DEPARTMENT OF COMPUTER SCIENCE
University of Pittsburgh
Technical Report TR-02-100
Learning Subjective Language

by
Janyce Wiebe, Theresa Wilson,
Rebecca Bruce, Matthew Bell, and Melanie Martin

Abstract

Subjectivity in natural language refers to aspects of language used to express opinions, evaluations, and speculations (Banfield, 1982; Wiebe, 1994). There are numerous NLP applications for which subjectivity analysis is relevant, including information extraction and text categorization. The goal of this work is learning subjective language from corpora. The first part of the paper explores annotating subjectivity at different levels (expression, sentence, document) and producing annotated corpora. In the second part of the paper, clues of subjectivity are generated and tested, including low-frequency words, collocations, and adjectives and verbs identified using distributional similarity. The third part of the paper examines the features working together in concert. The features, generated from different datasets using different procedures, exhibit consistency in performance in that they all do better and worse on the same datasets. In addition, we show that the density of subjectivity clues in the surrounding context strongly affects how likely it is that a word is subjective, and give the results of an annotation study assessing the subjectivity of sentences with high-density features. Finally, the clues are used to perform opinion-piece recognition (a type of text categorization and genre detection), to demonstrate the utility of the knowledge acquired in this paper.

1 Introduction

Subjectivity in natural language refers to aspects of language used to express opinions, evaluations, and speculations (Banfield, 1982; Wiebe, 1994). *Subjectivity tagging* is distinguishing sentences used to present opinions and other forms of subjectivity (*subjective sentences*) from sentences used to objectively present factual information (*objective sentences*). This task is especially relevant for news reporting and internet forums, in which opinions of various agents are expressed. There are numerous applications for which subjectivity tagging is relevant. Two are information retrieval and information extraction. Current extraction and retrieval technology focuses almost exclusively on the subject matter of documents. However, additional aspects of a document influence its relevance, including the evidential status of the material presented, and the attitudes expressed about the topic (Kessler et al., 1997). Knowledge of subjective language would also be useful in flame recognition (Spertus, 1997; Kaufer, 2000), email classification (Aone et al., 2000), intellectual attribution in text (Teufel and Moens, 2000), recognizing speaker role in radio broadcasts (Barzilay et al., 2000), review mining (Terveen et al., 1997), generation and style (Hovy, 1987), clustering documents by ideological point of view (Sack, 1995), answering questions from multiple perspectives, and any other application that would benefit from knowledge of how opinionated the language is, and whether or not the writer purports to objectively present factual material.

To employ subjectivity analysis in applications, good linguistic clues must be found. As with many pragmatic and discourse distinctions, existing lexical resources are not comprehensively coded for subjectivity. The goal of our current work is learning subjective language from corpora. This paper contributes to this goal by empirically examining subjectivity, generating and testing clues of subjectivity and contextual features, and using the knowledge gained to recognize subjective sentences and opinionated documents.

After background on subjectivity is provided (in Sections 2 and 3) and the datasets used in this work are identified (Section 4), Section 5 explores annotating subjectivity at multiple levels (expression, sentence, document) and producing annotated corpora. Annotator agreement is analyzed to understand and assess the viability of such annotations.

In the next part of the paper (section 6), clues of subjectivity are generated and tested. A mixture of data is used: Wall Street Journal and newsgroup data; expression-level as well as document-level annotations; and our manual annotations as well as existing annotations in the Wall Street Journal (*Editorials, Letters to the Editor*, *Arts & Leisure*, and *Viewpoints*). A variety of types of features are explored.

An important feature for subjectivity is the set of words that appear just once in the corpus. A term in the literature for such words is *hapax legomena*; here, they will be referred to as *unique* words. The set of all unique words is a feature with high frequency and significantly higher precision than baseline (Section 6.2).

Many clues of subjectivity are collocations rather than single lexemes. We demonstrate a straightforward method for automatically identifying collocational clues of subjectivity in texts (section 6.3). The method is first used to identify collocations composed of fixed sequences of words which, when they appear together, tend to be subjective. These include expressions such as *of the century* and *get out of here*. Interestingly, many of these collocations include non-content words that are typically on the stop lists of many NLP systems (e.g., *of, the, get, out, here* in the above examples). This paper shows the usefulness of these stop words for subjectivity recognition.

As argued by Smadja (1993), it is important to identify more general forms of collocations. This paper addresses one unusual type of generalization: one or more positions in the collocation may be filled by any word that is unique in the test data. Rather than discarding unique words, we provide evidence that they are informative for subjectivity recognition, especially in the context of collocations: the collocational features with unique fillers are the highest-precision feature we generate.

Unigram lexical features are also important. We identify adjective and verb features using the results of a method for clustering words according to distributional similarity (Lin, 1998) (section 6.4). The method uses existing annotations in the Wall Street Journal as training data; much similar data is freely available on-line. Distributional similarity is often used in NLP to construct semantic resources from corpora. The hypothesis behind its use in this paper is that two words may be distributionally similar because they are both potentially subjective (e.g., *tragic*, *sad*, and *poignant* are identified from *bizarre*). Another reason to use distributional similarity is to improve estimates of unseen events. That is also an aspect of its use in this work: words are selected or discarded based on the precision of it together with its N most similar neighbors.

In the third part of the paper, we examine the features working together in concert (section 7). The features, generated from different datasets using different procedures, exhibit an important consistency in performance in that they perform better and worse on the same datasets (section 7.1). In addition, we find that the density of subjectivity clues in the surrounding context strongly affects how likely it is that a word is subjective (Sections 7.2 and 7.3) and provide the results of an annotation study assessing the subjectivity of sentences with high-density features (Section 7.4). Finally, we use the clues together to perform opinion-piece recognition (a type of text categorization and genre detection), to demonstrate the utility of the knowledge acquired in this paper. Using an instance-based learning algorithm and leave-one-out cross validation, the system achieves significant results.

The final part of the paper includes a discussion of related work (section 8), and conclusions and plans for future work (section 9). The sentences included in the annotation study of Section 7.4 are given in the appendix, together with the tags assigned by the judges.

This paper extends and integrates the research reported in (Wiebe et al., 2001b; Wiebe et al., 2001a; Wiebe, 2000).

We will make our features available to other researchers, so they can experiment with subjective language in their applications.

2 Subjectivity

Sentence (1) is an example of a simple subjective sentence, and (2) is an example of a simple objective sentence: [1]

(1) At several different layers, it's a fascinating tale.
(2) Bell Industries Inc. increased its quarterly to 10 cents from 7 cents a share.

