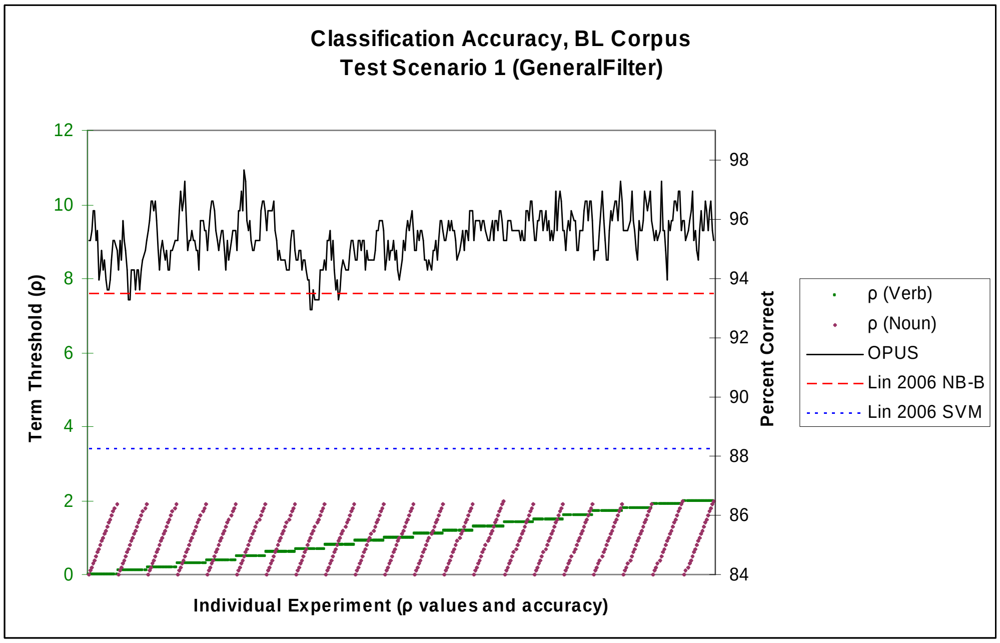


Figure 10 – SVM Classifier Accuracy using OPUS features for the BL Corpus, Test Scenario 1, with the GeneralFilter applied

My results for test scenario 2 exhibit the same pattern as Lin's, with my classifiers generally improving upon Lin's SVM accuracy and achieving about the same accuracy as Lin's Naïve Bayes models. Figure 11 summarizes the results with the DeterminerFilter. Average accuracy was 81.96%, and the maximum accuracy obtained was 83.84%. Results improved with the GeneralFilter, as shown in Figure 12. Average accuracy in this case was 83.12%, and the maximum accuracy obtained was 85.86%.

[32] This can be advantageous in situations where feature selection is critical, as in when memory and speed are at a premium given a task with sufficiently large feature vectors and/or large training sets (Brank et al. 2002).

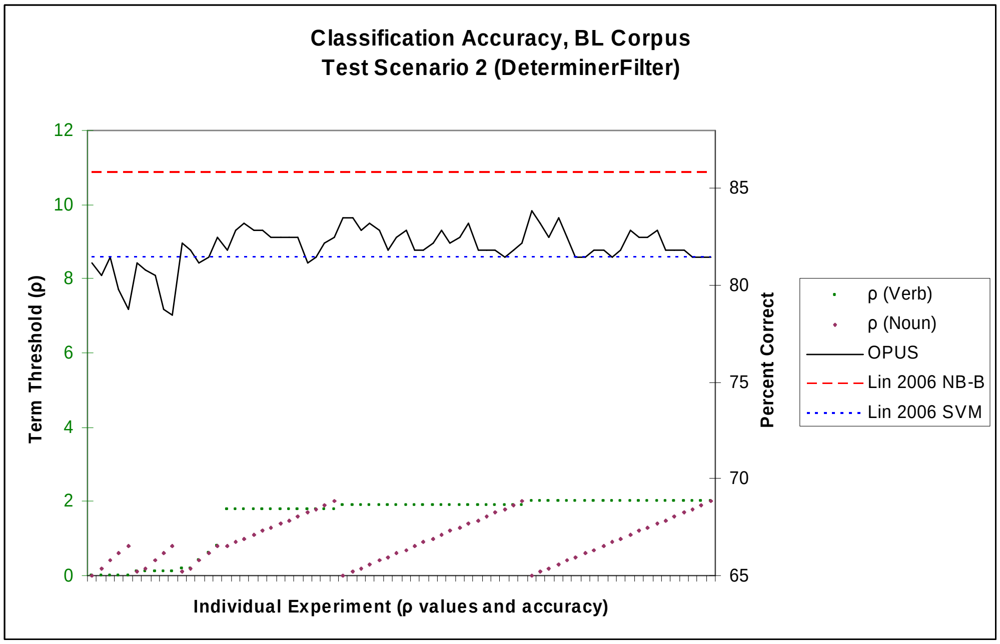


Figure 11 – SVM Classifier Accuracy using OPUS features for the BL Corpus, Test Scenario 2, with the DeterminerFilter applied

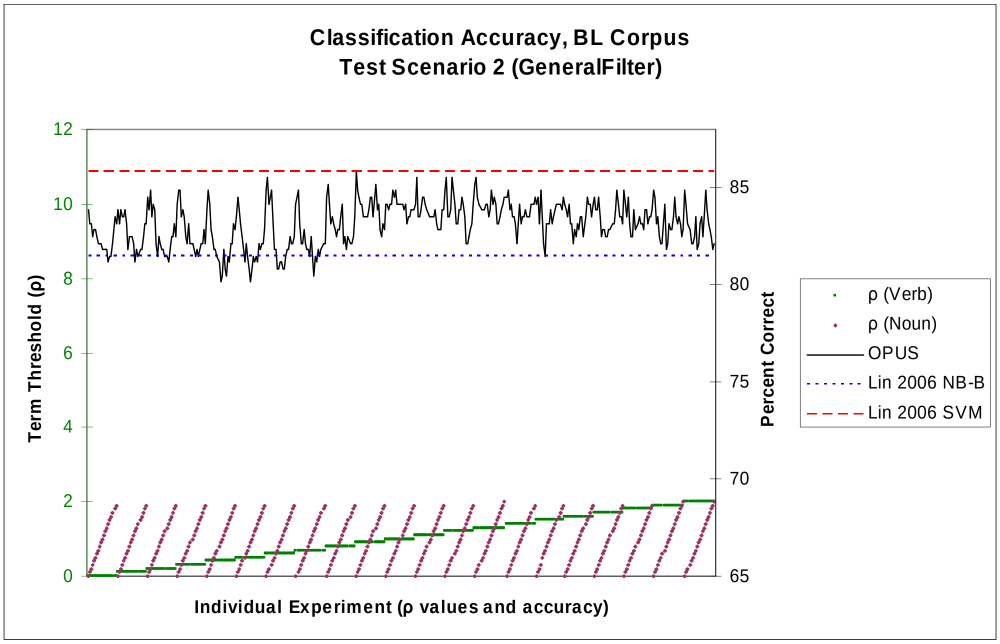


Figure 12 – SVM Classifier Accuracy using OPUS features for the BL Corpus, Test Scenario 2, with the GeneralFilter applied

4.1.5 Comparing Unigrams with OPUS Features

Given the high accuracy achieved with Lin’s unigram-based classifiers, a natural question to consider is how much the OPUS features actually contribute. Recall from Chapter 3 that each extracted grammatical relation results in two features, one for the governor term and one for the dependent term, tagged by the relation that joins them. Thus the set of terms that appear in the OPUS feature sets is the union of the domain-relevant terms automatically extracted from the corpus with all the terms that occur with them in the extracted relations. It is possible that it is merely the particular set of terms that results that leads to more accurate classifier models.

To examine the contribution of OPUS features, I compared the OPUS feature-based SVM classifiers against their corresponding unigram-based SVM classifiers. I will call the latter classifier type an OD-Unigram classifier (for OPUS-Derived Unigram). As a concrete example of what goes into an OD-Unigram feature set, consider the sentence *Sharon ordered the strike*. Suppose further that the verb *order* is in the list of domain-relevant verbs for an experiment, and that it is the only term in the sentence that occurs in either the verb or noun list of domain-relevant terms. The parser makes the following relations available for this sentence:

```
nsubj(order, Sharon)
det(strike, the)
dobj(order, strike)
```

The `det` relation is filtered out, and the `nsubj` and `dobj` relations are retained. This results in four features in an OPUS feature vector:

```
1. order-nsubj
2. nsubj-Sharon
3. order-dobj
4. dobj-strike
```

Thus there are three unigrams reflected in these features: Sharon, order, strike. The feature set of an OD-Unigram classifier is collected by gathering all such unigrams as are reflected in the OPUS feature set of a given experiment. I built OD-Unigram classifiers for the unigram lists corresponding to the same set of experiments described in section 4.1.3 (i.e. over the same ranges and combinations of values of ρ).

The OPUS classifiers perform better than the OD-Unigram classifiers in all but eight of the 423 experiments. The accuracy of the OD-Unigram classifiers is essentially flat at the level of Lin's NB-B classifier, averaging 93.49% (see Figure 13). Comparing classification accuracy across matched pairs of experiments, a Wilcoxon Matched-Pairs Signed-Ranks Test shows the overall difference in accuracy between the OPUS classifiers and the OD-Unigram classifiers to be highly significant ($W_+ = 86633$, $W_- = 103$, $N = 416$, $p << 0.001$).[33]

[33] I used the implementation available at http://www.fon.hum.uva.nl/Service/Statistics/Signed_Rank_Test.html

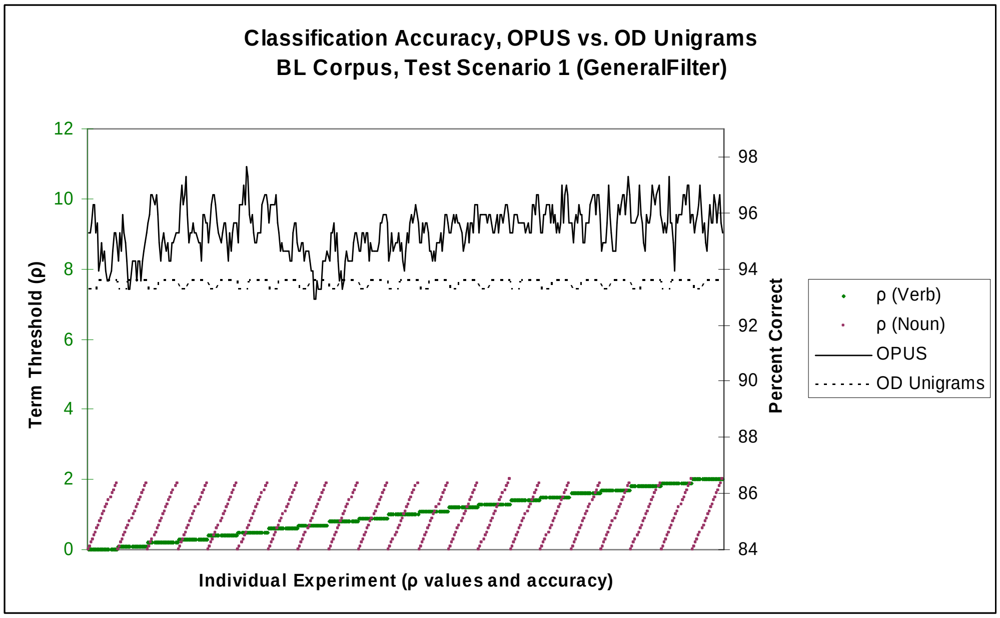


Figure 13 – Accuracy of OPUS SVM classifiers compared with the corresponding OD-Unigram SVM classifiers, Test Scenario 1

A similar pattern is seen in test scenario 2, though similar to the comparison of my results relative to Lin's, it is a bit less pronounced. The OD-Unigram classifiers have average accuracy of 81.75%, and are essentially flat at this level, which is quite near Lin's SVM classifier accuracy of 81.48% (the latter being based on the full vocabulary of the corpus). Over the same ranges for values of ρ, the OPUS classifiers perform better in all but 31 of the 423 experiments (see Figure 14). Comparing classification accuracy across matched pairs of experiments, a Wilcoxon Matched-Pairs Signed-Ranks Test shows the overall difference in accuracy between the OPUS classifiers and the OD-Unigram classifiers to be highly significant ($W_+ = 75104$, $W_- = 2711$, $N = 394$, $p << 0.001$).

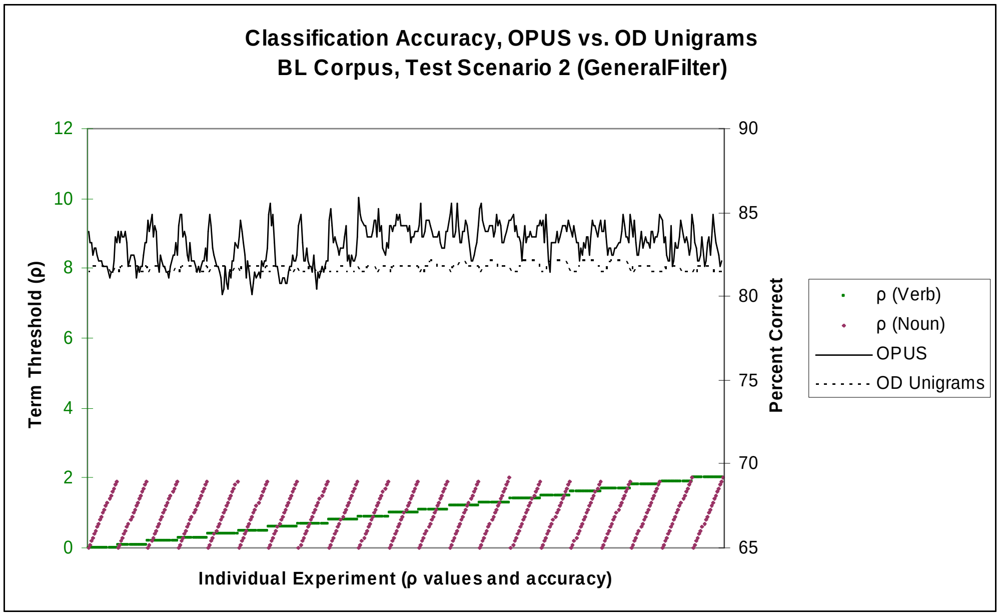


Figure 14 – Accuracy of OPUS SVM classifiers compared with the corresponding OD-Unigram SVM classifiers, Test Scenario 2

4.1.6 Analysis and Discussion

Lin et al. (2006) remark that “perspective is reflected in the words that are chosen by the writers.” They also note that in the BL Corpus, there is a great deal of overlap in the set of very frequent (non-stop) words used by authors from both perspectives, but that the two sides emphasize those words differently. In unigram models, emphasis equates to simple frequency. Lin highlights as a particular example that the word “palestinian” is used more frequently than the word “israel” in documents written from the Palestinian perspective, with the reverse being true for documents written from the Israeli perspective. The unigram-based classifier results that we have seen show that these facts can be learned and exploited by learning algorithms to excellent effect.

But note that it is not a given that the words “palestinian” and “israel” would have the distributions that they have here. The relatively muted nature of the documents in the BL Corpus means that its Palestinian authors refer to Israel by name. The phrase “Zionist entity” appears only once in the corpus, in a document of Palestinian perspective, but within the same document the author also uses the term “Israeli.” Extrapolating from the distribution of the words “palestinian” and “israel,” one might conjecture that the term “sharon” would appear more frequently in Israeli documents, and the term “arafat” more frequently in Palestinian documents. In the BL Corpus, the reality is quite different. Israeli documents mention both Sharon (51%) and Arafat (163%) more often than Palestinian documents. As found in Pang and Lee (2002), human intuitions about what words best emphasize a particular perspective can sometimes be highly mistaken.

I have shown that emphasis can be found at a source that we intuitively expect it to be found: in the actual usages of words, not just their mere presence. Authors from different perspectives discussing the same topic will encode the same events with different emphasis. The results in Chapter 2 showed that different event encodings will suggest different degrees of the underlying semantic components of transitivity. Components such as volition, punctuality, aspect, and kinesis, for example, can alter the implied ascriptions to event participants. Learning algorithms can effectively exploit the observable proxies for these semantic components in OPUS features.