The main types of subjectivity are:

- *Evaluation.* This category includes emotions such as hope and hatred as well as evaluations, judgments, and opinions. Examples of expressions involving positive evaluation are *enthused*, *wonderful*, and *great product!* Examples involving negative evaluation are *complained*, *you idiot!* and *terrible product.*

- *Speculation.* This category includes anything that removes the presupposition of events occurring or states holding, such as speculation and uncertainty. Examples of speculative expressions are *speculated* and *maybe.*

Following are examples of strong negative evaluative language from a corpus of Usenet newsgroup messages:

(3a) I had in mind your facts, buddy, not hers.
(3b) Nice touch. "Alleges" whenever facts posted are not in your persona of what is "real".

Following is an example of opinionated, editorial language, taken from an editorial in the Wall Street Journal:

(4) We stand in awe of the Woodstock generation's ability to be unceasingly fascinated by the subject of itself.

Sentences (5) and (6) illustrate the fact that sentences about speech events may be subjective or objective:

(5) Northwest Airlines settled the remaining lawsuits filed on behalf of 156 people killed in a 1987 crash, but claims against the jetliner's maker are being pursued, a federal judge said.
(6) "The cost of health care is eroding our standard of living and sapping industrial strength," complains Walter Maher, a Chrysler health-and-benefits specialist.

[1] The term *subjectivity* is due to Ann Banfield (1982). For references to work on subjectivity, please see (Banfield, 1982; Fludernik, 1993; Wiebe, 1994; Stein and Wright, 1995).

In (5), the material about lawsuits and claims is presented as factual information, and a federal judge is given as the source of information. In (6), in contrast, a complaint is presented. An NLP system performing information extraction on (6) should not treat the material in the quoted string as factual information, with the complainer as a source of information, whereas a corresponding treatment of sentence (5) would be appropriate.

Subjective sentences often contain individual expressions of subjectivity. Examples are *fascinating* in (1), and *eroding*, *sapping*, and *complains* in (6). The following paragraphs mention aspects of subjective expressions that are relevant for NLP applications.

First, although some expressions, such as *!*, are subjective in all contexts, many, such as *sapping* and *eroding*, may or may not be subjective, depending on the context in which they appear. A *potential subjective element* (*PSE*) is a linguistic element that may be used to express subjectivity. A *subjective element* is an instance of a potential subjective element, in a particular context, that is indeed subjective in that context (Wiebe, 1994).

Second, a subjective element expresses the subjectivity of a *source*, who may be the writer or someone mentioned in the text. For example, the source of *fascinating* in (1) is the writer, while the source of the subjective elements in (6) is Maher. In addition, a subjective element has a *target*, i.e., what the subjectivity is about or directed toward. In (1), the target is a tale; in (6), the target of Maher's subjectivity is the cost of health care. These are examples of *object-centric subjectivity*, which is about an object mentioned in the text (other examples: "I love this project"; "The software is horrible"). Subjectivity may also be *addressee-oriented*, i.e., directed toward the listener or reader (e.g., "You are a sweetheart").

Third, there may be multiple subjective elements in a sentence, possibly of different types and attributed to different sources and targets. For example, in (4), subjectivity of the Woodstock generation is described (specifically, its fascination with itself). In addition, subjectivity of the writer is expressed (e.g., "we stand in awe"). Finally, PSEs may be complex expressions such as "village idiot", "powers that be", "You" *NP*, and "What a" *NP*. There is a great variety of such expressions, including many studied under the rubric of idioms (see, for example, (Nunberg et al., 1994)).

Knowledge of subjective language would be useful in NLP applications that would benefit from knowledge of how opinionated the language is and whether or not the writer purports to objectively present factual material. Genre detection/text categorization is one area where knowledge of subjectivity would be useful. Examples include filtering inflammatory messages from newsgroups and listservs (Spertus, 1997; Kaufer, 2000), and, in customer relationship management, recognizing customer complaints expressed in email.

Subjectivity analysis would be useful in applications desiring to filter subjective language from their input documents. Information extraction and question answering systems seeking facts could restrict their attention to sentences that at least purport to present facts. For a multi-document question answering or summarization system attempting to reconcile factual inconsistencies, the ability to recognize subjective language would allow it to avoid attempting to reconcile a proposition presented as a fact with one presented as an opinion.

Knowledge of subjective language would also be useful for segmenting documents into factual and opinionated segments. Many news reports, for example, are composed of segments presenting facts, and segments presenting various opinions about and reactions to the reported events. A good strategy for a summarization system might be to segment the text into such segments and cluster them to find the main points of view (the *supporting groups* of (Bergler, 1992)) for inclusion in the summary. Similarly, for the purpose of intellectual attribution (Teufel and Moens, 2000), it would be useful to recognize segments presenting others' points of view. For question answering, questions may ask for answers from multiple perspectives (e.g., What opinions are being expressed in the world press about President Bush's Axis of Evil comments?). The ability to recognize document segments presenting opinions about the topic would be needed by such a system.

With colleagues, we have performed preliminary experiments with positive results in filtering subjectivity to increase the precision of information extraction[2] and in producing summaries balanced by point of view[3]. In addition, this paper presents positive results for recognizing opinionated documents. In addition to our efforts, we hope that the work presented in this paper will encourage others to experiment with subjective language in their applications.

3 Previous Work on Subjectivity Tagging

In previous work (Wiebe et al., 1999; Bruce and Wiebe, 1999), a corpus of sentences from the Wall Street Journal Treebank Corpus (Marcus et al., 1993) was manually annotated with subjectivity classifications by multiple judges. The judges were instructed to classify a sentence as subjective if it contains a significant expression of subjectivity, attributed to either the writer or someone mentioned in the text, and to classify the sentence as objective, otherwise. The judges rated the certainty of their answers on a scale from 0 to 3. Agreement was summarized in terms of Cohen's κ (Cohen, 1960), which compares the total probability of agreement to that expected if the taggers' classifications were statistically independent (i.e., "chance agreement"). After two rounds of tagging by three judges, an average pairwise κ value of 0.69 was achieved on a test set. For the sentences rated certain by the judges (rating 2 or 3), the average pairwise κ value is 0.88.

The EM learning algorithm was used to produce corrected tags representing the consensus opinions of the taggers (Goodman, 1974; Dawid and Skene, 1979). An automatic system to perform subjectivity tagging was developed using the new tags as training and testing data. In 10-fold cross validation experiments, a probabilistic classifier obtained an average accuracy on subjectivity tagging of 72.17%, more than 20 percentage points higher than a baseline accuracy obtained by always choosing the more frequent class. Five part-of-speech features, two lexical features, and a paragraph feature were used. Interestingly, the system performs better on the sentences for which the judges are certain. If only those sentences for which all judges rated their certainty as 2 or 3 are considered, the system's average accuracy across folds rises to 81.5%. These results suggested to us that there are clues of subjectivity that might be learned automatically from text. This motivated the work reported in the current paper.

[2] with Ellen Riloff

[3] with Regina Barzilay

4 Datasets Used in this Paper

For reference, this section identifies all of the datasets used in this paper.

First, **WSJ-SE1** and **WSJ-SE2** are each composed of 500 sentences from the Wall Street Journal Treebank corpus (Marcus et al., 1993). These are the datasets used in (Wiebe et al., 1999; Bruce and Wiebe, 1999). The concatenation of the two datasets will be referred to as **WSJ-SE**. The annotations of this data used in this paper are expression-level, subjective-element annotations by two judges, **M** and **D**, as described in Section 5.3. **WSJ-SE1-M** will refer to M's annotations of WSJ-SE1; **WSJ-SE1-D** will refer to D's annotations of WSJ-SE1; **WSJ-SE2-M** will refer to M's annotations of WSJ-SE2; and **WSJ-SE2-D** will refer to D's annotations of WSJ-SE2.

Corpus **NG** is a corpus of 1140 Usenet newsgroup messages, balanced among the categories alt, sci, comp, and rec in the Usenet hierarchy. **NG-SE** is a 15413-word subset of **NG**, and **NG-FE** is a 362-message, 88210-word subset of **NG**. NG-SE and NG-FE do not overlap. Corpus **NG-SE** is annotated by one judge, **M**, for subjective elements, as described in section 5.3. Corpus **NG-FE** is annotated at the document level for flames, as described in section 5.2, and annotated at the expression-level for flame elements by two judges, **MM** and **R**, as described in Section 5.3.

Finally, eight files from the Treebank corpus are used: W9-2 (160,552 words), W9-20 (162,331 words), W9-21 (147,272 words), W9-23 (159,535 words), W9-4 (161,380 words), W9-10 (161,055 words), W9-22 (159,776 words), and W9-33 (158,768 words). Sometimes they are used individually and sometimes they are grouped into two datasets of four files each: **OP1** is the concatenation of W9-4, W9-10, W9-22, and W9-33 and **OP2** is the concatenation of W9-2, W9-20, W9-21, and W9-23. OP1 contains a total of 640,975 words and OP2 contains a total of 629,690 words. Neither WSJ-SE1 nor WSJ-SE2 is included in OP1 or OP2.

The opinion-piece document level annotations of OP1 and OP2 have been manually refined, as described in section 5.4.

In addition, all instances of all of the PSEs described in Section 6 in OP1 and OP2 have been automatically identified. All training to define the PSE instances in OP1 was performed on data separate from OP1, and all training to define the PSE instances in OP2 was performed on data separate from OP2.

All corpora have been stemmed (Karp et al., 1992) and part-of-speech tagged (Brill, 1992).

5 Annotations and Observations

This section introduces the idea of annotating subjectivity at multiple levels (document, sentence, and expression), presents the results of annotations studies, and examines expression-level annotations to learn about subjectivity. Note that an additional annotation study is presented later in section 7.4.

5.1 Choices in Annotation

In expression-level annotation, the judges first identify the sentences they believe are subjective. They next identify the subjective elements in those sentences, i.e., the expressions they feel are responsible for the subjective classification.[4] For example (subjective elements are in parentheses):

> They paid (yet) more for (really good stuff).
> (Perhaps you'll forgive me) for reposting his response.

Ultimately, we would like to recognize all of the subjective elements in a text and their types, targets, and sources. However, both manual and automatic tagging at this level are difficult because the tags are very fine-grained, and there is no predetermined classification unit: a subjective element may be a single word or a large expression. In this work, we use subjective-element annotations as training data, examining them to form hypotheses, and mining them for features. We do not use the subjective-element annotations as test data in our evaluations.

Document-level subjectivity annotations are text categories of which subjectivity is a key aspect. We use three text categories: editorials (Kessler et al., 1997), reviews, and "flames", i.e., hostile messages (Spertus, 1997; Kaufer, 2000). For ease of discussion, we group editorials and reviews together under the term *opinion pieces.*

There are benefits to using such document-level annotations. First, they are directly relevant for many applications, such as filtering hostile messages and mining reviews from Internet forums. Second, they contain existing annotations we can exploit, such as editorials and arts reviews, identified as such by newspapers, as well as on-line product reviews accompanied by formal numerical ratings (for example, 4 on a scale from 1 to 5).

However, a challenging aspect of such data is that they are noisy. On the one hand, opinion pieces contain objective sentences. Editorials contain objective sentences presenting facts supporting the writer's argument, and reviews contain sentences objectively presenting facts about the product. On the other hand, non-opinion pieces contain subjective sentences. News reports present reactions to and attitudes toward reported events (van Dijk 1988); they often contain segments starting with expressions such as *critics claim* and *supporters argue.* In addition, quoted-speech sentences in which individuals express their subjectivity are often included (Barzilay et al., 2000). This noisiness of opinion-piece data must be considered when such data is used for training and testing.

Sentence-level annotations are also an important level of analysis. The sentence provides a pre-specified classification unit and, while sentence-level judgments are not as fine-grained as subjective-element judgments, they do not involve the large amount of noise we face with document-level annotations. This paper presents a sentence-level annotation study to assess the subjectivity of sentences identified by our system using density features (section 7.4). We anticipate that sentence-level analysis will be increasingly important in future work.

[4] We are grateful to Aravind Joshi for suggesting this level of annotation.

5.2 Flame Annotations

Flames in newsgroups or listservs are inflammatory messages. Later in this paper expression-level flame annotations are used as training data. Those annotations are based on a document-level flame annotation study, the results of which are presented in this section.

In this study, newsgroup messages were assigned the tags *flame* or *not-flame*. The corpus is corpus NG, which, recall, consists of 1140 Usenet newsgroup messages, balanced among the categories alt, sci, comp, and rec in the Usenet hierarchy. The corpus was divided, preserving the category balance, into a training set of 778 messages and a test set of 362 messages. (The NG-FE subset of NG is exactly this test set of 362 messages.)