The results summarized in Figure 13 and Figure 14 show that across a variety of vocabulary subsets, classifiers based on unigrams perform fairly uniformly. That level of performance is exceeded when OPUS features are used to distinguish different instances of those unigrams. In analyzing the better performance achieved by their Naïve Bayes models over their SVM model, Lin et al. (2006) speculate that it may be explained by the better performance of generative models over discriminative models when working with small training sets, as established by Ng and Jordan (2002). However, I nonetheless achieve sizeable improvements with SVM classifiers when using OPUS feature sets. The increase in accuracy is not uniform and thus appears to be highly sensitive to the particular sets of terms that are reflected in the feature sets. This pattern reflects what must be a wide variety of ways in which terms differ in usage between the two perspectives studied here.

The results of the BL Corpus experiments mark the second domain in which I have demonstrated the positive results of employing OPUS features. The marginal results for test scenario 2 suggest that I may have been brushing up against what might be the minimum size of a training set for my method to be effective. In section 4.2, I report results using a much larger and more topically diverse corpus.

4.2 Domain Extension 2: Congressional Speech

4.2.1 The Congressional Speech Corpus

In this section, I apply the use of OPUS features in experiments on a corpus of United States Congressional floor debate speeches, hereafter referred to as the CS corpus (for Congressional Speech). The CS corpus was developed by Lillian Lee and colleagues at Cornell University and generously made publicly available to the research community in December 2006.[34] They conducted document-level sentiment classification experiments with the corpus, including extensions of their maximum flow/minimum cut graph model approach (Pang and Lee 2004), and published their results in Thomas, Pang, and Lee (2006), hereafter referred to as TPL06.

The CS corpus has a number of properties that make it both attractive for and suitable to the current work. First, Congressional speech, and political discourse more

[34] See http://www.cs.cornell.edu/home/llee/data/convote.html for more information and to download the corpus.

generally, is a domain that is of interest to the general public that also generates intense interest and study across a wide variety of academic disciplines. Outside of natural language processing, for example, Congressional speech in particular has been studied in political science, economics, and journalism (Gentzkow and Shapiro 2006a, Groseclose and Milyo 2003, Quinn et al. 2006). The rise of political blogging further adds to the available volumes of electronic political text exhibiting perspective and sentiment. Automated or semi-automated analysis of such text for perspective and sentiment is thus potentially an application of considerable value.

In addition to being the subject of a good deal of study, Congressional transcript data is publicly available. The CS corpus and the Congressional speech data studied by Gentzkow and Shapiro (2006a) are both drawn from transcripts for the year 2005, and thus are likely to have at least some overlap. However, to my knowledge, the only published research to date that specifically utilizes the CS corpus is that of the corpus's originators. Given that, my experimental design and evaluation follow closely that of TPL06 in order to facilitate the most direct comparison possible.

TPL06 conducted classification experiments where the task was to determine if each individual speech segment (in the test set) represented text expressing support or opposition (a vote of YEA or NAY) to pending legislation. TPL06 used standard SVM-based classification techniques which they combined with a novel graph model classification framework that incorporates inter-speech segment relationships. In this section, I describe the corpus, summarize TPL06's methods and results, and report on my improvements to their results. These improvements were achieved using both an improved SVM classifier based on the OPUS features which are the focus of this dissertation, as well as a novel classifier combination method applied within TPL06's graph model framework.

4.2.2 The Corpus

This section summarizes the methods used to prepare the CS corpus as described by TPL06. The CS corpus was built from publicly available data extracted from the GovTrack website (http://govtrack.us). GovTrack is an independent website, run by Joshua Tauberer, which gathers together and makes available diverse publicly available information related to Congress, its members, and their legislative and fundraising activities, including voting records. GovTrack provides floor-debate speeches in a very convenient format, separated into distinct HTML files for each debate. TPL06 were thus able to relatively easily build the corpus as sets of speech segments that are gathered into specific debates.

Each "document" in the corpus consists of an individual, uninterrupted segment of speech at the podium on the floor of the U.S. House of Representatives. Each such document is referred to as a "speech segment" because there is a dialogue-like structure to debates among representatives on the floor. Speech segments are grouped together into debates, as indicated by the original transcripts. A sample speech segment is shown here:

> Mr. Speaker, I thank the gentleman from Massachusetts for yielding me this time. Because the Central American Free Trade Agreement can not pass on its merits, its supporters are attempting a last-minute bid to win desperately needed votes later this evening, probably very late this evening, on the Central American Free Trade Agreement. This bill before us purports to address the imbalanced trade relationship with china. We all know it will not do that. But what it is is just another cynical attempt to buy what is very well documented in this nation's pro-free trade, pro-CAFTA media, very well documented in the media; this is just another cynical attempt to buy votes on CAFTA, among other cynical attempts to buy votes on CAFTA. This fails, as the gentleman from Maryland (Mr. Cardin) said, as the gentleman from New Jersey members of Congress should be troubled that this bill has been introduced only in order to push through another trade priority. We should not have to approve a job-killing trade deal with Central America in order to get the chance to vote on a toothless China bill. I will say that again: we should not have to approve a job-killing trade deal with Central America in order to get a chance to vote on this toothless China bill. There are no assurances even that the Senate has plans to consider this half measure, and it is surely unlikely to ever become law. Aggressively counteracting China’s unfair trade practices should be a top trade priority. The gentleman from Michigan (Mr. Levin) and the gentleman from Maryland (Mr. Cardin), members of the committee on ways and means, they want it to be, but it should have nothing to do with CAFTA. Unfortunately, for the past 5 years, the administration has done nothing to curb China’s illegal trade activities. It is always words over action. In the past 5 years, our government has refused to enforce domestic trade laws with regard to China, failed to take advantage of WTO mechanisms to challenge China’s violations of international trade rules, balked at taking any concrete action on China 's manipulation of its currency; what I hear from my manufacturers in Akron , in Lorain , and in Elyria almost every week. Our government has proposed eliminating funding for China enforcement activities and our government's proposed Congressional efforts to address China 's unfair trade practices through legislation. This bill fails to resolve these problems. Instead of demanding action, it calls for more reports and more studies to tell us what we already know, that China is simply not playing fair. Congress may get only one chance, Mr. Speaker, to act on China trade this year. Wasting that opportunity on this ineffective bill is a betrayal of America’s working families, of our small manufacturers, and of our long-term economic security. Congress should not be fooled by this lose-lose proposition. A toothless bill on China will not make CAFTA any better.

TPL06 initially extracted all available transcripts of floor debates in the U.S. House of Representatives for the year 2005. They also collected the voting records for all roll-call votes during that year. TPL06 used this vote data in two main ways. First, the ground-truth labelings of speech segments were determined automatically by correlating the YEA or NAY label of a speech segment with the final vote of the speaker of the segment on the bill under debate. Secondly, TPL06 were able to focus the corpus on debates regarding “controversial” bills, which they defined operationally as debates in which the losing side generated at least 20% of the speeches. This was an important step to take because we are interested primarily in analyzing the language used in situations where the two sides in a debate have substantive differences, and a non-trivial number of floor speeches can be about tangential subjects or uncontroversial topics (e.g. recognizing Flag Day or similar).

TPL06 took several additional steps in the preparation of the CS corpus that are idiosyncratic to the nature of Congressional speech:

> *We automatically discarded those speech segments belonging to a class of formulaic, generally one-sentence utterances focused on the yielding of time on the house floor (for example, "Madam Speaker, I am pleased to yield 5 minutes to the gentleman from Massachusetts"), as such speech segments are clearly off-topic.*[35] *We also removed speech segments containing the term "amendment", since we found during initial inspection that these speeches generally reflect a speaker's opinion on an amendment, and this opinion may differ from the speaker's opinion on the underlying bill under discussion.*

TPL06 created training, test, and development (parameter-tuning) splits of the corpus by randomly selecting debates for each split. Each split represents about 70%, 20%, and 10% of the data, respectively. Table 25 provides a summary of the corpus. In another important step, TPL06 ensured that speech segments remained grouped by debate; they required that all speech segments from an individual debate appear in the same split of the corpus. I consider this requirement to serve two main purposes. First, the task in all these experiments is to classify the perspective of speech segments with respect to their expression of support or opposition to legislation, *not* with respect to the *topic* of the legislation. The task is binary pro/con classification irrespective of the topic. By keeping entire debates within a single split of the corpus, the focus is kept on sentiment classification because to at least some degree it has been ensured that speech segments on the same topic are not present in both the training set and the test set. This makes it less likely that features reflective of topic will be responsible for classification performance on the test set. Second, as noted by TPL06 and described below, the use of graph modeling within the classification framework developed by TPL06 is intended to take advantage of relationships between speech segments that are defined only within a debate. In support of direct comparisons, my experiments use exactly the corpus splits defined by TPL06.

	corpus total	training set	test set	development set
speech segments	3857	2740	860	257
debates	53	38	10	5
average number of speech segments	72.8	72.1	86.0	51.4
average number of speakers per debate	32.1	30.9	41.1	22.6

Table 25 - Descriptive Statistics for the Congressional Speech Corpus

The remainder of this chapter is organized as follows. First, I discuss the results of experiments with the CS corpus that build SVM classifiers using OPUS features, and compare my results to TPL06. I then present the graph minimum cut framework developed by TPL06. Following that, I report the results I achieve within that framework, again comparing to TPL06, and then extend it with a novel method of

[35] Note that many (potentially formulaic or off-topic) single sentence speech segments remain in the corpus, as they do not contain the term "yield," which was the simple criterion applied to automatically remove such speech segments. For example, one such speech segment consists of only the text *"Mr. Speaker, I demand a recorded vote."*

classifier combination. A discussion and summary of contributions concludes the chapter.

4.2.3 Initial SVM Classification of the CS Corpus

This section presents the methods and results of classification experiments with the CS corpus prior to the introduction of graph models with inter-document relationships.

I begin by reviewing previous results with the CS Corpus. TPL06 provide two simple baselines, which they improve upon with an SVM classifier. The first baseline is a majority baseline, where all items are classified according to whether the YEA or NAY votes were in the majority. The second baseline was intended to test if “the task can be reduced to simple lexical checks,” as stated by TPL06. In that baseline, they used the signed difference between the number of terms containing the stem “support” and the number of terms containing the stem “oppos” where the majority class was chosen in the event of a difference of zero. Their SVM classifier is based on a feature vector of simple unigrams, with no feature selection of any kind. The unigrams are not stemmed, stop words are not removed, and all tokens in the corpus are features. They used binary presence-of-feature values for each feature. The performance of these classifiers on the CS corpus test set is shown in Table 26. The unigram-based SVM, hereafter referred to as U-SVM, clearly outperforms the two baseline classifiers.

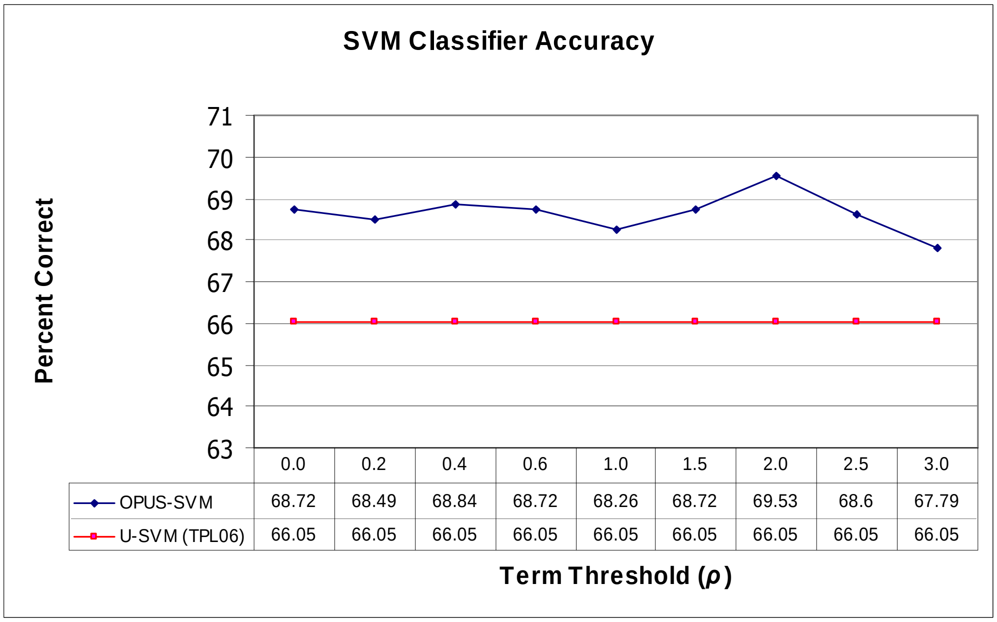
YEA or NAY classification of speech segments	Classification Accuracy
Majority baseline	58.37
“support” – “oppos”	62.67
U-SVM	66.05

Table 26 - Classification Accuracy for baseline and SVM classifiers from TPL06, in percent

I take the U-SVM classifier as my initial point of comparison. I ran experiments using SVM classifiers built in the same manner that I used for the BitterLemons corpus (Section 4.1).[36] In initial experiments, I used domain-relevant term lists that were extracted from all terms in the CS corpus, i.e. any term with a minimum frequency of one was eligible to be included if its relative frequency ratio was positive (see Section 3.x). The SVM classifiers were then trained on feature vectors populated by the OPUS features of the domain-relevant terms, extracted from the grammatical relations in which they occurred. I tested term threshold (ρ) values for the domain-

[36] There are some differences, though not in terms of how the features are defined. In all experiments in SVM classification with the CS corpus, I used the SVMLight (Joachims 1999) implementation of the learning algorithm rather than WEKA’s. This choice was necessitated for two reasons. First, TPL06 used SVMLight, and following their method, I used all default parameter values for SVMLight (which includes using linear kernels). Second, WEKA is unable to handle the memory requirements of a corpus this size, given the much larger set of training and test instances and the concomitantly larger feature vectors. Aside from this difference, the experimental method for the grammatical relation based SVM classifiers I build is the same as that described for the BL corpus.

relevant term lists that ranged between 0.0 and 3.0, with the same value used for both the verb and noun term lists. The first experiments used the DeterminerFilter as introduced in Section 4.1.3. For all values of ρ tested, the corresponding SVM models outperformed the U-SVM model. In Figure 15, the accuracy achieved in these experiments is graphed relative to U-SVM.[37]



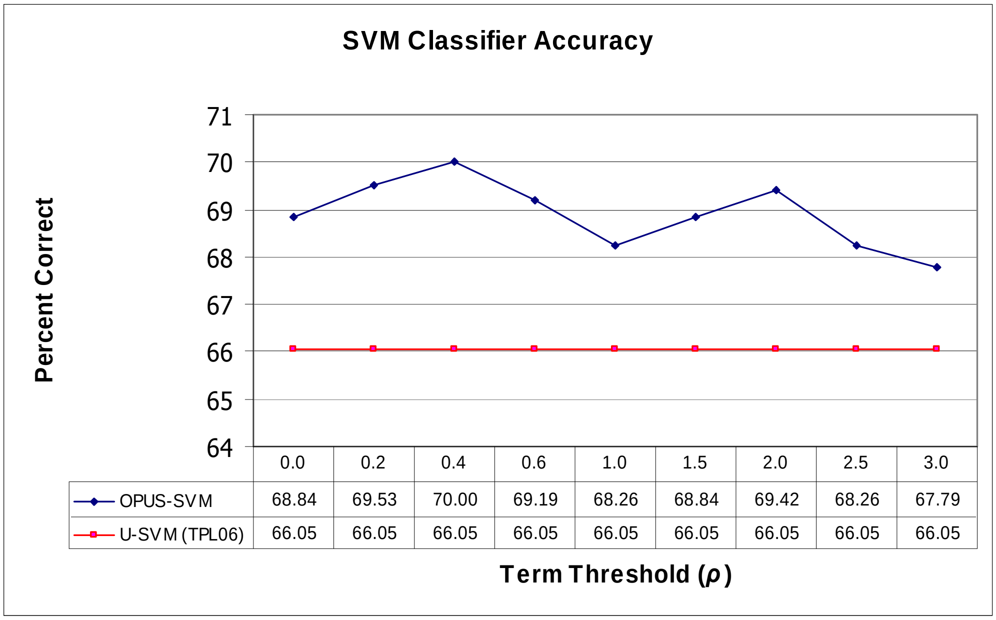
	0.0	0.2	0.4	0.6	1.0	1.5	2.0	2.5	3.0
OPUS-SVM	68.72	68.49	68.84	68.72	68.26	68.72	69.53	68.6	67.79
U-SVM (TPL06)	66.05	66.05	66.05	66.05	66.05	66.05	66.05	66.05	66.05

Figure 15 - YEA-NAY SVM classifier accuracy for various values of ρ, using the DeterminerFilter

The next set of experiments substituted the GeneralFilter for the DeterminerFilter. The results are summarized in Figure 16. For all values of ρ except two (ρ =2.0 and 2.5), these results are equal to or better than the results for the determiner relation filter, and thus again also outperformed the U-SVM model. The best result, at ρ = 0.4, achieved 70.00% accuracy and is a significant improvement ($n_+ = 104$, $n_- =138$, $p < 0.05$).[38]

[37] A Wilcoxon Matched-Pairs Signed-Ranks Test over the matched pairs of experiments shows the differences to be significant ($W_+ = 45$, $W_- = 0$, $N = 9$, $p < 0.01$).