The annotators were instructed to mark a message as a flame if the "main intention of the message is a personal attack, containing insulting or abusive language." A number of policy decisions are made in the instructions, dealing primarily with included messages (part or all of a previous message that is included in the current message). Some additional issues addressed in the instructions are to whom the attack is directed, nonsense, sarcasm, humor, rants, and raves.

During the training phase, two annotators, **MM** and **R**, participated in multiple rounds of tagging, revising the annotation instructions as they proceeded. During the testing phase, MM and R independently annotated the test set, achieving a κ value on these messages of 0.69. A third annotator, **L**, trained on 492 messages from the training set, and then annotated 88 of the messages in the test set. The pairwise κ values on this set of 88 are: MM & R: 0.80; MM & L: 0.75; R & L: 0.79; for an average pairwise κ of 0.78.

Spertus (1997) also performed flame annotation. As in our data, the distribution of flames to non-flames in her data is highly skewed in favor of non-flames. In both hers and our studies, the percentage agreement results are high, as expected with such a skewed distribution. Spertus reports 98% agreement on non-inflammatory messages and 64% agreement on inflammatory messages. Our percentage agreement results are comparable. For example, the percentage agreement between MM and R on all 362 messages in the testing phase was 92% and the pairwise percentage agreement on the set of 88 messages was MM & R: 93%; MM & L: 91%; R & L: 91%; for an average pairwise percentage agreement of 92%.

5.3 Subjective-Element Annotations

This section analyzes subjective-element annotations performed on the datasets WSJ-SE1, WSJ-SE2, NG-FE, and NG-SE.

Recall that WSJ-SE1 and WSJ-SE2 were manually annotated at the sentence level as described in (Wiebe et al., 1999; Bruce and Wiebe, 1999).

For this paper, two annotators (D and M) were asked to identify the subjective elements in these datasets. Specifically, the taggers were given the subjective sentences identified in the previous study, and asked to put brackets around the words they believe cause the sentence to be classified as subjective. A single round of tagging was performed, without training or any communication between annotators.

Note that inflammatory language is a kind of subjective language. NG-FE is a subset of

	All Words	Nouns	Verbs	Modals	Adj's	Adverbs	Det's
NG-FE	0.4657	0.5213	0.4571	0.4008	0.5011	0.3576	0.4286
WSJ-SE1	0.4228	0.3999	0.4235	0.6992	0.6000	0.4328	0.2661
WSJ-SE2	0.3703	0.3705	0.4261	0.4298	0.4294	0.2256	0.1234

Table 1: κ values for word agreement

the Usenet newsgroup corpus used in the document-level flame-annotation study described in section 5.2 (specifically, it is the 362-message test set of that study). For this study, R and MM were asked to identify the *flame elements* in NG-FE. Flame elements are the subset of subjective elements that are perceived to be inflammatory. R and MM were asked to do this in all 362 messages, even those not identified as flames, because messages that were not judged to be flames at the message (document) level may contain some individual inflammatory phrases. Again, a single round of tagging was performed, without training or communication between annotators.

In addition, tagger M performed subjective-element annotations on the 15413-word newsgroup corpus NG-SE.

In datasets WSJ-SE and NG-SE, the taggers were also asked to specify one of five subjective element types: $e+$ (positive evaluative), $e-$ (negative evaluative), $e?$ (some other type of evaluation), u (uncertainty), and o (none of the above), with the option to assign multiple types to an instance.

5.3.1 Agreement Among Taggers

There are techniques for analyzing agreement when annotations involve segment boundaries (Litman and Passonneau, 1995; Marcu et al., 1999), but our focus in this paper is on words. Thus, our analyses are at the word level: each word is classified as either appearing in a subjective element or not. Punctuation and numbers are excluded from the analyses.

Table 1 provides κ values for word agreement for the flame-element annotations in NG-FE and for the subjective-element annotations in WSJ-SE1 and WSJ-SE2. As mentioned above, only a single round of tagging was performed, without training. The agreement results reflect this fact: they are not high by Krippendorf's (1980) suggested standards (κ values over .80 support definite conclusions, while those between .67 and .80 support tentative conclusions). As discussed in (DiEugenio, 2000), Krippendorf (p. 147) observes that "if it is an exploratory study without serious consequences [the] level of reliability may be relaxed considerably, but it should not be so low that the findings can no longer be taken seriously." In our case, the agreement is much higher than that expected by chance. As seen in later sections of the paper, these annotations proved very useful for exploring hypotheses and generating features.

Observation of the data suggests one significant source of disagreement between taggers, exhibited in the subjective elements identified for a single sentence in WSJ-SE1:

D: (e+ played the role well) (e? obligatory ragged jeans a thicket of long hair

	All Words	Nouns	Verbs	Modals	Adj's	Adverbs	Det's
WSJ-SE1	0.53	0.16	0.38	0.49	0.81	0.38	0.00
WSJ-SE2	0.48	0.00	0.42	0.40	0.81	0.77	0.40

Table 2: κ values for subjective-element type agreement in the WSJ data

and rejection of all things conventional)
M: played the role (e+ well) (e? obligatory) (e- ragged) jeans a (e? thicket) of long hair and (e- rejection) of (e- all things conventional)

Judge D consistently identifies entire phrases as subjective, while judge M prefers to select discrete lexical items. Similarly, within the flame-element data, there are many instances where both taggers identify the same segment of a sentence as forming a subjective element but disagree on the boundaries of that segment. For example:

R: (classic case of you deliberately misinterpreting my comments)
MM: (you deliberately misinterpreting my comments)

These patterns of partial agreement are also evident in the κ values for words from specific syntactic categories (see Table 1 again). In the WSJ data, agreement on determiners is particularly low because they are often included as part of a subjective element by tagger D but excluded from subjective elements by tagger M.

In the WSJ experiments, the taggers most frequently agree on the selection of modals and adjectives, while in the flame experiment, agreement was highest on nouns and adjectives. The high agreement on adjectives in both genres is consistent with results from other work (Bruce and Wiebe, 1999; Wiebe et al., 1999), but high agreement on nouns in the flame data versus high agreement on modals in the WSJ data was unexpected. In the future, we plan to examine the use of nouns and modals in the different datasets to try to understand these differences.