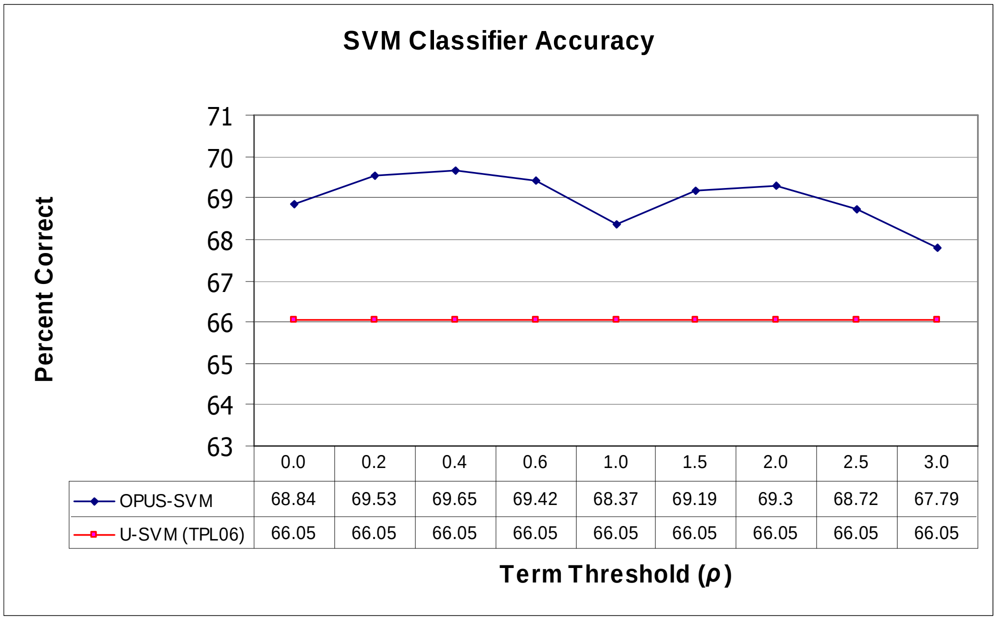
[38] A Wilcoxon Matched-Pairs Signed-Ranks Test over the matched pairs of experiments shows the differences to be significant ($W_+ = 45$, $W_- = 0$, $N = 9$, $p < 0.01$).



	0.0	0.2	0.4	0.6	1.0	1.5	2.0	2.5	3.0
OPUS-SVM	68.84	69.53	70.00	69.19	68.26	68.84	69.42	68.26	67.79
U-SVM (TPL06)	66.05	66.05	66.05	66.05	66.05	66.05	66.05	66.05	66.05

Figure 16 - YEA-NAY SVM classifier accuracy for various values of *ρ*, using the GeneralFilter

A third set of experiments used domain-relevant term lists extracted with the constraint that the terms have a minimum frequency of 25 in the CS corpus. Using the general relation filter, the results are shown in Figure 17.[39]



	0.0	0.2	0.4	0.6	1.0	1.5	2.0	2.5	3.0
OPUS-SVM	68.84	69.53	69.65	69.42	68.37	69.19	69.3	68.72	67.79
U-SVM (TPL06)	66.05	66.05	66.05	66.05	66.05	66.05	66.05	66.05	66.05

Figure 17 - YEA-NAY SVM classifier accuracy for various values of ρ, using the GeneralFilter and domain-relevant terms with a minimum frequency of 25.

Taken together, these results demonstrate that OPUS features consistently outperform unigram features in SVM classifiers for our task. This result is robust across various

[39] A Wilcoxon Matched-Pairs Signed-Ranks Test over the matched pairs of experiments shows the differences to be significant ($W_+ = 45$, $W_- = 0$, $N = 9$, $p < 0.01$).

parameter settings for ρ as well as two different corpus frequency threshold values. The results also primarily favor the use of the GeneralFilter, which sustains or improves accuracy relative to the DeterminerFilter, by removing grammatical relations that are not directly related to argument structure, adjuncts, or NP-internal structure (see the discussion in Section 4.1.3).

4.2.4 Modeling Inter-Speech Segment Relationships

A number of tasks in machine learning and natural language processing can benefit from the consideration of inter-item relationships, rather than considering each item independently. Semi-supervised methods in machine learning, for example, attempt to label unlabeled training items by considering various similarity measures between a labeled item and some number of unlabeled items. This is of considerable value given that it is often the case that labeled items are available in smaller numbers and can be expensive to produce while unlabeled items can be collected cheaply and in larger quantities (Blum and Chawla 2001).

The use of *minimum cuts* in graphs has been shown to be a natural and effective means of exploiting inter-item relationships. In this section we explore the use of graph minimum cuts as a method to further improve the results on our classification task with the CS corpus. The use of minimum cuts in graphs for various natural language processing tasks is an increasingly active area of research (Goldberg and Zhu 2006; Pang and Lee 2004; Thomas, Pang and Lee 2006). Here I summarize the method and describe the specific instantiation of it for our experimental task as originated by TPL06. I then introduce a novel extension to the method, a new classifier combination technique, which achieves the best results thus far on the task.

In my work with the CS corpus to this point, I have used only intra-item features within a straightforward SVM framework. However, TPL06 identify, based on the dialogue-like nature of the CS corpus, two particularly interesting inter-item relationships.[40] The first is *same-speaker* relations. In most cases, all speech segments by the same speaker within a debate can be reasonably expected to express consistent sentiment, YEA or NAY, within the debate. And, in fact, the method by which ground-truth labels were assigned to speech segments ensures that they are labeled as such, whether actually true or not. Second, in many speech segments, speakers make reference to other House members, often for the purpose of expressing agreement with their position on the bill under debate. Thus there exist *agreement* relations between items by different speakers.

TPL06 exploit these relations using the following classification framework. Let x_1, x_2,…, x_n be the set of speech segments within a debate. Let c represent the class, YEA or NAY, with which each item may be labeled. We define a non-negative function $ind_{svm}(x,c)$ which provides a score indicating the strength of preference that item x be classified as c. The function is so named because it is defined precisely in terms of

[40] Many other possible relations exist, including numerous measures of document similarity as studied in work on semi-supervised classification.

information obtained from the SVM models of the previous section, which are based on the features for *individual* items. Next, assume we have *weighted relations*, with wgt(r) indicating the weight of relation r between some pairs of items. A non-negative weight indicates the degree to which it is preferable that the two items within the relation receive the same classification label. These relations can represent the same-speaker and agreement links between speech segments as described above. Any class labeling C = c(x_i),c(x_2),…,c(x_n) for the segments x_i in a debate can then be assigned the *partition cost* defined by Equation 2.

$$\mathrm{cost(c)} = \sum_{x} ind_{svm}(x, \bar{c}(x)) + \sum_{x,x':c(x)\neq c(x')} \sum_{r\,between\,x,x'} wgt(r)$$

Equation 2 - Cost of class assignment for speech segments in a debate

Here, $\bar{c}$ *(x)* is the class opposite the class *c(x)* that is assigned to *x* by the given labeling. Over all items *x*, Equation 2 sums the individual score assigned to items for the class label not chosen, plus the sum of the weights of any relations *x* has to other items that end up not receiving the same label as *x*. Minimizing this partition cost thus represents the optimal way to label speech segments such that individual item scores can drive label assignments while strong relations between items work to keep those items from receiving different labels.

This optimization problem seems intractable, as there are 2^n possible binary partitions of the x_i's. If we represent the problem as a graph, however, we can take advantage of well developed graph-theoretic algorithms that provide a solution. We construct the graph as follows. The set of vertices in a graph *G* is defined as $\{v_1,v_2,\ldots,v_n,s,t\}$, where each v_i corresponds to a speech segment x_i in a debate. The nodes *s* and *t*, called the *source* and the *sink* respectively, correspond to the two classes of our binary classification task. For all v_i, two arcs are added to the graph connecting each v_i to the source and the sink. If we consider the source to represent class *c* and the sink to represent class $\bar{c}$, the weights on the arcs to the source are given the value of the function $\mathrm{ind}_{\mathrm{svm}}(x,c)$ and the weights on the arcs to the sink are given the value $\mathrm{ind}_{\mathrm{svm}}(x,\bar{c}\,)$. Arcs between the v_i, representing same-speaker and agreement arcs between speech segments, are weighted with the wgt(r) values. An example of such a graph is shown in Figure 18. Note that the wgt(r) arcs schematically represent both same-speaker and agreement arcs, and thus not all pairs of v_i nodes are connected by them.

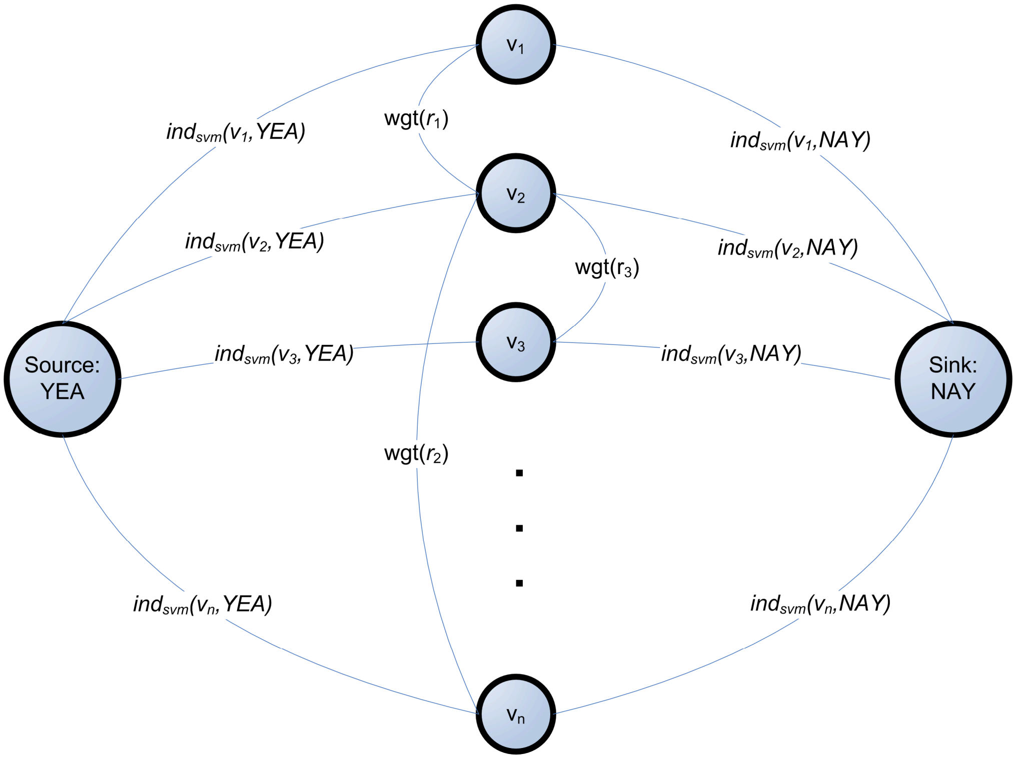


Figure 18 – The YEA-NAY classification task, including inter-item relations, modeled as a graph. The wgt(*r*) arcs represent same-speaker and between-speaker agreement links.

Pang and Lee (2004) provide a succinct definition of a graph minimum cut:

> *A* cut *(S, T) of graph G is a partition of its nodes into sets S = {s} ⋃ S′ and T = {t} ⋃ T′, where s ∉ S′, t ∉ T′. Its* cost *cost(S, T) is the sum of the weights of all arcs crossing from S to T. A minimum cut of G is one of minimum cost.*

A minimum cut of G is thus a partition of the graph into two disjoint subgraphs, one containing the source node *s* and a subset *S* of v_i nodes, and the other containing the sink node *t* and a subset *T* of v_i nodes. Cuts correspond to a partition of nodes than has a cost equal to the partition cost of Equation 2. The solution to the optimization problem thus reduces to finding minimum cuts.

Computing the minimum cut(s) of a graph like the one in Figure 18 is a well-studied problem and thus several algorithms exist for the corresponding *maximum flow* problem. This graph-theoretic problem has practical application to many network distribution problems, including transportation flow modeling and electronic network modeling. Given a graph of the sort in Figure 18, if we consider the weights assigned to each arc as a maximum capacity between its vertices, the solution to the maximum flow problem answers the question *What is the maximum flow from the source to the sink, subject to the capacities of the individual arcs across all paths from the source to the sink?* It has been shown that the maximum flow solution corresponds directly to the minimum cut, and thus algorithms that solve this problem can provide the set of arcs in the minimum cut. Typically, the solution to the maximum flow problem will have some arcs in the graph that ‘operate’ at less than capacity (i.e. whose weight is reduced in the solution). These changes in weights influence the minimum cut, and thus allow for the relabeling of items in our classification task.

Restating the definition a different way, the minimum cut can be thought of as the set of arcs that, if removed, would partition the graph in two (subject to the constraints described above), and for which the sum of the capacities of the removed arcs is a minimum. A useful (and real) analogy is to think of a wartime situation in which one side would like to disable its enemy’s supply distribution network at lowest cost. Given an appropriate model of the enemy’s network, the arcs of the minimum cut would indicate the optimal set of supply routes to cut.

At least four algorithms exist for the maximum flow problem. Chekuri (1997) provides experimental comparisons among some of them. We have used Andrew Goldberg’s HIPR program, an efficient implementation of the *push-relabel* algorithm (Cherkassky and Goldberg 1997). [41,42] These algorithms have been shown to compute an exact solution in polynomial time. In fact, as Pang and Lee (2004) note, running time can be near linear in practice.

In the public CS corpus distribution, along with the speech segments themselves, Lee et al. provide the graph arc weights and supporting spreadsheets used to calculate arc weights for the debate graphs. Their intention was to invite comparison with other experiments on this corpus, and the sharing of this data greatly facilitates this opportunity.

There are three types of weighed arcs in debate graphs: individual SVM score arcs, same-speaker arcs, and agreement arcs. Here we describe each in detail, including the precise method used by TPL06 to calculate the weights for each type.

1. The individual SVM scores [$\text{ind}_{\text{svm}}(x,C)$] are based on the signed distance $d(x)$ from the feature vector representing x to the decision plane of the SVM classifier model. If the SVM model produces a positive value for this distance for an item, it is classified as YEA. Negative values indicate a classification of NAY.

   Each distance value is normalized by dividing it by the standard deviation of all scores in the debate containing the speech segment. For each speech segment, the sum of the weights on the arcs to the source and the sink is always 10000. The specific weights are determined as follows:

[41] The C-language software, available free for research purposes, can be downloaded at http://www.avglab.com/andrew/soft.html.
[42] Our graphs are described in DIMACS format, the standard input format to HIPR as well as other implementations of maximum flow algorithms. See http://lpsolve.sourceforge.net/5.5/DIMACS_maxf.htm for a detailed description

- A speech segment with a normalized score at or below -2 is assigned an arc weight of 0 to the source and an arc weight of 10000 to the sink. That is to say, normalized scores that are heavily in the negative direction are fully weighted in the direction of the NAY class.
- A speech segment with a normalized score at or above +2 is assigned an arc weight of 10000 to the source and an arc weight of 0 to the sink. That is to say, normalized scores that are heavily in the positive direction are fully weighted in the direction of the YEA class.
- A speech segment with normalized score between -2 and +2 is assigned arc weights as follows:

  weight_of_arc_from_source = (normalized score + 2) * 2500.

  The weight of the arc to the sink is computed by subtracting the value above from 10000, i.e.

  $\text{ind}_{\text{svm}}(x, \bar{c}) \overset{def}{=} 10000 - \text{ind}_{\text{svm}}(x,c)$

2. Same-speaker arcs are assigned an effectively infinite weight. The specific value is not relevant, as long as it is high enough to ensure that a minimum cut of the graph will never separate two speech segments from the same speaker.

3. Agreement arcs are weighted based on a separate agreement classifier model built by TPL06. They consider only explicit by-name references to other speakers, which simplifies the task of building this classifier and ensures no error in identifying the existence of the references themselves. Then, to determine if the reference is an instance of agreement, they built an SVM classifier based on a presence-of-unigram feature vector derived from a text window 30 tokens before and 20 tokens after the reference, including the reference itself.[43] As an illustrative example, consider the speech segment excerpt shown in Figure 19. This excerpt contains numerous by-name references to other members of Congress, all of which are clearly instances of agreement between the speaker and the named members. The text window around the reference to Congressman Larsen is highlighted.