Turning to subjective-element type (e.g, *e+* for positive evaluative), agreement is comparable to that for word agreement. Table 2 shows κ values for type agreement, with the three evaluative tags combined into one tag. Only the words classified as belonging to subjective elements by both taggers are considered in the analysis. In cases where one or both of the annotators assigned multiple tags, their tags are counted as matching if they have a type in common. All multiply assigned tags are *uncertain* in combination with an *evaluative* tag. They are not common: each tagger assigned multiple tags to fewer than 7% of the subjective instances.

There were few assignments to type *other*: 1.7% of M's tags in NG-SE, and less than 1% of both M's and D's tags in WSJ-SE.

The question arises, when a word appears in multiple subjective elements, are those subjective elements all the same type? Table 3 shows that a significant portion are used in more

WSJ-SE-M		WSJ-SE-D		NG-SE	
MultInst	MultType	MultInst	MultType	MultInst	MultType
413	.17	378	.16	571	.29

Table 3: Word-POS-Types used in multiple types of subjective elements

			WSJ-SE						NG-FE			
	D		M		Agree		Agree		R		MM	
	Num	P	Num	P	Num	P	Num	P	Num	P	Num	P
all	18341	.07	18341	.08	16857	.04	15413	.15	86279	.01	88210	.02
unique	2615	.14	2615	.20	2522	.15	2348	.17	5060	.07	4836	.03
unique/ all		2.0		2.5		3.75		1.13		7.0		1.5

Table 4: Proportions (P) of unique words in subjective elements

than one type. Each item considered in the table is a word-POS pair that appears more than once in a subjective element (instances assigned multiple types are excluded from this analysis, to avoid counting things twice). The figures shown are the total number of word-POS items that appear more than once (the columns labeled *MultInst*) in subjective elements, and the proportion of those items that appear in more than one type of subjective element (the columns labeled *MultType*). These results highlight the need for contextual disambiguation. For example, one thinks of *great* as a positive evaluative term, but its polarity depends on the context; it can be used negatively evaluatively in a context such as "Just great."

Further analysis of subjectivity type can be found in (Hatzivassiloglou and McKeown, 1997; Hatzivassiloglou and Wiebe, 2000; Wiebe et al., 2001a). Subjectivity type is not explored further in this paper.

5.3.2 Uniqueness

Based on previous work (Wiebe et al., 1998), we hypothesized that low-frequency words are associated with subjectivity. It appears that people are creative when they are being opinionated. Table 4 provides evidence that the number of unique words (words that appear just once in the corpus) in subjective elements is higher than expected. The first row gives information for all words and the second gives information for words that appear just once. The figures in the *Num* columns are total counts, and the figures in the *P* columns are the proportions that appear in subjective elements. The *Agree* columns give information for the subset of the corresponding dataset upon which the two annotators agree. For example, in WSJ-SE, D and M agree on 16857 of the words (row 1, column 5). The proportion of those words in subjective elements is .04 (row 1, column 6). There are 2615 unique words in WSJ-SE (row 2, columns 1 and 3). D included 14% (column 2) and M included 20% (column 4) of them in subjective elements. Of those 2615 unique words, 2522 of them are in the set of words upon which D and M agree (column 5). 15% of those are in subjective elements (column 6).

The bottom row gives ratios of the proportion figures in row 2 to the proportion figures in row 1. This row shows that the proportion of unique words that are subjective is higher than the proportion of all words that are subjective. In all cases but the *Agree* figure for flame elements, the difference in proportions is dramatic: in WSJ-SE, the proportion of unique words is 200% higher in subjective elements in D's annotations; 250% higher in M's annotations; and 375% higher for the agreement subset. In the flame-element data (NG-FE), the performance is less consistent between taggers. While R's proportion of unique words in flame elements is fully 700% higher, MM's is only 150% higher. Their agreement subset is only 13% higher. Even though this figure is not as high, the difference as well as all the other differences are highly statistically significant.

The disproportionate number of low-frequency words in subjective elements has important implications for learning subjective language.

First, we do not want to discard low-frequency words, as many NLP systems do, because we do not want to discard a valuable source of information. However, statistical techniques are not reliable for learning low-frequency features, so attempting to learn individual lexical clues would be problematic. Thus, the strategy in this work is to identify sets of words and phrases, rather than individual ones; specifically, we seek sets with precision greater than the baseline precision, with more than a handful of occurrences. We identify a number of such sets in the second part of the paper (Section 6) and, in the third part of the paper, show they can be used together to improve recognition of opinions.

Second, the higher-than-expected proportions of unique words raises the possibility of using uniqueness as an informative feature of subjectivity. This is explored in section 6.2.

5.4 Opinion-Piece Document Annotation for Testing

Document-level opinion-piece data is used throughout this paper as test as well as training data. The class *opinion-piece* is the union of *Editorials, Letters to the Editor*, *Arts & Leisure*, and *Viewpoints* in the Wall Street Journal. An inspection of some data revealed that some editorials and reviews are not marked as such. For example, there are articles whose purpose is to present an argument rather than cover a news story, but they are not explicitly labeled as editorials by the Wall Street Journal. Thus, the opinion-piece data has been manually refined. The annotation instructions are simply to identify any additional opinion pieces that are not marked as such. To test the reliability of this annotation, two judges independently annotated two Wall Street Journal files, W9-22 and W9-33, each approximately 160K words. This is an "annotation lite" task: with no training, the annotators achieved κ values of 0.94 and 0.95, and each spent an average of three hours per WSJ file. The agreement data for W9-22 is given in Table 5 in the form of a contingency table.

All of the files composing OP1 and OP2 were manually annotated as described in the previous paragraph. Recall that OP1 and OP2 are each the concatenation of four Wall Street Journal Treebank files, for a total of 640,975 words and 629,690 words, respectively.

		Tagger 2		
		Op	*Not Op*	
Tagger 1	*Op*	$n_{11} = 23$	$n_{12} = 0$	$n_{1+} = 23$
	Not Op	$n_{21} = 2$	$n_{22} = 268$	$n_{2+} = 270$
		$n_{+1} = 25$	$n_{+2} = 268$	$n_{++} = 293$

Table 5: Opinion-piece agreement in W9-22

6 Generating and Testing Features

6.1 Introduction

As discussed above in section 5.3.2, in this work, features are sets of words and phrases, rather than individual words and phrases. A number of different procedures are used to create them, and both the subjective-element data and the opinion-piece data are used for training. When the opinion-piece data is used for training, the existing classifications are used (*Editorials, Letters to the Editor*, *Arts & Leisure*, and *Viewpoints*). When the opinion-piece data is used for testing, the manually refined classifications are used (i.e., the existing annotations augmented with the manually identified opinionated documents, as described in section 5.4).

The precision of a set S with respect to opinion pieces is:

$$prec(S) = \frac{\text{number of instances of members of S in opinion pieces}}{\text{total number of instances of members of S in the data}}$$

As discussed above in Section 5.1, while there are good reasons to use opinion-piece data for training and testing, we cannot expect absolutely high precisions. First, as mentioned in that section, opinion pieces contain objective sentences and non-opinion pieces contain subjective sentences. For example, in WSJ-SE, which has been annotated at the sentence and document levels, 70% of the sentences in opinion pieces are subjective and 30% are objective. In non-opinion pieces, 44% of the sentences are subjective and only 56% are objective. Second, the distribution of opinions and non-opinions is highly skewed in favor of non-opinions. For example, in table 5, tagger 1 classifies only 23 of 293 articles as opinion pieces. Finally, we are assessing PSEs, which are only potentially subjective; many have objective as well as subjective uses.

In this work, increases in precision over a baseline precision are used as evidence that promising sets of PSEs have been found. The main baseline for comparison that we use is the number of word instances in opinion pieces, divided by the total number of word instances:

$$baselinePrecision = \frac{\text{number of word instances in opinion pieces}}{\text{total number of word instances}}$$

	W9-10		W9-22		W9-33		W9-04	
	freq	+prec	freq	+prec	freq	+prec	freq	+prec
unique words	4763	+.16	4274	+.11	4567	+.11	4794	+.15
baseline precision		.18		.13		.14		.19

Table 6: Frequencies and increases in precision for words that appear exactly once

6.2 Unique Words as Features

In section 5.3.2 we showed that there are more than expected unique words in subjective elements. Here we provide evidence that unique words are informative for recognizing subjectivity. We extend our analysis of unique words by measuring their precision with respect to opinion-piece classifications.

We compare the precision of the set of unique words to the precision of the set of all words in the corpus. That is, the following is compared to the baseline precision ($basePrecision$):

$$prec(uniques) = \frac{\text{number of uniques in opinion pieces}}{\text{total number of uniques}}$$

Table 6 shows the increases over baseline realized by the set of unique words in the four WSJ datasets composing OP1, with numbers removed. Baseline precisions are listed at the bottom of the table. The *freq* columns give total frequencies. The *+prec* columns show the improvements in precision over baseline. For example, in W9-10, unique words have precision .34: .18 baseline plus an improvement over baseline of .16. In addition to achieving good improvements over baseline, *the set of unique words is a frequent feature.*

The question arises, how does corpus size affect the precision of the set of unique words? Presumably, uniqueness in a larger corpus is more meaningful than uniqueness in a smaller one. The results in Figure 1 provide evidence that it is. The Y axis in Figure 1 represents increase in precision over baseline and the X axis represents corpus size. Five graphs are plotted, one for the set of words that appear exactly once, one for the set of words that appear exactly twice ($freq2$), one for the set of words that appear exactly three times ($freq3$), etc.

To create corpora of various sizes, we concatenated the eight WSJ datasets composing OP1 and OP2 to create a single large corpus of over 1 million words. In this large corpus, 9% of the documents are opinion pieces. We selected random samples of various sizes from this combined corpus, preserving the 9% distribution of opinion pieces in each random sample. Each data point in figure 1 is an average over 25 random samples of the same size; samples of size $20, 40, \ldots, 2420, 2440$ documents were generated. Over the 25 random samples, the smallest corpus (containing 20 documents) averages 9,617 words. For the largest corpus (containing 2440 documents) the average size is 1,225,186 words.

As can be seen, the precision of unique and other low-frequency words increases with corpus size, with increases tapering off at the largest corpus size tested. Although not as dramatic, words with frequency 2 also realize a nice increase in precision over baseline. Even words of frequency 3, 4, and 5 show modest increases.

To help us understand low-frequency words in large as opposed to small datasets, we can consider the following analogy. With collectible trading cards, rare cards are the most valuable. However, looking in only a few packs of cards will not tell us if any of our cards are valuable. It is only by looking at many packs of cards that we realize which are the rare ones. It is only in samples of sufficient size that uniqueness is informative.

An interesting question is, what happens to the frequency of unique words as the corpus size gets larger? On the one hand, we expect more second occurrences as the corpus size increases. On the other hand, from Zipf's law we know that a large number of unique words can be expected in any corpus (Zipf, 1935). In Figure 2, the increasing line is the number of unique words in the corpus. The decreasing line is the proportion of the corpus made up of unique words. Together, Figures 1 and 2 show that precision and frequency continue to increase with corpus size, but the rates of these increases slow.

The results in this section suggest that, when working with a small test corpus (or when classifying individual documents as they are received), an NLP system using uniqueness features to recognize subjectivity should determine uniqueness with respect to the test data augmented with an additional store of (unannotated) data.