[43] They determined this text window size through tuning on the development set.

> It is not even clear we can move it to another bill at this point. Yet, it is the only bill standing, and it is a bipartisan effort to try to address this scourge that is crossing the country. I thank Chairman Sensenbrenner; also majority leader Roy Blunt, who has been an early leader in this charge; Chairman Barton of the energy and commerce committee for his willingness to have this. I would also thank the several members who have worked so hard to make this comprehensive anti-meth legislation happen. In particular, I would like to thank Representatives Mark Kennedy, Darlene Hooley of Oregon, Dave Reichert and John Peterson, because they provided much of the content of this comprehensive bill and their consistently strong leadership on the house floor. I would also like to thank the four co-chairmen of the congressional meth caucus, Congressmen Larsen, Calvert, Boswell and Cannon, for their staffs' assistance in putting this together so we could have a bipartisan effort. Congressman Tom Osborne has crusaded on this house floor and across the country on behalf of anti-meth legislation, as has Congressmen Baird, Wamp, Boozman, King, Gordon and so many others. This would not be happening today if we did not have this bipartisan coalition, and I hope it becomes law.

Figure 19 – A speech segment excerpt showing by-name references. The reference classifier text window around a reference to Congressman Larsen is highlighted.

Using a similar ground-truth labeling method as for the speech segments themselves, each reference in the training set is labeled as an instance of agreement if the two speakers voted the same way on the bill considered in the debate. A reference is otherwise labeled as a disagreement. The agreement classifier achieves accuracy of about 80%. Only references categorized as agreements (those with positive score) are used. We use the agreement weights provided by TPL06 without change, though we extend their experimental use of them.

Two free parameters are employed when calculating agreement weights. The first, θ_{agr}, is used in normalizing the scores according to the formula in Equation 3:

$$normalized\_score = \frac{raw\_svm\_score - \Theta_{agr}}{std.\ dev.\ of\ all\ reference\ scores\ in\ the\ debate}$$

Equation 3 – Normalization of agreement weight

The parameter θ_{agr} serves to vary the precision of the agreement references used. That is to say, as θ_{agr} increases, more agreement instances will fail to achieve a positive normalized score, meaning that only instances with higher confidence (higher raw SVM score) are retained. This can be valuable because false positives in the agreement arcs can draw a speech segment toward an erroneous change in label, whereas false negatives only leave agreement information unused. TLP06 report results for the values $\theta_{agr}= 0$ and $\theta_{agr} = \mu$, where μ = the mean agreement score assigned by the SVM agreement classifier to all references within a debate (resulting in the use of only above-average scores). We report additional results for $\theta_{agr} = 1.5\mu$, but do not explore additional values for θ_{agr} because at higher thresholds the number of positively valued reference weights shrinks substantially.

For references with a positive normalized score, the final arc weight is computed employing α, the second free parameter, as in Equation 4:

$$arc\_weight = normalized\_score \times 2500 \times \alpha$$

Equation 4 – Final arc weight for an agreement link

The parameter α serves to scale the weights, and is tuned on the corpus development set to a value that maximizes accuracy. Given the final arc weight value, vertices representing a single speech segment for each of the two speakers involved in the reference are joined with two directed arcs, one in each direction, both with the given arc weight. Which vertices are chosen is arbitrary, and only a single pair of vertices representing the two speakers are connected in this manner, because the infinite-weights of the same-speaker arcs propagate the agreement arcs to all vertices associated with the speakers.

To evaluate classification performance, a graph for each debate in the test set is constructed as just described. The minimum cut is computed, from which the final classification decisions are derived, and the resulting accuracy is computed. TPL06 report the results in Table 27, an augmented version of Table 26, and we add the final line showing accuracy with $\theta_{agr} = 1.5\mu$.

YEA or NAY classification of speech segments	Classification Accuracy
Majority baseline	58.37
“support” – “oppos”	62.67
SVM (unigrams)	66.05
SVM with same-speaker arcs	67.21
SVM with same-speaker arcs and agreement arcs, $\theta_{agr} = 0$	70.81
SVM with same-speaker arcs and agreement arcs, $\theta_{agr} = \mu$	70.81
SVM with same-speaker arcs and agreement arcs, $\theta_{agr} = 1.5\mu$	67.33

Table 27 - Classification Accuracy for baseline, SVM, and graph model classifiers, in percent

The CS corpus distribution provides all of the raw and derived SVM scores, as well as all of the raw and derived agreement scores needed to produce the results in rows 3 through 6 of Table 27. It does not provide the actual constructions, the sets of predictions made, nor certain other specifics (e.g. what value in practice achieves the infinite weighting of same-speaker arcs). Because I aim for the most direct comparison possible, in order to validate that my experimental setup faithfully adheres to TPL06, I ran the corpus distribution data for the experiments reported in rows 3 through 6 through my implementation of the experiment. I exactly reproduced the accuracy scores as reported by TPL06 to three decimal places. I can thus be confident that my implementation is a faithful instantiation of the experimental setup.

The results in Table 27 show that the addition of same-speaker arcs improves accuracy, and the addition of agreement arcs further improves accuracy. As the precision of the agreement arcs is increased, accuracy remains unchanged at $\theta_{agr} = \mu$ (though the set of predictions differs) and then drops for $\theta_{agr} = 1.5\mu$. We consider these facts more closely below in light of additional experiments. The increase from 66.05% to 67.21% with the addition of same-speaker arcs is not a significant difference. The improvement to 70.81% percent accuracy with the addition of agreement arcs over the SVM alone (66.05%) is significant ($n_+ = 155$, $n_- = 196$, $p < 0.05$) as is the improvement over the same-speaker results (67.21%; $n_+ = 49$, $n_- = 80$, $p < 0.01$).[44] The drop to 67.33% is also significant ($n_+ = 78$, $n_- = 48$, $p < 0.01$).

4.2.5 Joining OPUS Features with Inter-Speech Segment Relationships

Buoyed by the improvements in YEA-NAY classification presented in section 4.2.3, my expectation was that these improvements would carry through to graph minimum cut classifier models that integrate inter-speech segment relationships. A classifier that performs better in isolation would probably provide better individual classifier scores from which we derive the weights for the arcs connecting items to the source and sink.

To investigate this possibility, I executed the same experiment as summarized in Table 27, substituting the initial SVM classification scores of TPL06’s unigram feature vectors with my SVM scores from the best performing OPUS-based feature vectors. These results are shown in Table 28, repeating part of Table 27 for convenience.

[44] TPL06 do not report significance for these results. I determined significance based on my version of the experiments, again using the Sign Test.

YEA or NAY classification of speech segments	U-SVM	OPUS-SVM
SVM only	66.05	70.00*
SVM arcs plus same-speaker arcs	67.21	70.81*
SVM arcs plus same-speaker arcs and agreement arcs, $\theta_{agr} = 0$	70.81	68.37
SVM arcs plus same-speaker arcs and agreement arcs, $\theta_{agr} = \mu$	70.81	70.93
SVM arcs plus same-speaker arcs and agreement arcs, $\theta_{agr} = 1.5\mu$	67.33	70.12

Table 28 - Classification accuracy, comparing graph model classifiers using two different base SVM classifiers, in percent. Asterisks denote statistical significance between the U-SVM and OPUS-SVM classifier accuracies, ($p < 0.05$)

As noted earlier, our OPUS feature-based SVM model, OPUS-SVM, performs significantly better than U-SVM, and this remains true when adding same-speaker arcs to each model. We observe a different pattern when adding agreement arcs to OPUS-SVM. Considering the accuracies reported for OPUS-SVM in column 3 of Table 28 in isolation, they do not differ significantly from each other, except for the value 68.37% (at $\theta_{agr} = 0$) which is a difference that is highly significant ($n_+ = 29$, $n_- = 8$, $p < 0.001$). These results suggest that the same-speaker and agreement arcs are not significantly helping the better-performing OPUS-SVM. Moreover, the lowest precision set of agreement arcs actually damages performance significantly. As the precision of agreement arcs increases, the OPUS-SVM performs slightly better, and appears to maintain its performance while successfully incorporating limited, high-precision agreement information. U-SVM appears more dependent on agreement arcs for improvement, as evidenced by the dropoff where the highest threshold value $\theta_{agr} = 1.5\mu$ further restricts the agreement information available. If speakers who are truly in agreement on a debate have similarities in the way they speak in that debate, then the linguistically motivated OPUS-SVM model alone might have exploited some of that information, rendering redundant the (high-precision) agreement arc information.

4.2.6 Classifier Combination

Both U-SVM and OPUS-SVM show varying degrees of benefit from the addition of inter-speech segment relations, but overall the two models exhibit different patterns of performance. It thus may be advantageous to explore combinations of the two classifiers. Classifier combination has been a heavily researched problem, but the graph modeling paradigm employed here invites an approach which to my knowledge is novel in the min-cut framework: build graphs with multiple ind$_{\text{svm}}$ arcs from the source and to the sink for each v_i, one for each distinct SVM model.[45] This idea is illustrated in Figure 20 (cf. Figure 18). I will call this classifier combination method the *graph union* method.

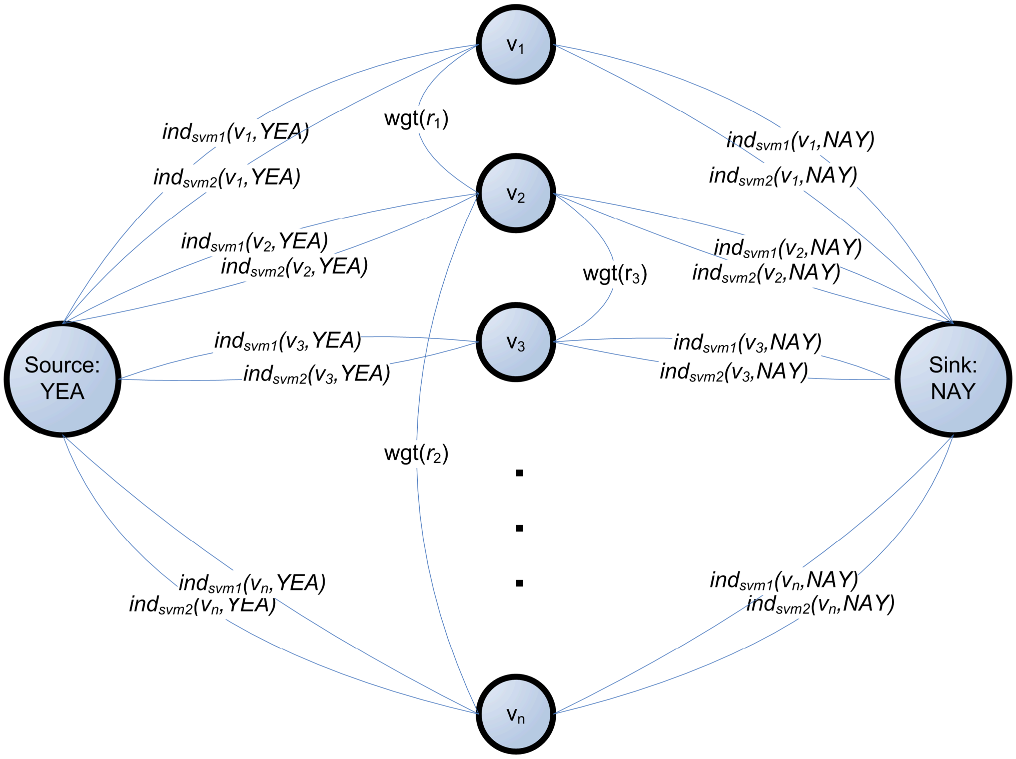


Figure 20 – The YEA-NAY classification task, including inter-item relations, modeled as a graph, using the graph union classifier combination method implemented with multiple ind$_{\text{svm}}$ arcs.

I ran experiments testing this approach. I began with a straightforward combination of the two SVM models, and then successively added the same-speaker arcs and then the agreement arcs with our three threshold values. The results are shown in Table 29.

[45] This idea is of course not restricted to SVM classifiers – an arc weight derived from any kind of classifier model could be added this way.

YEA or NAY classification of speech segments	**Graph Union of U-SVM and OPUS-SVM**
SVM arcs	70.35
SVM arcs plus same-speaker arcs	73.37*
SVM arcs plus same-speaker arcs and agreement arcs, $\theta_{agr} = 0$	73.49
SVM arcs plus same-speaker arcs and agreement arcs, $\theta_{agr} = \mu$	73.85
SVM arcs plus same-speaker arcs and agreement arcs, $\theta_{agr} = 1.5\mu$	74.19

Table 29 – Classification accuracy for the graph union of two SVM models, in percent. Asterisk indicates statistically significant difference.

The accuracy achieved with only the arcs associated with the two SVM models increases slightly to 70.35%. This does not differ significantly from the OPUS-SVM alone (70.00%, Table 28), but is improved over U-SVM alone at 66.05%, ($n_+ = 155$, $n_- = 192$, $p \leq 0.0533$). Adding same-speaker arcs, the result jumps significantly to 73.37% ($n_+ = 39$, $n_- = 65$, $p < 0.02$) and continues to increase as agreement arcs are added with increasing precision. The increased accuracies achieved with the agreement arcs are not significant amongst each other or to the same-speaker result. All of these results, however, remain significantly better than the graph union of just the arcs from the two SVM models, and significantly better than the corresponding results for each SVM model in isolation. The graph union method of classifier combination thus appears to be an effective way to take advantage of the individual strengths of each SVM classifier and the information available from same-speaker and (high-precision) agreement arcs. These results are the highest accuracy yet reported on this task.

4.2.7 Discussion and Contributions

In this chapter I extended the method described in Chapter 3, and applied OPUS features to classification tasks in two additional domains, using the BitterLemons and Congressional Speech corpora. In both cases I have demonstrated benefit from the deeper levels of linguistic processing, including the extraction of features reflective of argument structure, as motivated by the lexical semantics literature and the psycholinguistic results of Chapter 2.

Pang and Lee (2004), in introducing the minimum cut framework for sentiment classification, note that “it is perfectly legitimate to use knowledge-rich algorithms employing deep linguistic knowledge about sentiment indicators to derive the individual scores. And we could also simultaneously use knowledge lean methods to assign the association scores.” I have successfully taken up that challenge here.

Pang and Lee further explain that a major strength of their classification framework is the ability to combine intra-item and inter-item information in a principled fashion, rather than attempting to place inter-item information, perhaps in some synthetic form, within the kind of individual item feature vectors that serve as input to “traditional” classification algorithms such as SVM or Naïve Bayes. Furthermore, Blum and Chawla (2001) show that the arc weighting functions used in the graph minimum cut framework are of crucial importance, having substantial impact on the quality of the classification performance. I have shown here that improved individual classifier scores result in better arc-weighting functions. Additionally, the distinct feature sets that are the basis for these improved arc-weighting functions are best combined by building them into better individual classifiers and then combining them using the graph union approach. The graph union method performed better overall, and achieves its best performance when combining the best underlying SVM models. (See *Appendix 6 – Graph Union vs. Feature Union* for additional evidence supporting the graph union method).

Another result of my experiments with the CS corpus is that the addition of the kinds of inter-item relations considered here, same-speaker and (machine-learned) agreement links, do not uniformly lead to improvements, though the overall impression that they are helpful is maintained. Exploiting inter-item information should continue to be an interesting area of research, given that some common kinds of texts for which we would like to identify perspective and sentiment are precisely those that are likely to exhibit repeated authorship and inter-author agreement (and disagreement), such as forums, blogs, and other periodic writings by diarists and columnists. The rise of social networking sites such as MySpace and Facebook provide another potentially rich source of inter-author relationships that can be identified and exploited with similar or perhaps even better certainty than was possible with the Congressional speech corpus.

5 Summary of Contributions and Future Work

5.1 Summary of Contributions

The vast majority of work in sentiment analysis has been specifically targeted at the detection of subjective statements and the mining of opinions. In this dissertation I have addressed a different but related problem that to date has received relatively little attention in NLP research: detecting implicit sentiment, or spin, in text. Drawing on ideas from the literature in lexical semantics, I first established a relationship predictive of sentiment for components of meaning that are thought to be drivers of verbal argument selection and linking (Dowty 1991) and to be arbiters of what is foregrounded or backgrounded in discourse (Hopper and Thompson, 1980). I then used observable proxies for the expression of these meaning components, a set of linguistically motivated features, as input to machine learning methods for building statistical classifier models. This approach yielded models that performed significantly better than baseline models, and achieved the best performance yet published on the classification tasks I executed. I demonstrated the robustness of the approach by successfully applying it to three distinct text domains under a number of different experimental conditions. Sentiment analysis has been acknowledged as a more difficult problem than topic classification (Pang et al., 2002), and my approach establishes a new front on which to pursue this difficult problem.