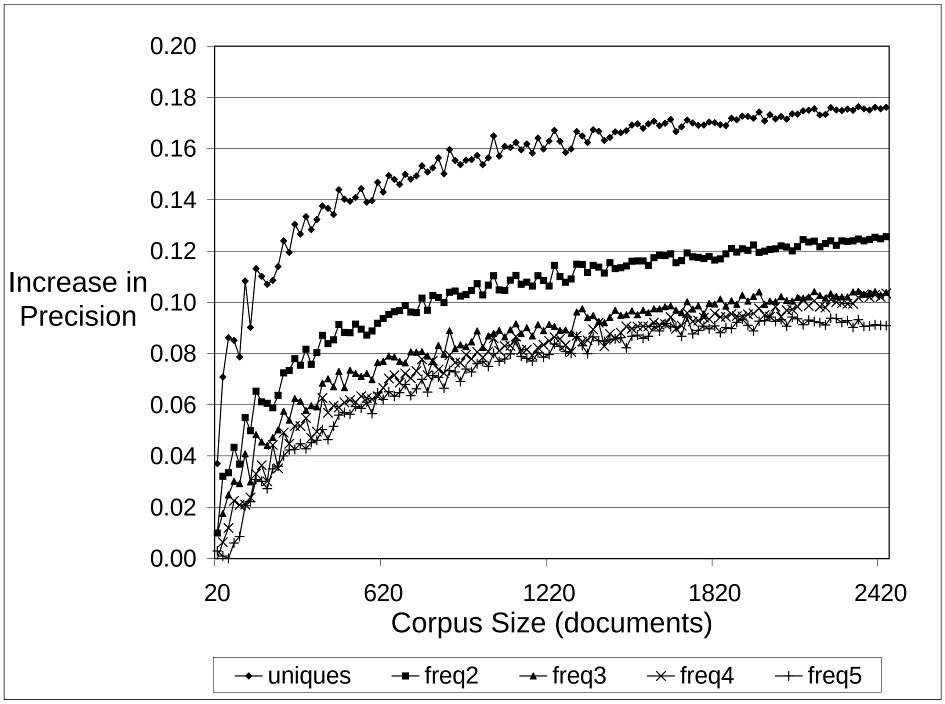

Figure 1: Precision of low-frequency words as corpus size increases

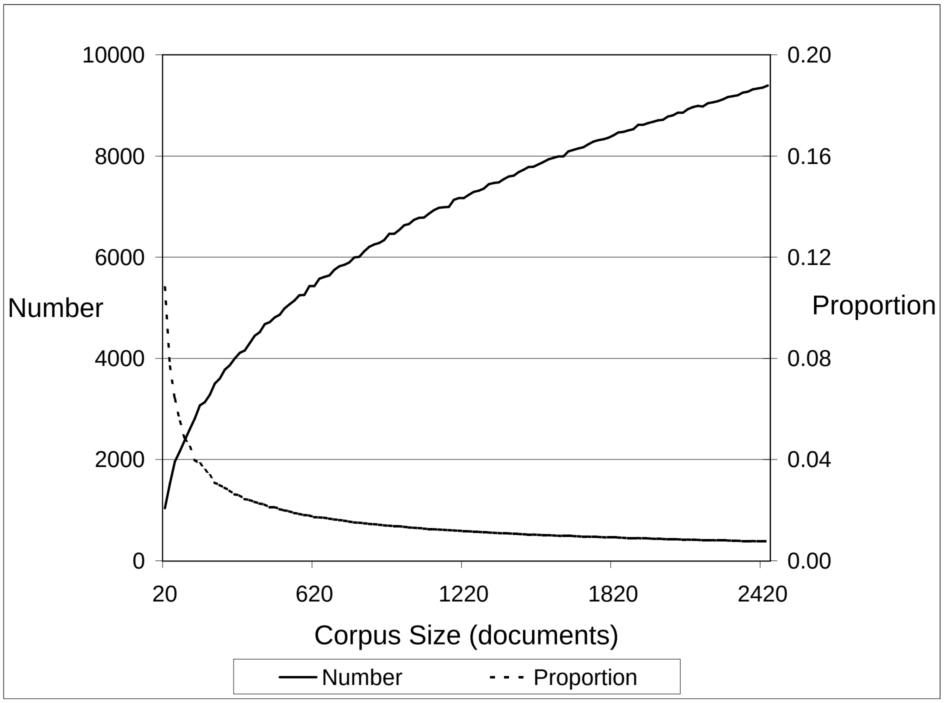

Figure 2: Number of words that are unique in WSJ-concat as corpus size increases

6.3 Identifying potentially subjective collocations from subjective-element (SE) and flame-element (FE) annotations

Much previous work in mining collocations from texts (e.g., (Smajda, 1993; Lin, 1999; Biber, 1993)) was directed at extending lexicographic resources to serve machine translation and word sense disambiguation. In this work, data is mined for potentially subjective collocations.

Collocations were selected from NG-SE, NG-FE, and WSJ-SE, using the union of the annotators' tags for the datasets tagged by multiple taggers (i.e., NG-FE and WSJ-SE). The following method is used to identify collocations. First, we extract all 1-grams, 2-grams, 3-grams, and 4-grams from the training data and calculate the precision of each. The precision of an n-gram is the number of subjective instances of that n-gram divided by the total number of instances of that n-gram. An instance of an n-gram is subjective if each word occurs in a subjective element. Each position is filled with a word|part-of-speech pair. An example of a 3-gram is: (in-*prep* the-*det* can-*noun*), which matches trigrams consisting of preposition *in* followed by determiner *the* ending with noun *can*.

Potentially subjective collocations were selected based on their precision, using two criteria. First, the precision of the n-gram must be at least 0.1. Second, the precision of the n-gram must be greater than the maximum precision of its constituents.

Specifically, let $(W1, W2)$ be a bi-gram consisting of consecutive words $W1$ and $W2$. $(W1,W2)$ is identified to be a potential subjective element if $prec(W1, W2) >= 0.1$ and:

$$prec(W1, W2) > max(prec(W1), prec(W2))$$

For trigrams, we extend the second condition as follows. Let $(W1, W2, W3)$ be a trigram consisting of consecutive words $W1$, $W2$, and $W3$. The condition is then:

$$prec(W1, W2, W3) > max(prec(W1, W2), prec(W3)) \text{ or}$$

$$prec(W1, W2, W3) > max(prec(W1), prec(W2, W3))$$

The selection of 4-grams is similar to the selection of 3-grams, comparing the 4-gram first with the maximum of the precisions of word $W1$ and trigram ($W2$, $W3$, $W4$) and then with the maximum of the precisions of trigram ($W1$,$W2$,$W3$) and word $W4$. The n-gram collocations identified as above will be called *fixed-n-grams*, i.e., they are fixed sequences of words of length n.

Note that Samuel et al. (1998) use a similar filter to eliminate redundant n-gram features for dialog act recognition: they eliminate n-grams that contain substrings with the same or better entropy scores.

It is common in studies of collocations to omit low-frequency words and expressions from analysis, because they give rise to invalid or unrealistic statistical measures (Church and Yarowsky, 1990). One example is the entropy measure used by Samuel et al. (1998), who also exclude low-frequency n-grams. However, various studies show that retaining low-frequency words and events is valuable for particular tasks. (Weeber et al., 2000) recognizes the importance of low-frequency words for information extraction and, in (Copeck et al., 2000),

	W9-10		W9-22		W9-33		W9-04	
	freq	+prec	freq	+prec	freq	+prec	freq	+prec
fixed-2-grams	1972	+.07	1933	+.04	1839	+.05	1840	+.07
ugen-2-grams	256	+.26	261	+.17	254	+.17	281	+.21
fixed-3-grams	243	+.09	214	+.05	262	+.05	213	+.08
ugen-3-grams	133	+.27	147	+.16	133	+.15	148	+.29
fixed-4-grams	17	+.06	12	+.18	14	-.07	18	+.15
ugen-4-grams	3	+.82	15	+.27	13	+.25	13	+.12
baseline precision		.18		.13		.14		.19

Table 7: Frequencies and increases in precision of collocations

unique words are included as one of a number of text categorization features. (Daelemans et al., 1999) cite a number of studies: (Bod, 1995) for parsing, (Collins and Brooks, 1995) for prepositional-phrase attachment, and (Dagan et al., 1997) for word sense disambiguation. In addition to using unique words as unigram clues for subjectivity (see Section 6.2 above), we have also found a valuable use for unique words in extracting and selecting *generalized collocations*.