Here I summarize the specific contributions of this dissertation, by chapter:

Chapter 2: The analysis of the experimental results in Section 2.4.2 confirmed the hypothesis that manipulation of event encodings in specific ways yields specific effects on the sentiment perceived by the readers of those encodings. Different encodings were shown to exhibit varying degrees of the semantic components of Transitivity. Drawing from the experimental data in Section 2.4.1, regression models demonstrated the predictive power that these matters of degree have for the implicit sentiment attributed to those encodings. These results established a link between elements of meaning and surface form, setting the stage for the classification experiments in Chapters 3 and 4. Additionally, the psycholinguistic results of Chapter 2 contribute additional evidence that the semantic components of Transitivity are psychologically real, and that they can provide insights into distributional facts such as the frequency with which certain words appear in certain constructions.

Chapter 3: The classification experiments in Chapter 3 successfully applied the results of Chapter 2 to a practical problem. In these experiments, I extracted observable features of event encodings related to killing according to the methods described in Section 3.2. I then used them to build classifier models in which machine learning algorithms were able to detect patterns of usage that were distinct for each side in the death penalty debate. These models outperformed baseline models. Notably, they did so without any explicit attempt to distinguish opinionated or subjective language from objective language. In section 3.6 I applied a technique for corpus relevant term identification, and demonstrated that I could obtain further improvement in classification accuracy with a fully automatic end-to-end process. I obtained the best accuracies using OPUS features for automatically derived terms, outperforming unigram and bigram baselines as well as OPUS features for manually selected terms.

Chapter 4: In this chapter I demonstrated the robustness of the method in Chapter 3 by applying it to corpora from two additional domains. In both cases I demonstrated benefit from the linguistically motivated features inspired by the results of Chapter 2. In Section 4.1.4 I reported experimental results that are the best yet achieved for the BL Corpus. In Section 4.2.3 I reported significantly improved classifier models for the CS corpus, prior to the introduction of same-speaker or agreement links. In 4.2.6, I introduced a novel classifier combination method, the graph union method, that successfully extended the graph minimum cut classification approach to achieve further significant improvements in classification for the CS Corpus, again achieving the best results yet reported. The graph union classifier combination method contributes a particularly promising vehicle for exploration into different methods of producing OPUS features and combining them with baseline features and inter-item relationships, which are commonly available in sentiment-oriented task settings.

5.2 Future Work

5.2.1 Extension to Additional Corpora and Domains

The three corpora that I have worked with here were each useful individually, and contrasted well amongst each other. The DP corpus was probably the most topically focused one, and had appreciable amounts of narrative text, mostly of crime events and court proceedings. Narrative is the type of text most precisely considered by Hopper and Thompson (1980). The BL Corpus was also reasonably topic focused, though it possibly exhibited more diversity within the domain of the Palestinian-Israeli conflict. While ongoing issues such as the occupation, the intifada, and the status of Jerusalem might elicit long term discussion, the BitterLemons web site is more focused on weekly events. Thus it seems less likely to show the discussion of long term debates in the way that the DP Corpus does for issues like deterrence and recidivism. The CS Corpus is no doubt the most topically diverse corpus, which likely explains why classification accuracies were the lowest for this corpus.

A research program further exploring and extending the generality of my method would proceed with further experimentation within the domains I have considered in this dissertation, as well as within entirely different domains.

Work with each of the corpora used here could be extended, beginning perhaps with the following:

- The DP Corpus was constructed from documents downloaded in 2005. Since that time, the death penalty issue has seen significant developments such as moratoria on executions and several important court decisions. Extending the corpus would provide interesting opportunities for additional experiments, perhaps testing models built on older text against test sets of newer text. Related to the DP Corpus, the death penalty memoranda prepared by Alberto Gonzalez, Jr. when he was legal counsel to Texas Governor George W. Bush have been made public and have become the topic of much discussion. There have been claims, for example, that the memoranda exhibit a systematic omission of certain kinds of facts. It would be interesting to experiment with these documents, perhaps using Lin and Hauptmann's (2006) document collection perspective divergence metric, to see if the memoranda as a collection can be measured as being more closely aligned with one side or the other in the DP Corpus.
- BitterLemons.org continues its weekly format and there is about two years' worth of additional material available since the time BL Corpus was constructed. Here too it would be interesting to test models built on older data against test sets of newer data. Additionally, the union of all the data would increase the size and therefore the interest of the corpus overall. Moreover, while the guest-authored half of the BL Corpus represents many different writers, there is a non-trivial number of guest authors who have made multiple weekly contributions. This fact fairly begs us to try to integrate the SVM classifier models built for the BL Corpus into a graph minimum cut framework with same-"speaker" links for these repeated guest author items.
- Congressional floor debate speeches of course never end, and the Congressional Record assures the continued public availability of debate transcripts. Thus the CS Corpus could be extended and used for additional experimentation. Additionally, the CS Corpus has items labeled by party affiliation, which could make for another potentially enlightening set of classification experiments. Moreover, following Gentzkow and Shapiro (2006a) and Quinn et al. (2006), it would be interesting to use my approach to pursue work similar to that done in political economy for tracking the spread of spin, sometimes referred to as the dissemination of talking points or "the echo-chamber effect." Another line of pursuit would be to build better agreement classifiers for defining agreement links, using my OPUS-feature based method.

Many possibilities exist for experimentation within additional domains. Among the set of different areas of potential interest for experimentation with respect to spin, some have corpora fairly readily available:

- Bill O'Reilly Transcripts, which have been analyzed with respect to O'Reilly's use of propaganda techniques (Conway et al., 2007)
- Transcripts from the defunct Crossfire show on CNN, which features two participants from each side, explicitly identified as having a left- or right-leaning perspective.

5.2.2 Methodological Extensions

The psycholinguistic investigation reported in Chapter 2 provides a basis for additional exploration into the relationships among variations in lexical and constructional meaning and sentiment. Numerous other constructions could be examined in psycholinguistic experimentation in order to determine their contribution to perceived sentiment when employed in particular event encodings. In addition to the constructions mentioned in Section 2.2, prepositional phrases and resultatives would likely be important constructions in the study of sentiment, given what seem to be their strong links to semantic components such as kinesis, change of state, and affectedness. Any clear results from such an investigation could inform improved relation filters, or perhaps relation weighting, building on the relation filters discussed in Section 4.1.3.

The degree to which the purely structural aspects of constructions contribute to the conveyance of perceived sentiment could be investigated by employing nonsense words as in Kako (2006a, 2006b). Context, perhaps using the vignette approach from Section 2.4.2.1, could be manipulated to modulate the interpretation of the nonsense verbs.

More generally, constructions could be directly employed as features, in the same vein as the TRANS feature in Section 3.2.2, rather than indirectly through bilexical grammatical relations. The machine learning framework used here would serve as the substrate for a constructions-level linguistic analysis of the link between surface encoding and sentiment. Data sparsity could be addressed in a number of ways, including smoothing with backoff methods based on verb class or WordNet synsets. Experiments extending my approach to sentence or passage-level classification would likely benefit from a feature space enhanced in these ways.

As mentioned in Section 1.1, machine learning methods can be quite effective, even in an error-filled feature space, if the errors are at least reasonably systematic. Nonetheless, it is likely that the linguistically-informed features that drove the document-level sentiment classification improvements reported here could produce further improvements if the features contained fewer errors. Experimentation with other tools, parsers for example, would be useful for comparing their relative performance in such tasks as my implicit sentiment classification task. A parser that could easily have its lexicon augmented would be able to better address domain relevant multi-word terms and named entities. Additional levels of processing, such as anaphora resolution, could enhance the richness of the feature data extracted.

The procedure for the identification of domain relevant terms could certainly be enhanced as well. Simple measures would include normalizing British and American spelling differences in the BNC data, as mentioned in Section 3.5, and implementing better smoothing for terms that do not occur in the reference corpus. Terms could be post-filtered relative to a dictionary or WordNet, to reduce noise. Finally, the relative frequency ratio itself, as the core criterion for domain relevance, could be replaced with a more sophisticated method such as the language model approach developed by Tomokiyo and Hurst (2003). Success in automatically identifying the terms for which linguistic features are extracted is all the more important for another endeavor: expanding my approach to additional languages.

Appendix 1 – Experimental Sentences for Experiment 1

High transitive frequency Externally Caused Change of State Verbs

The huge radiators dissipated the heat.
The scandal frayed party unity.
The sunset reddened the evening sky.
The cook thawed the holiday turkey.
The solution oxidized the scrap metal.
The weather stiffened his joints.

High transitive frequency Internally Caused Change of State Verbs

The intense sun blistered the paint.
Acid rain corroded the building.
The storm eroded the beach.
The yeast cultures fermented the beer.
The plants sprouted tender buds.
The violent storms swelled the sea.

Low transitive frequency Externally Caused Change of State Verbs

The police abated violent crime.
The stroke atrophied the right brain.
The fire alarm awoke the residents.
The French chef crumbled the cheese.
Religious extremists exploded the bomb.
The famous violinist vibrated the strings.

Low transitive frequency Internally Caused Change of State Verbs

The plants bloomed yellow blossoms.
The heavy traffic deteriorated the bridge.
The rare disease rotted the potatoes.
The constant rain rusted the car.
The regulations stagnated private investments.
The intense heat wilted the crowd.

Active-Form Transitive Kill Verb Sentences

Terrorists killed eight marketgoers.
The rebel killed the army officer.
Terrorists slaughtered nine hostages.
The city council member assassinated a rival candidate.

Gunmen shot the opposition leader.
The daughters poisoned the blind woman.

Active-Form Ergative Kill Verb Sentences

The man in custody strangled the deputy.
The woman smothered her grandmother.
The local man choked the vagrant.
The teenager drowned the young boy.
The man suffocated the 24-year old woman.
The caretaker starved the severely ill woman.

Nominalized-Form Transitive Kill Verb Sentences

The explosion killed eight marketgoers.
The attack killed the army officer.
The slaughter killed nine hostages.
The assassination plot killed a rival candidate.
The shooting killed the opposition leader.
The poisoning killed the blind woman.

Nominalized-Form Ergative Kill Verb Sentences

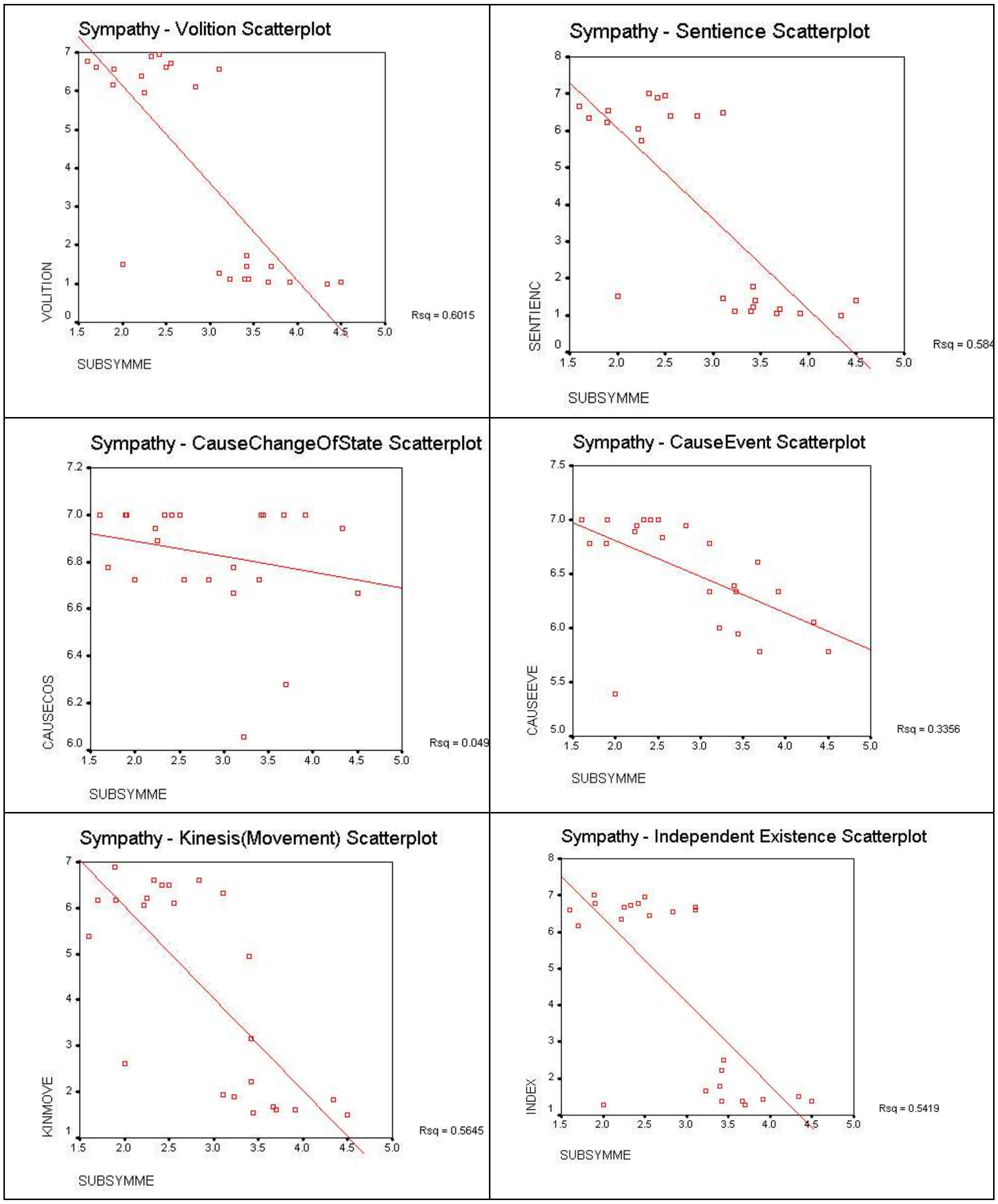
The strangling killed the deputy.
The smothering killed the grandmother.
The choking killed the vagrant.
The drowning killed the young boy.
Suffocation killed the 24-year old woman.
Starvation killed the severely ill woman.

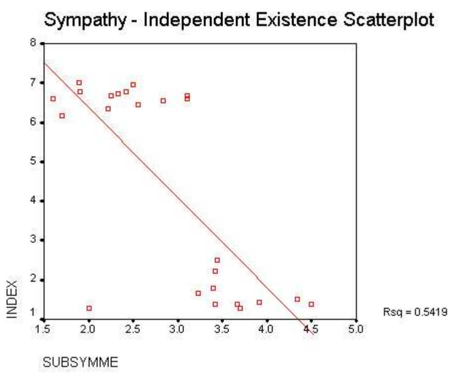
Appendix 2 – Experimental Stimulus Headlines

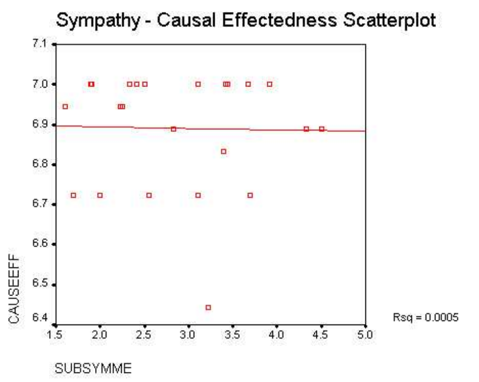
verb	**Transitive Effective Headline**	**Nominalized Transitive Headline**	**Passive Headline**
kill	Terrorists kill eight marketgoers	Explosion kills eight marketgoers	Eight marketgoers are killed
kill	Rebel kills army officer.	Attack kills army officer	Army officer is killed
slaughter	Terrorists slaughter nine hostages	Slaughter kills nine hostages	Nine hostages are slaughtered
assassinate	City council member assassinates rival candidate	Assassination plot kills rival candidate.	Rival candidate is assassinated.
shoot	Gunmen shoot opposition leader	Shooting kills opposition leader	Opposition leader is shot
poison	Daughters poison blind woman	Poisoning kills blind woman	Blind woman is poisoned
strangle	Man in custody strangles deputy	Strangling kills deputy	Deputy is strangled
smother	woman smothers grandmother	Smothering kills grandmother	Grandmother is smothered
choke	Local man chokes vagrant	Choking kills vagrant	Vagrant is choked
drown	Teenager drowns young boy	Drowning kills young boy	Young boy is drowned
suffocate	Man suffocates 24-year old woman	Suffocation kills 24-year old woman	Twenty-four year old woman is suffocated
starve	Caretaker starves severely ill woman	Starvation kills severely ill woman	Severely ill woman is starved

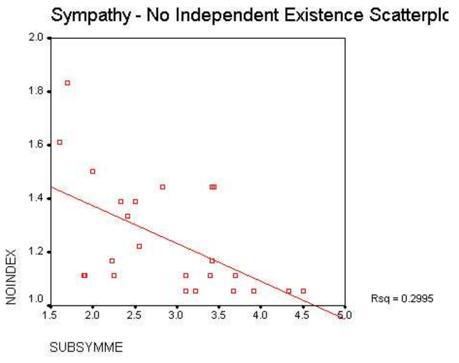
Appendix 3 – Scatter Plots

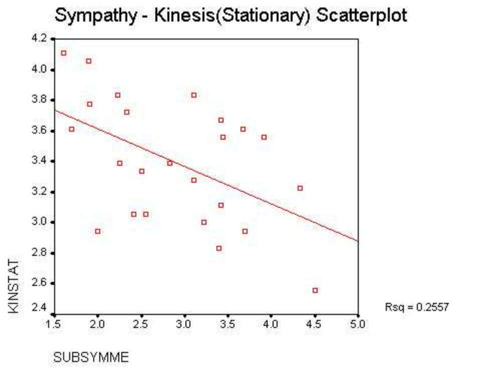
Scatter plots for Independent Variables (Semantic Components) by Dependent Variable (Sympathy for Perpetrator)

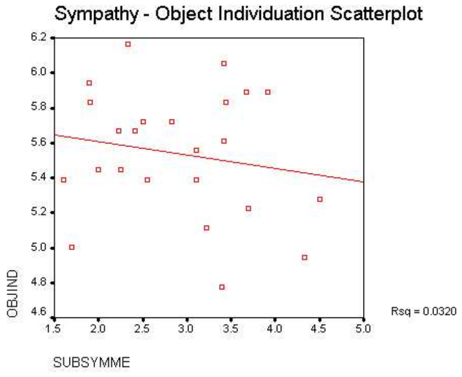


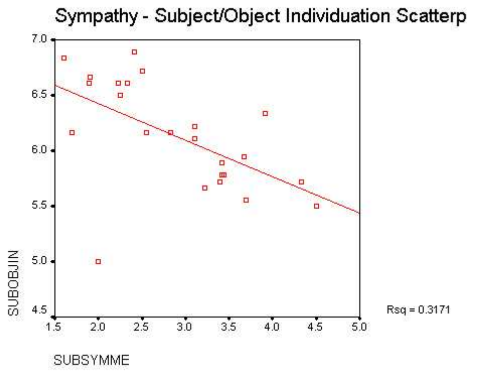

Sympathy - Independent Existence Scatterplot
INDEX
8
7
6
5
4
3
2
1
1.5
2.0
2.5
3.0
3.5
4.0
4.5
5.0
SUBSYMME
Rsq = 0.5419

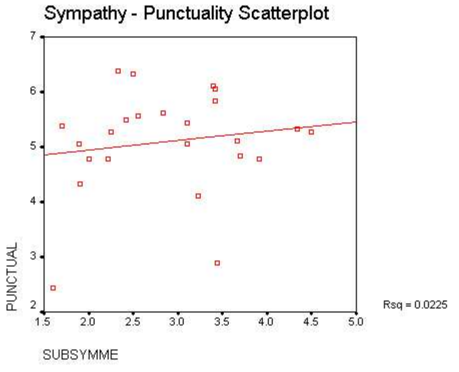

Sympathy - Causal Effectedness Scatterplot
CAUSEEFF
7.1
7.0
6.9
6.8
6.7
6.6
6.5
6.4
1.5
2.0
2.5
3.0
3.5
4.0
4.5
5.0
SUBSYMME
Rsq = 0.0005

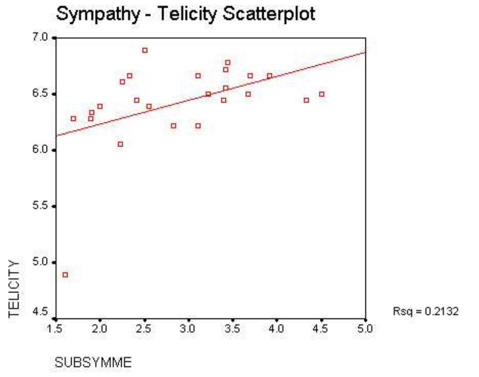

Sympathy - No Independent Existence Scatterplc
NOINDEX
2.0
1.8
1.6
1.4
1.2
1.0
1.5
2.0
2.5
3.0
3.5
4.0
4.5
5.0
SUBSYMME
Rsq = 0.2995


Sympathy - Kinesis(Stationary) Scatterplot
KINSTAT
4.2
4.0
3.8
3.6
3.4
3.2
3.0
2.8
2.6
2.4
1.5
2.0
2.5
3.0
3.5
4.0
4.5
5.0
SUBSYMME
Rsq = 0.2557


Sympathy - Object Individuation Scatterplot
OBJIND
6.2
6.0
5.8
5.6
5.4
5.2
5.0
4.8
4.6
1.5
2.0
2.5
3.0
3.5
4.0
4.5
5.0
SUBSYMME
Rsq = 0.0320


Sympathy - Subject/Object Individuation Scatterp
SUBOBJIN
7.0
6.5
6.0
5.5
5.0
4.5
1.5
2.0
2.5
3.0
3.5
4.0
4.5
5.0
SUBSYMME
Rsq = 0.3171


Sympathy - Punctuality Scatterplot
PUNCTUAL
7
6
5
4
3
2
1.5
2.0
2.5
3.0
3.5
4.0
4.5
5.0
SUBSYMME
Rsq = 0.0225


Sympathy - Telicity Scatterplot
TELICITY
7.0
6.5
6.0
5.5
5.0
4.5
1.5
2.0
2.5
3.0
3.5
4.0
4.5
5.0
SUBSYMME
Rsq = 0.2132

Appendix 4 – Grammatical Relations of the Stanford Parser

Typed Dependency / Grammatical Relation		**Description and Example(s)**
aux (auxiliary)		An auxiliary of a clause is a non-main verb of the clause. Example: "Reagan has died" → aux(died, has)
	auxpass (passive auxiliary)	A passive auxiliary of a clause is a non-main verb of the clause which contains the passive information. Example: "Kennedy has been killed" → auxpass(killed, been)
	cop (copula)	A copula is the relation between the complement of a copular verb and the copular verb. Examples: "Bill is big" → cop(big, is) "Bill is an honest man" → cop(man, is)
conj (conjunct)		A conjunct is the relation between two elements connected by a conjunction word. Example: "Bill is big and honest" → conj(big, honest)
cc (coordination)		A coordination is the relation between an element and a conjunction. Example: "Bill is big and honest." → cc(big, and)
pred (predicate)		The predicate of a clause is the main VP of that clause; the predicate of a subject is the predicate of the clause to which the subject belongs. Example: "Reagan died" → pred(Reagan, died)
arg (argument)		An argument of a VP is a subject or complement of that VP; an argument of a clause is an argument of the VP which is the predicate of that clause. Example: "Clinton defeated Dole" → arg(defeated, Clinton), arg(defeated, Dole)
	subj (subject)	The subject of a VP is the noun or clause that performs or experiences the VP; the subject of a clause is the subject of the VP which is the predicate of that clause. Examples: "Clinton defeated Dole" → subj(defeated, Clinton) "What she said is untrue" → subj(is, What she said)

Typed Dependency / Grammatical Relation				**Description and Example(s)**
		nsubj (nominal subject)		A nominal subject is a subject which is a noun phrase. Example: "Clinton defeated Dole" → nsubj(defeated, Clinton),
			nsubjpass (passive nominal subject)	A nominal passive subject is a subject of a passive which is an noun phrase. Example: "Dole was defeated by Clinton" → nsubjpass(defeated, Dole)
		csubj (clausal subject)		A clausal subject is a subject which is a clause. Examples: subject is "what she said" in both examples "What she said makes sense" → csubj(makes, said) "What she said is untrue" → csubj(untrue, said)
	comp (complement)			A complement of a VP is any object (direct or indirect) of that VP, or a clause or adjectival phrase which functions like an object; a complement of a clause is a complement of the VP which is the predicate of that clause. Examples: "She gave me a raise" → comp(gave, me), comp(gave, a raise) "I like to swim" → comp(like, to swim)
		obj (object)		An object of a VP is any direct object or indirect object of that VP; an object of a clause is an object of the VP which is the predicate of that clause. Example: "She gave me a raise" → obj(gave, me), obj(gave, raise)
			dobj (direct object)	The direct object of a VP is the noun phrase which is the (accusative) object of the verb; the direct object of a clause is the direct object of the VP which is the predicate of that clause. Example: "She gave me a raise" → dobj(gave, raise)
			iobj (indirect object)	The indirect object of a VP is the noun phrase which is the (dative) object of the verb; the indirect object of a clause is the indirect object of the VP which is the predicate of that clause. Example: "She gave me a raise" → iobj(gave, me)

Typed Dependency / Grammatical Relation			**Description and Example(s)**
		pobj (object of preposition)	The object of a preposition is the head of a noun phrase following the preposition. (The preposition in turn may be modifying a noun, verb, etc.)[46] Example: "I sat on the chair" → pobj(on, chair)
	attr (attributive)		Example: "What is that?" → attr(is, What)
	ccomp (clausal complement with internal subject)		A clausal complement of a VP or an ADJP is a clause with internal subject which functions like an object of the verb or of the adjective; a clausal complement of a clause is the clausal complement of the VP or of the ADJP which is the predicate of that clause. Such clausal complements are usually finite (though there are occasional remnant English subjunctives). Examples: "He says that you like to swim" → ccomp(says, like) "I am certain that he did it" → ccomp(certain, did)
	xcomp (clausal complement with external subject)		An xcomp complement of a VP or an ADJP is a clausal complement with an external subject. (For now, only "TO-clause" are recognized, as well as participial clauses.) These xcomps are always non-finite. Example: "I like to swim" → xcomp(like, swim) "I am ready to leave" → xcomp(ready, leave)
	compl (complementizer)		A complementizer of a clausal complement is the word introducing it. Example: "He says that you like to swim" → complm(like, that)
	mark (marker – word introducing an advcl)		A marker of an adverbial clausal complement is the word introducing it. Example: "U.S. forces have been engaged in intense fighting after insurgents launched simultaneous attacks" → mark(launched, after)
	rel (relative – word introducing a rcmod)		A relative of a relative clause is the head word of the WH-phrase introducing it. Examples: "I saw the man that you love" → rel(love, that) "I saw the man whose wife you love" → rel(love, wife)

[46] I use the Stanford Parser's "collapsed" form of the relations which in some cases reduces two relations to one. For prepositions, the `prep` and `pobj` relations are combined. For example, the phrase "reason for a change" produces the relations `prep(reason-2, for-3)` and `pobj(for-3, change-6)`. These relations are "collapsed" to the single relation `prep_for(reason, change)`.

Typed Dependency / Grammatical Relation			**Description and Example(s)**
		acomp (adjectival complement)	An adjectival complement of a VP is a adjectival phrase which functions like an object of the verb; an adjectival complement of a clause is the adjectival complement of the VP which is the predicate of that clause. Example: "She looks very beautiful" → acomp(looks, very beautiful)
	agent (agent)		The agent of a passive VP is the complement introduced by "by" and doing the action. Examples: "The man has been killed by the police" → agent(killed, police)
ref (referent)			The "referent" grammatical relation. A referent of NP is a relative word introducing a relative clause modifying the NP. Example: "I saw the man that you love" → ref(man, that)
expl (expletive)			This relation captures an existential there. Example: "There is a statue in the corner" → expl(is, there)
punct (punctuation)			This is used for any piece of punctuation in a clause, if punctuation is being retained in the typed dependencies. Example: "Go home!" → punct(Go, !)
mod (modifier)			A modifier of a VP is any constituent that serves to modify the meaning of the VP (but is not an ARGUMENT of that VP); a modifier of a clause is a modifier of the VP which is the predicate of that clause. Examples: "I swam in the pool last night" → mod(swam, in the pool), mod(swam, last night)
	advcl (adverbial clause modifier)		An adverbial clause modifier of a VP is a clause modifying the verb (temporal clauses, consequences, conditional clauses, etc.) Example: "The accident happened as night was falling" → advcl(happened, falling) "If you know who did it, you should tell the teacher" → advcl(know, tell)

Typed Dependency / Grammatical Relation		Description and Example(s)
	purpcl (purpose clause modifier)	A purpose clause modifier of a VP is a clause headed by "(in order) to" specifying a purpose. Note: at present we only recognize ones that have "in order to" as otherwise we can't give our surface representations distinguish these from xcomp's. We can also recognize "to" clauses introduced by "be VBN". Example: "He talked to the president in order to secure the account" → purpcl(talked, secure)
	tmod (temporal modifier)	A temporal modifier of a VP or an ADJP is any constituent that serves to modify the meaning of the VP or the ADJP by specifying a time; a temporal modifier of a clause is an temporal modifier of the VP which is the predicate of that clause. Examples: "I swam in the pool last night" → tmod(swam, last night)
	rcmod (relative clause modifier)	A relative clause modifier of an NP is a relative clause modifying the NP. The link points from the head noun of the NP to the head of the relative clause, normally a verb. Examples: "I saw the man that you love" → rcmod(man, love) "I saw the book which you bought" → rcmod(book, bought)
	amod (adjectival modifier)	An adjectival modifier of an NP is any adjectival phrase that serves to modify the meaning of the NP. Examples: "Sam eats red meat" → amod(meat, red)
	infmod (infinitival modifier)	Example: "points to establish are ..." → infmod(points, establish)
	partmod (participial modifier)	A participial modifier of an NP or VP is a VP[part] that serves to modify the meaning of the NP or NP. Examples: "truffles picked during the spring are tasty" → partmod(truffles, picked) "Bill picked Fred for the team demonstrating his incompetence" → partmod(picked, demonstrating)
	num (numeric modifier)	A numeric modifier of an NP is any number phrase that serves to modify the meaning of the NP. Examples: "Sam eats 3 sheep" → num(sheep, 3)

Typed Dependency / Grammatical Relation		Description and Example(s)
	number (element of compound number)	A compound number modifier is a part of a number phrase or currency amount. Example: I lost $ 3.2 billion" → number($, billion3)
	appos (appositional modifier)	An appositional modifier of an NP is an NP that serves to modify the meaning of the NP. Examples: "Sam, my brother, eats red meat" → appos(Sam, brother)
	nn (noun compound modifier)	A noun compound modifier of an NP is any noun that serves to modify the head noun. Note that this has all nouns modify the rightmost a la Penn headship rules. There is no intelligent noun compound analysis. Example: "Oil price futures" nn(futures, oil) nn(futures, price)
	abbrev (abbreviation modifier)	An abbreviation modifier of an NP is an NP that serves to abbreviate the NP. Examples: "The Australian Broadcasting Corporation (ABC)" → abbrev(Corporation, ABC)
	advmod (adverbial modifier)	An adverbial modifier of a word is an RB or ADVP that serves to modify the meaning of the word. Examples: "genetically modified food" → advmod(modified, genetically)
	neg (negation modifier)	The negation modifier is the relation between a negation word and the word it modifies. Examples: "Bill is not a scientist" → neg(scientist, not) "Bill doesn't drive" → neg(drive, n't)
	poss (possession modifier)	Example: "their offices" → poss(offices, their) "Bill 's clothes" → poss(clothes, Bill)
	possessive (possessive modifier – 's)	Example: "John's book" → possessive(John, 's)
	prt (phrasal verb particle)	The "phrasal verb particle" relation identifies phrasal verbs. Examples: "They shut down the station." → prt(shut, down)
	det (determiner)	Examples: "The man is here" → det(man,the) "In which city do you live?" → det(city,which)
	predet (predeterminer)	Examples: "All the boys are here" → predet(boys,all)
	preconj (preconjunct)	Examples: "Both the boys and the girls are here" → preconj(boys,both)

Typed Dependency / Grammatical Relation		Description and Example(s)
	prep (prepositional modifier)	A prepositional modifier of an NP or VP is any prepositional phrase that serves to modify the meaning of the NP or VP. Examples: "I saw a cat in a hat" → prep(cat, in) "I saw a cat with a telescope" → prep(saw, with)
sdep (semantic dependent)		The "semantic dependent" grammatical relation has been introduced as a supertype for the controlling subject relation.
	xsubj (controlling subject)	Examples: "Tom likes to eat fish" → xsubj(eat, Tom)

Table 30 - Grammatical relations available from the Stanford Parser, taken from de Marneffe et al. (2006) and the javadoc documentation available from the parser's distribution

Appendix 5 – Development of the Death Penalty Corpus

The Web represents a vast resource of text for language and language technology research, but the data comes with well known problems.[47] This appendix provides details on the procedures I have used to prepare web-based data for research purposes, and the rationale behind those procedures. The document is based on my experience preparing data for the DP Corpus.

The idea behind the DP Corpus is to provide textual material representing the viewpoints of both the pro-death penalty and anti-death penalty movements. The specific web sites from which documents were collected were found through Google searches, and are discussed in Section 3.3.2.

The corpus includes documents that were downloaded from other sites through links. The "crawling" or "spidering" was carried out using the wget tool (http://www.gnu.org/software/wget). Link depth was limited to five levels. Roughly 1000 (non-image) documents from each side of the death penalty debate were downloaded.

Rationale for Cleaning Procedures

After downloading the initial documents, it quickly became clear that several issues common to web-based data indicated the need for both automated and manual cleaning and filtering of the documents. The issues summarized here provide the rationale for the reasonably intensive cleansing of the data that I carried out:

1. Repetitive headers, footers, and other navigational material (aka HTML "furniture"). My interest is in the textual commentary, prose, and real language use in the documents. But these repetitive elements can be too easily picked up by learning algorithms and dominate the results. Additionally, models based on these textual elements would certainly not transfer well to documents from other origins. These needed to be removed.
2. Some documents had virtually no real content in them. These are often forms, purely navigational elements, and the like. Some documents contained tables of numeric data, with virtually no sentential content. These are outside the scope of interest. Of course, image files, stylesheets and the like, must also be removed.
3. Some documents could be plausibly judged as neutral and/or off topic. The task we are focusing on is sentiment detection, not topic classification.

[47] There is an active community developing tools to help manage web-based corpus research. For example, see http://www.sketchengine.co.uk/WebBootCaT.htm. There is also a series of ACL workshops devoted entirely to building corpora from the web (see http://cental.fltr.ucl.ac.be/wac3). This year's workshop includes a competitive shared task devoted specifically to clean up procedures for web-derived corpus documents, highlighting the shared concerns motivating the procedures described here.

4. Some documents were exact copies of other documents found under different names.

Data Preparation and Cleansing Procedures

The first step in the data preparation procedure was to extract the text from all downloaded documents, including HTML, MS Word and PDF documents. The raw text was then sentence-delimited.

The first pass at data cleaning was an automatic, heuristic series of filters:

1. These file types were retained: MS Office documents, PDF, html, xml, xhtml, and many other content-full formats (e.g. WordPerfect, RTF). Non-text file types were removed (.jpg, gif, .css).
2. Some boilerplate data filtering was employed. When text was extracted and sentence-delimited, a small set of items was removed (e.g. copyright notices).
3. An additional layer of filters worked to automatically weed out useless documents. For each document, I performed the following tests:
   - If the file had less than ten total sentences, I rejected the file. Ideally this threshold would be made configurable, because clearly this would not work for documents such as message posts.
   - If the difference between the longest and shortest sentence was less than five tokens, I rejected the file. The motivation here is that there is not a large enough range in sentence length for it to be prose. It was likely tabular data, for example.
   - If more than half the sentences had length less than three, I rejected the file. Again, the file is probably tabular data, mis-tagged media, or something else un-prose-like.
   - If two times the standard deviation of sentence token length was less than five (i.e. 95% of the sentences are less than five tokens different in length from the mean token length), I rejected the file. This was another approach to eliminating tabular or list data files.
   - For files that pass all these tests, I eliminated sentences of token length less than three. I also eliminated each sentence for which 90% or more of its tokens are punctuation or digits.

Running this regimen against the DP Corpus, a visual inspection all of the rejected material indicated it was virtually all tabular data, lists, headers, footers, navigation links, long runs of punctuation and other noise. The precision of what was eliminated was very high, though recall turned out to be fairly low. Rejection rates of documents based on this procedure for the DP Corpus as a whole ranged from two to seven percent. For accepted documents, between four and 11 present of sentences were rejected.

After experimenting with the data produced from the procedure defined so far, I found that large amounts of HTML furniture remained in the files. Additionally, random inspection of individual files found lingering navigation files, forms, tabular data, and off-topic documents. Because of this, a manual inspection and filtering of the documents was conducted. The procedure was as follows:

1. I generated a spreadsheet that provided, for each document, hyperlinks to both the original document and its sentence-delimited extracted text. For each document, columns were provided for:
    - sentiment (PRO, CON, neutral, N/A)
    - general document tone (legal, scientific, journalistic, op-ed – this is informal)
    - Special notes or comments
    - Boolean flag indicating whether to retain the document
    - Boolean flag indicating if the sentence-delimited file was manually edited in this phase
2. I inspected each document in it original form and judged it as to its sentiment, contentfull-ness and overall relevance. Off-topic or purely navigational files, forms, duplicates, etc. are marked to be eliminated. Marginal documents are marked with special notes (e.g. biblical excerpts). Retained documents were hand-edited if deemed necessary, in their sentence-delimited version, for repetitive footer or navigational material that had survived to this point, and marked thusly.

The manual procedure just described was conducted for the PRO half of the DP Corpus in its entirety, and for about 100 documents from the ANTI half. As there were 596 documents retained from the PRO half of the corpus, in order to expedite matters the top 596 documents from the ANTI half of the corpus by size (subject to the indications of the 100 files that had already been inspected) were taken to be the complementary half of the corpus. I had found that in general, the non-substantive documents tend to be the smallest and thus the larger documents were retained.

A final automated procedure was applied, necessitated in part because the time consuming manual procedure was not carried out in its entirety: All sentences across each half of the corpus were scanned for duplicate sentences. These in almost every case were additional footers, table headers or other formatting material, and in some cases appeared to be publishing errors. Additionally, this method was used to find remaining documents that were entire duplicates of each other. In these cases, one copy of the document was retained and the others were eliminated. The final corpus has 1152 documents, split evenly between PRO and CON.

A valuable enhancement to the DP Corpus would be to conduct annotation for ground truth labels by one or more additional human annotators. This would facilitate the calculation of inter-annotator agreement and provide an upper bound on the document level sentiment classification task for this corpus. However, I believe with great confidence that inter-annotator agreement would be exceptionally high. While I did find neutral or off-topic documents, I found no documents that would be labeled with a ground truth value of PRO or CON that was the opposite of that which was associated with its web site of origin.

Appendix 6 – Graph Union vs. Feature Union

To further examine the value of the graph union approach to classifier combination introduced in Section 4.2.6, I considered an alternative: build a single SVM classifier using a *feature union* of the features used to build the two separate SVM models. For OPUS-SVM, I naturally have the feature vectors available, but for U-SVM, I have only the raw hyperplane distances and derived scores as provided by TPL06. Thus I built my own unigram-based SVM classifier, trained on the training set of the CS corpus. It is difficult to exactly reproduce the tokenization as used by TPL06 to produce U-SVM so this model, U-SVM$_2$, is not the same. For that reason, I ran experiments testing both the feature union of U-SVM$_2$ and OPUS-SVM and the graph union of the two.

Matt Thomas (p.c.) provided the token list used to build U-SVM, so we can at least consider an overall comparison of the features sets. U-SVM's feature vectors are based on a set of 32,012 tokens, and U-SVM$_2$'s features vectors are based on a set of 23,936 tokens. The U-SVM token set has 99.38% recall of the U-SVM$_2$ token set (74.31% precision), so the smaller feature set is very nearly a proper subset of the larger. A qualitative inspection of the differences indicates that the tokenization differs primarily with regard to punctuation. For example, tokens in the U-SVM set appear to not separate the sentence-ending period from the final word of the sentence, which naturally would inflate the size of the token set, as we see here.

YEA or NAY classification of speech segments	U-SVM$_2$	Feature Union U-SVM$_2$ + OPUS-SVM	Graph Union U-SVM$_2$ + OPUS-SVM
SVM	66.51	67.67	68.37
SVM with same-speaker arcs	67.33	67.91	69.42*
SVM with same-speaker arcs and agreement arcs, $\theta_{agr} = 0$	67.09	64.77	68.72*
SVM with same-speaker arcs and agreement arcs, $\theta_{agr} = \mu$	65.00	65.93	69.42*
SVM with same-speaker arcs and agreement arcs, $\theta_{agr} = 1.5\mu$	66.63	66.63	67.79

Table 31 - Classification accuracy for the graph union of two SVM models, in percent. Asterisks indicate significant improvements of Graph Union over Feature Union.

Table 31 summarizes the results of experiments with U-SVM$_2$ alone, and as a feature union and graph union with OPUS-SVM.

Accuracy for the U-SVM$_2$ classifier alone is initially (not significantly) higher than U-SVM, but it fails to benefit as much from the addition of inter-item relations. The more interesting result is that the graph union of the classifiers uniformly performs better than the feature union of the classifiers. This improvement is significant for the same-speaker condition ($n_+ = 4$, $n_- = 17$, $p < 0.01$) and the agreement conditions for $\theta_{agr} = 0$ ($n_+ = 3$, $n_- = 37$, $p << 0.01$) and $\theta_{agr} = \mu$ ($n_+ = 11$, $n_- = 41$, $p << 0.01$). Overall these results support the conclusion that the novel graph union method has appreciable value.

Bibliography

Allport, G.W. (1935). Attitudes. In C.M. Murchison (ed.), *Handbook of Social. Psychology*. Winchester, MA: Clark University Press

Baroni, M., & Vegnaduzzo, S. (2004). Identifying subjective adjectives through web-based mutual information. In *Proceedings of KONVENS-04* (pp. 17–24). Vienna. Retrieved from http://sslmit.unibo.it/baroni/publications/konvens2004/wmiKONV.pdf

Bem, D. J. (1970). *Beliefs, attitudes and human affairs*. Belmont, CA: Wadsworth.

Benamara, F., Cesarano, C., Picariello, A., Reforgiato, D., & Subrahmanian, V.S. (2007). Sentiment analysis: adjectives and adverbs are better than adjectives alone. In *Proceedings of ICWSM 2007*. Retrived from http://oasys.umiacs.umd.edu/oasys/papers/icwsmV2.pdf

Bethard, S., Yu, H., Thornton, A., Hatzivassiloglou, V., & Jurafsky, D. (2004). Automatic extraction of opinion propositions and their holders. In J. G. Shanahan, J. Wiebe, & Y. Qu (Eds.), *Proceedings of the AAAI Spring Symposium on Exploring Attitude and Affect in Text: Theories and Applications*, Stanford:ACM. Retrieved from http://www.stanford.edu/~jurafsky/SS404BethardS.pdf

Blum, A., & Chawla, S. (2001). Learning from labeled and unlabeled data using graph mincuts. In *Proceedings of ICML* (pp.19–26).

Bolinger, D. L. (1968). Entailment and the meaning of structures. *Glossa, 2*, 119–127.

Brank, J., Grobelnik, M., Milić-Frayling, N., & Mladenić, D. (2002). Feature selection using linear support vector machines. Microsoft Technical Report MSR-TR-2002-63.

Burstein, M. H., (1979). The use of object-specific knowledge in natural language processing. In *Proceedings of the 17th Annual Meeting of the Association for Computational Linguistics* (pp. 53-58). San Diego:ACL. Retrieved from http://acl.ldc.upenn.edu/P/P79/P79-1013.pdf.

Carroll, J., Minnen, G., & Briscoe, T. (1999). Corpus annotation for parser evaluation. In *Proceedings of the EACL workshop on Linguistically Interpreted Corpora (LINC),* (pp. 35-41). Retrieved from http://xxx.lanl.gov/PS_cache/cs/pdf/9907/9907013v1.pdf.

Cesarano, C., Dorr, B., Picariello, A., Reforgiato, D., Sagoff, A., & Subrahmanian, V.S. (2006, March). OASYS: An opinion analysis system. In *Proceedings of*

AAAI-2006 Spring Symposium on Computational Approaches to Analyzing Weblogs. http://oasys.umiacs.umd.edu/oasys/papers/oct31.ps

Cesarano, C., Picariello, A., Reforgiato, D., & Subrahmanian, V.S. (2007). The OASYS 2.0 opinion analysis system. Accepted as DEMO to ICWSM 2007. Retrieved from http://oasys.umiacs.umd.edu/oasys/papers/icwsm-demo.pdf

Chekuri, C. S., Goldberg, A. V., Karger, D. R., Levine, M. S., & Stein, C. (1997, January). Experimental study of minimum cut algorithms. In *8th Annual ACM-SIAM Symposium on Discrete Algorithms* (pp. 324-333).

Cherkassky, B.V., & Goldberg, A.V. (1997). On implementing push-relabel method for the maximum flow problem. *Algorithmica*, *19*, 390-410.

Chesley, P., Vincent, B., Xu, L., & Srihari, R. (2006). Using verbs and adjectives to automatically classify blog sentiment. In *Proceedings of AAAI-CAAW-06, the Spring Symposia on Computational Approaches to Analyzing Weblogs*, Stanford:AAAI. Retrieved from http://www.aaai.org/Papers/Symposia/Spring/2006/SS-06-03/SS06-03-005.pdf

Conway, M., Grabe, M. E., & Grieves, K. (2007). Villains, victims and the virtuous in Bill O’Reilly’s “No-Spin Zone”: Revisiting world war propaganda techniques. *Journalism Studies*, *8*, 197-223.

Damerau, F. J. (1993). Generating and evaluating domain-oriented multi-word terms from texts. *Information Processing and Management*, *29*, 433–447.

de Marneffe, M., MacCartney, B. & Manning, C. D. (2006). Generating typed dependency parses from phrase structure parses. In *Proceedings of LREC 2006*. Retrieved from http://www-nlp.stanford.edu/~wcmac/papers/td-lrec06.pdf.

Dowty, D. R. (1991). Thematic proto-roles and argument selection. *Language*, *67*, 547-619.

Dunning, T. (1993). Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics*, *19*, 61-74.

Durbin, S. D., Richter, J. N., & Warner, D. (2003). A system for affective rating of texts. In *Proceedings of the 3rd Workshop on Operational Text Classification, 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining Washington, DC*. Retrieved from http://www.cs.montana.edu/~richter/affective_rating.pdf.

Eagly, A. and Chaiken, S. (1993). *The psychology of attitudes*. Fort Worth, TX: Harcourt Brace Jovanovich.

Ekman, P. (1994). *Moods, emotions, and traits.* In P. Ekman and R. J. Davidson (Eds.), *The Nature of Emotion*. Oxford: Oxford University Press.

Filip, H., Tanenhaus, M. K., Carlson, G. N., Allopenna, P. D., & Blatt, J. (2001). Reduced relatives judged hard require constraint-based analyses. In S. Stevenson and P. Merlo (Eds.), *Lexical representations in sentence processing*. Cambridge: Cambridge University Press

Fishbein, M., & Ajzen, I. (1975). *Belief, attitude, intention, and behavior: an introduction to theory and research*. Reading, MA: Addison-Wesley.

Fisher, C., Gleitman, H., & Gleitman, L. R. (1991). On the semantic content of subcategorization frames. *Cognitive Psychology, 23*, 331-392.

Gamon, M. (2004a). Linguistic correlates of style: authorship classification with deep linguistic analysis features. In *Proceedings of COLING 2004* (pp. 611-617). Retrieved from http://research.microsoft.com/nlp/publications/coling2004_authorship.pdf.

Gamon, M. (2004b). Sentiment classification on customer feedback data: noisy data, large feature vectors, and the role of linguistic analysis. In *Proceedings of COLING 2004*, (pp 841-847). Retrieved from http://research.microsoft.com/nlp/publications/coling2004_sentiment.pdf.

Gamon, M., & Aue, A. (2005). Automatic identification of sentiment vocabulary: exploiting low association with known sentiment terms. In *Proceedings of the Workshop on Feature Engineering for Machine Learning in Natural Language Processing at ACL 2005*, (pp. 57-64). Retrieved from http://research.microsoft.com/nlp/publications/sentiment_ACL_workshop_2005.pdf.

Gentzkow, M., & Shapiro, J. (2006a, November). What drives media slant? evidence from U.S. newspapers. Retrieved from http://home.uchicago.edu/~jmshapir/biasmeas111306.pdf

Gentzkow, M., & Shapiro, J. (2006b). Media bias and reputation. *Journal of Political Economy*, *114*, 280-316. Retrieved from http://home.uchicago.edu/~jmshapir/bias.pdf.

Goldberg, A. E. (1995). *Constructions: A construction grammar approach to argument structure*. Chicago:University of Chicago Press.

Goldberg, A. B., & Zhu, J. (2006). Seeing stars when there aren't many stars: graph-based semi-supervised learning for sentiment categorization. In *TextGraphs: HLT/NAACL Workshop on Graph-based Algorithms for Natural Language Processing*. Retrieved from http://pages.cs.wisc.edu/~goldberg/publications/sslsa.pdf.

Grefenstette, G., Qu, Y., Shanahan, J. G., & Evans, D. A. (2004). Coupling niche browsers and affect analysis for an opinion mining application. In *Proceedings of RIAO 2004*. Retrieved from http://www.riao.org/Proceedings-2004/papers/0140.pdf

Grimshaw, J. (1990). *Argument structure*. Cambridge, MA:MIT Press.

Groseclose, T., & Milyo, J. (2005). A measure of media bias. *The Quarterly Journal of Economics, 120*, 1191-1237. Retrieved from http://www.polisci.ucla.edu/faculty/groseclose/pdfs/MediaBias.pdf.

Harden, B. (2006, July 26). On Puget Sound, it's orca vs. inc. *The Washington Post*, p. A3.

Hatzivassiloglou, V., & McKeown, K. R. (1997). Predicting the semantic orientation of adjectives. In *Proceedings of the 35th Annual Meeting of the ACL.* Somerset, NJ:Association for Computational Linguistics (pp. 174-181).

Hearst, M. (1992). Direction-Based Text Interpretation as an Information Access Refinement. In P. Jacbos (ed.) *Text-Based Intelligent Systems*, Hillsdale, NJ:Lawrence Erlbaum.

Hopper, P., & Thompson, S. A. (1980). Transitivity in grammar and discourse. *Language*, *57*, 251-295.

Joachims, T. (1998). Text categorization with support vector machines: learning with many relevant features. In *Proceedings of the European Conference on Machine Learning (ECML).* (pp. 137-142). Berlin: Springer.

Joachims, T. (1999). Making large-Scale SVM Learning Practical. In B. Schölkopf, C. Burges & A. Smola (Eds.), *Advances in Kernel Methods - Support Vector Learning*, (pp. 137-142) . Cambridge, MA:MIT-Press.

Kako, E. (2006a). Thematic role properties of subjects and objects. *Cognition*, *101,*1-42.

Kako, E. (2006b). The semantics of syntactic frames. *Language and Cognitive Processes*, *21*, 562-575.

Keller, F., & Alexopoulou, T. (2001). Phonology competes with syntax: experimental evidence for the interaction of word order and accent placement in the realization of information structure. *Cognition*, *79*, 301-372. Retrieved from http://homepages.inf.ed.ac.uk/keller/papers/cognition01.html

Keller, F., Corley, M., Corley, S., & Konieczny, L. (1998). WebExp: A java toolbox for web-based psychological experiments. Technical Report HCRC/TR-99, Human Communication Research Centre, University of Edinburgh. Retrieved from http://www.hcrc.ed.ac.uk/web_exp/doc/users_guide.html

Kilgarriff, A. (1997). Putting frequencies into the dictionary. *International Journal of Lexicography, 10*, 135-155. Retrieved from http://www.kilgarriff.co.uk/Publications/1996-K-IJLFreqs.pdf

King, T. H., Crouch, R., Riezler, S., Dalrymple, M., & Kaplan, R. (2003). The PARC 700 dependency bank. In *4th International Workshop on Linguistically Interpreted Corpora* (LINC-03). Retrieved from http://citeseer.ist.psu.edu/cache/papers/cs/29258/http:zSzzSzwww2.parc.comzSzistlzSzmemberszSzriezlerzSzPAPERSzSzLINC03.pdf/king03parc.pdf .

Klavans, J., & Kan, M. (1998). Role of verbs in document analysis. In *Proceedings of the Conference COLING-ACL*, (pp. 680-686). Montreal:ACL.

Klein, D., & Manning, C. D. (2002, December). Fast exact inference with a factored model for natural language parsing. In S. Becker, S. Thrun, & Klaus Obermayer (Eds.), *Advances in Neural Information Processing Systems 15*, (pp. 3-10). Cambridge, MA: MIT Press.

Klein, D., & Manning, C. D. (2003). Accurate unlexicalized parsing, In *Proceedings of the 41st Meeting of the Association for Computational Linguistics,* (pp. 423-430). Somerset, NJ:ACL. Retrieved from http://nlp.stanford.edu/pubs/unlexicalized-parsing.pdf.

Kullback, S., & Leibler, R. A. (1951). On information and sufficiency. *The Annals of Mathematical Statistics, 22*, 79–86.

Lakoff, G. (1987). *Women, fire and dangerous things: what categories reveal about the mind*. Chicago:University of Chicago Press.

Leech, G., Rayson, P., & Wilson, A. (2001). *Word frequencies in written and spoken English: based on the British National Corpus*. London:Longman.

Lemmens, M. (1998). *Lexical perspectives on transitivity and ergativity*. Amsterdam:John Benjamins.

Levin, B., & Rappaport Hovav, M. (1995). *Unaccusativity: at the syntax –lexical semantics interface*. Cambridge,MA:MIT Press.

Levin, B. (1993). *English verb classes and alternations: a preliminary investigation*. Chicago:University of Chicago Press.

Levy, R., & Andrew, G. (2006). Tregex and Tsurgeon: tools for querying and manipulating tree data structures. In *Proceedings of LREC 2006*. Retrieved from http://ling.ucsd.edu/~rlevy/papers/levy_andrew_lrec2006.pdf.

Lin, W., & Hauptmann, A. (2006). Are these documents written from different perspectives? A test of different perspectives based on statistical distribution divergence. In *Proceedings of ACL 2006*, (pp. 1057-1064). Sydney:ACL. Retrieved from http://idee.inf.cs.cmu.edu/cmu_home/papers/lin06do_these_docum_convey_differ_persp.pdf.

Lin, W., Wilson, T., Wiebe, J., & Hauptmann, A. (2006). Which side are you on? Identifying perspectives at the document and sentence levels. In *Proceedings of Tenth Conference on Natural Language Learning (CoNLL),* (pp. 109-116). Retrieved from http://acl.ldc.upenn.edu/W/W06/W06-2900.pdf.

Marcus, M., Santorini, B., & Marcinkiewicz, M. A. (1993). Building a large annotated corpus of English: the Penn Treebank. *Computational Linguistics*, 19, 313-330. Retrieved from http://acl.ldc.upenn.edu/J/J93/J93-2004.pdf.

McKoon, G., & Ratcliff, R. (2003). Meaning through syntax: language comprehension and the reduced relative clause construction. *Psychological Review, 110*, 490-525.

McKoon, G., & Ratcliff, R. (2007). Interactions of meaning and syntax: Implications for models of sentence comprehension. *Journal of Memory and Language, 56*, 270-290

McKoon, G., & MacFarland, T. (2000). Externally and internally caused change of state verbs. *Language*, *76*, 833-858.

McKoon, G., & MacFarland, T. (2002). Event templates in the lexical representations of verbs. *Cognitive Psychology, 45*, 1-44.

Mishne, G. (2005). Experiments with mood classification in blog posts. In *Style2005 - 1st workshop on stylistic analysis of text for information access, SIGIR 2005*. Retrieved from http://staff.science.uva.nl/~gilad/pubs/style2005-blogmoods.pdf

Mulder, M., Nijholt, A., den Uyl, M., & Terpstra, P. (2004). A lexical grammatical implementation of affect. In *Proceedings of TSD-04, the 7th international conference: text, speech and dialogue*, *Lecture notes in computer science,* 3206, (pp. 171–178). Brno, CZ:Springer-Verlag. Retrieved from http://wwwhome.cs.utwente.nl/~anijholt/artikelen/tsd2004.pdf

Nigam, K., & Hurst, M. (2004). Towards a robust metric of opinion. In *AAAI Spring Symposium on Exploring Attitude and Affect in Text*. Retrieved from http://www.kamalnigam.com/papers/metric-EAAT04.pdf.

Nigam, K. & Hurst, M. (2006). Towards a robust metric of polarity. In J. Shanahan, Y. Qu, & J. Wiebe. (Eds.). *Computing attitude and affect in text: theory and applications*. Dordrecht, The Netherlands: Springer.

Ng, A.Y., & Jordan, M. (2002). On discriminative vs. generative classifiers: A comparison of logistic regression and naive bayes. In *Advances in Neural Information Processing Systems 14*. Cambridge, MA: MIT Press. Retrieved from http://citeseer.ist.psu.edu/cache/papers/cs/26676/http:zSzzSzwww-2.cs.cmu.eduzSzGroupszSzNIPSzSzNIPS2001zSzpaperszSzpsgzzSzAA28.pdf/on-discriminative-vs-generative.pdf.

Oard, D., Elsayed, T., Wang, J., Wu, Y., Zhang, P., Abels, E., Lin, J., Soergel, D. (2006). TREC 2006 at Maryland: blog, enterprise, legal and QA tracks. In *Proceedings of TREC 2006*. Retrieved from http://trec.nist.gov/pubs/trec15/papers/umd.blog.ent.legal.qa.final.pdf.

Olsen, M. B., & Resnik, P. (1997). Implicit object constructions and the (in)transitivity continuum. In *33rd Regional Meeting of the Chicago Linguistics Society*, (pp. 327-336). Retrieved from http://umiacs.umd.edu/~resnik/pubs/cls-97.paper.pdf.

Oskamp, S. (1991). *Attitude and opinions*, 2nd Ed. Englewood Cliffs, New Jersey: Prentice Hall.

Pang, B., & Lee, L. (2004). A sentimental education: sentiment analysis using subjectivity summarization based on minimum cuts. In *Proceedings of the Association for Computational Linguistics (ACL-2004)*, (pp. 271-278). Barcelona:ACL. Retrieved from http://acl.ldc.upenn.edu/P/P04/P04-1035.pdf.

Pang, B., Lee, L., & Vaithyanathan, S. (2002). Thumbs up? Sentiment classification using machine learning techniques. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP-2002)*, (pp. 79-86). Philadelphia:ACL. Retrieved from http://acl.ldc.upenn.edu/W/W02/W02-1011.pdf.

Pedler, J. (2003). A corpus-based update of a computer-usable dictionary. In *Proceedings of the 8th International Symposium on Social Communication* (pp. 487-492).

Pullum, G. K. (2003, December 17). Passive voice and bias in Reuter headlines about Israelis and Palestinians. Retrieved from http://itre.cis.upenn.edu/~myl/languagelog/archives/000236.html

Pustejovsky, J. (1991). The generative lexicon. *Computational Linguistics*, *17*, 409-441.

Quinn, K., Monroe, B., Colaresi, M., Crespin, M., & Radev, D. (2006). An automated method of topic-coding legislative speech over time with application to the 105th-108th U.S. Senate". Draft retrieved from http://www.people.fas.harvard.edu/~kquinn/papers/TopicsMethodDavis.pdf

Resnik, P. (1997, April 4-5) Selectional preference and sense disambiguation. Presented at the ACL SIGLEX Workshop on Tagging Text with Lexical Semantics: Why, What, and How? Washington, D.C. in conjunction with ANLP-97. Retrieved from http://www.ims.uni-stuttgart.de/~light/tueb_html/resnik2.ps.

Resnik, P., & Elkiss, A. (2005). The Linguist's Search Engine: An overview. In *Proceedings of the ACL Interactive Poster and Demonstration Sessions*, (pp. 31-36). Ann Arbor:ACL. Retrieved from http://acl.ldc.upenn.edu/P/P05/P05-3009.pdf.

Read, J. (2004). Recognising affect in text using pointwise-mutual information. (Masters thesis, University of Sussex, 2004). Retrieved from http://www.informatics.sussex.ac.uk/users/jlr24/papers/read-us04.pdf

Read, J. (2005). Using emoticons to reduce dependency in machine learning techniques for sentiment classification. In *Proceedings of the ACL 2005 Student Research Workshop*, (pp. 43-48). Ann Arbor:ACL. Retrieved from http://acl.ldc.upenn.edu/P/P05/P05-2008.pdf.

Riloff, E., & Wiebe, J. (2003). Learning extraction patterns for subjective expressions. In *Proceedings of the 2003 Conference on Empirical Methods in Natural Language Processing (EMNLP-03)*. Retrieved from http://acl.ldc.upenn.edu/W/W03/W03-1014.pdf.

Shulman, S., & Schlosberg, D. (2002). Electronic rulemaking: New frontiers in public participation. Paper presented at the annual meeting of the American Political Science Association, Boston Marriott Copley Place, Sheraton Boston & Hynes Convention Center, Boston, Massachusetts. Retrieved from http://www.allacademic.com/meta/p66302_index.html

Sinclair, J. (1991). *Corpus, concordance, collocation*. Oxfort: Oxford University Press.

Stefanowitsch, A., & Gries, S. Th. (2003). Collostructions: investigating the interaction of words and constructions. *International Journal of Corpus Linguistics*, *8*, 209-243.

Stone, P.J. (1997). Thematic text analysis: new agendas for analyzing text content. In C. Roberts (ed.), *Text Analysis for the Social Sciences*, (pp. 35-54). Mahwah:NJ:Lawrence Erlbaum Associates.

Subasic, P., & Huettner, A. (2001). Affect Analysis of Text Using Fuzzy Semantic Typing. In *IEEE Transactions on Fuzzy Systems*, *9*, (pp. 483-496). Retrieved from http://www.clairvoyancecorp.com/talks/FuzzyTypingCP.pdf

Takamura, H., Inui, T., & Okumura, M. (2005). Extracting emotional polarity of words using spin model. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, (pp. 133-140). Ann Arbor:ACL. Retrieved from http://acl.ldc.upenn.edu/P/P05/P05-1017.pdf

Thomas, M., Pang, B., & Lee, L. (2006). Get out the vote: determining support or opposition from Congressional floor-debate transcripts. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, (pp. 327-335). Sydney:ACL. Retrieved from http://www.aclweb.org/anthology/W/W06/W06-1639, http://www.cs.cornell.edu/home/llee/papers/tpl-convote.dec06.pdf

Tomokiyo, T., & Hurst, M. (2003). A language model approach to keyphrase extraction. In *Proceedings of the ACL 2003 workshop on Multiword expressions: analysis, acquisition and treatment*. Retrieved from http://acl.ldc.upenn.edu/W/W03/W03-1805.pdf.

Turney, P. D., & Littman, M. L. (2003). Measuring praise and criticism: Inference of semantic orientation from association. *ACM Transactions on Information Systems*, *21*, 315-346.

Wiebe, J., Bruce, R., & O’Hara, T. (1999). “Development and use of a gold standard data set for subjectivity classifications”. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics (ACL-99)*, (pp. 246–253). Retrieved from http://acl.ldc.upenn.edu/P/P99/P99-1032.pdf.

Wierzbicka, A. (1980). *Lingua mentalis: the semantics of natural language*. New York:Academic Press.

Wilson, T., Wiebe, J., & Hoffmann, P. (2005). Recognizing contextual polarity in phrase-level sentiment analysis. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, (pp. 347-354). Vancouver:ACL. Retrieved from http://www.aclweb.org/anthology/H/H05/H05-1044.

Wilson, T., Wiebe, J., & Hwa, R. (2006). Recognizing strong and weak opinion clauses. *Computational Intelligence*, *22*, 73-99.

Wilson, T., Wiebe, J., & Hwa, R. (2004). Just how mad are you? Finding strong and weak opinion clauses. In *Proceedings of the Nineteenth National Conference on Artificial Intelligence (AAAI-2004).* Retrieved from http://homepages.inf.ed.ac.uk/twilson/pubs/hltemnlp05.pdf.

Witten, I. H., & Frank, E. (2005). *Data mining: Practical machine learning tools and techniques*, 2nd Edition. San Francisco:Morgan Kaufmann.

Wright, S. (2001). *Internally caused and externally caused change of state verbs*. (Retrieved from ProQuest Dissertations & Theses).

Wu, Y., Oard, D., & Sobroff, I. (2006). An exploratory study of the W3C mailing list test collection for retrieval of emails with pro/con arguments. Presented at *The Third Conference on Email and Anti-Spam*, July 27-28, 2006, Mountain View, California. Retrieved from www.ceas.cc/2006/26.pdf.

Yi, J., Nasukawa, T., Bunescu, R., & Niblack, W. (2003). Sentiment Analyzer: Extracting Sentiments about a Given Topic using Natural Language Processing Techniques. In *ICDM '03: Proceedings of the Third IEEE International Conference on Data Mining*, (pp. 427-434). Washington: IEEE Computer Society.