To find and select generalized collocations, we first find every word that appears just once in the corpus and replace it with a new word, 'UNIQUE'. In essence, we treat the set of single-instance words as a single, frequently-occurring word. The above method for extracting and selecting n-grams is then used to obtain the potentially subjective collocations with positions filled by UNIQUE. Hereafter we will refer to these collocations as *ugen-n-grams* (**u**nique **gen**eralized-n-grams).

To test the ugen-n-grams extracted from the subjective-element training data (i.e., WSJ-SE,NG-SE, and NG-FE) using the method outlined above, we assess their precision with respect to opinion piece data. As with the training data, all unique words in the test data are replaced by "UNIQUE", so that, when matching a ugen-n-gram against the test data, the "UNIQUE" fillers match words that are unique in the *test* data. The test data are the four WSJ datasets composing OP1 (W9-10, W9-22, W9-33, and W9-04). Recall that WSJ-SE is not included in this data (or in OP2).

Once again, the baseline for comparison is the number of word instances in opinion pieces, divided by the total number of word instances.

Table 7 shows the results of testing the fixed-n-gram and the ugen-n-gram patterns on the four WSJ corpora. The *freq* columns give total frequencies, and the *+prec* columns show the improvements in precision from the baseline. The baseline precisions are given at the bottom of the table. The ugen-n-grams show large increases in precision.

These results show that the method used to extract and select potentially subjective n-grams is promising.

The question arises, how do the sets of instances extracted from the test data by the fixed-n-grams compare to the sets of instances extracted from the test data by the ugen-n-grams?

W9-10	2grams	3grams	4grams
intersecting instances	4	2	0
%overlap	0.0016	0.0049	0
W9-22			
intersecting instances	4	0	0
%overlap	0.0016	0	0
W9-33			
intersecting instances	0	0	0
%overlap	0	0	0
W9-04			
intersecting instances	0	0	0
%overlap	0	0	0

Table 8: Overlap between fixed-n-grams and ugen-n-grams, n=2,3,4

Are they relatively similar, or are the two applications of the method recognizing different PSEs?

To address this question, we examined their intersection. Taking all the potentially subjective n-gram instances identified in the test data, we compared the set of fixed-n-grams to the set of ugen-n-grams. Specifically, we checked for fixed-n-gram instances contained within ugen-n-gram instances. Contrary to our expectation, this intersection is very small. In the four datasets, there are a total of 8,577 fixed-n-grams instances. Only 59 of these, less than 1%, are contained within ugen-n-gram instances. This remarkably small intersection indicates that we have identified two different types of potentially subjective collocations.

To test the hypothesis that many fixed-n-gram collocations contain stop-list words, we counted how many of the collocations that have instances in the test data contain stop words, using the stop-word list available from the ACL NLP/CL Universe website (http://perun.si.umich.edu/ radev/u/d Many contain them. There are 978 fixed-n-gram and 186 ugen-n-gram patterns with instances in the test data. 35% of the fixed-n-gram patterns and 34% of the ugen-n-gram patterns contain one or more stop words from the list. Recent work by van der Wouden (2001) also supports the result that stop words can be an important component of certain types of collocations.

Examples of collocations are given in Tables 9, 10, and 11. Table 9 shows fixed-3-grams that appear in opinion pieces in at least two of the WSJ test datasets. In Tables 10 and 11, the collocations and the instances shown were randomly selected from those that appear in OP1 (both opinion and non-opinion pieces). 'U' stands for 'UNIQUE'.

Note that the instances in Tables 9, 10, and 11 are from the *test* data. The instances in the training data from which the ugen patterns were derived include different unique words.

a-*det* long-*adj* way-*noun*	a-*det* sort-*noun* of-*prep*
an-*det* example-*noun* of-*prep*	as-*prep* he-*pronoun* be-*verb*
be-*verb* in-*prep* the-*det*	be-*verb* it-*pronoun* that-*prep*
be-*verb* the-*det* case-*noun*	can-*model* do-*verb* be-*verb*
have-*verb* to-*prep* pay-*verb*	he-*pronoun* be-*verb* a-*det*
here-*adverb* are-*verb* some-*det*	in-*prep* the-*det* middle-*noun*
it-*pronoun* be-*verb* time-*noun*	it-*pronoun* should-*model* be-*verb*
of-*prep* the-*det* century-*noun*	one-*noun* of-*prep* his-*pronoun*
rest-*noun* of-*prep* the-*det*	rest-*noun* of-*prep* us-*pronoun*
seem-*verb* to-*prep* be-*verb*	should-*model* have-*verb* be-*verb*
some-*det* of-*prep* us-*pronoun*	the-*det* country-*noun* be-POS
the-*det* difference-*noun* between-*prep*	the-*det* kind-*noun* of-*prep*
the-*det* middle-*noun* of-*prep*	the-*det* need-*noun* to-*prep*
the-*det* other-*adj* hand-*noun*	the-*det* quality-*noun* of-*prep*
the-*det* rest-*noun* of-*prep*	to-*prep* be-*verb* the-*det*
to-*prep* do-*verb* so-*adverb*	to-*prep* say-*verb* about-*prep*.

Table 9: Fixed-3-Gram Collocations in Opinion Pieces in $\geq$ 2 WSJ Test Sets

6.4 Generating features from Document-Level Annotations Using Distributional Similarity

This section uses the opinion-piece annotations to expand our set of PSEs beyond those that can be derived from the subjective-element annotations.

The approach is based on *distributional similarity*, where words are judged to be more or less similar based on their distributional patterning in text.

Distributional similarity is most commonly used in NLP for two purposes: to create dictionaries and thesauri from corpora (see, for example, (Lin, 1998; Riloff and Jones, 1999)) and to smooth parameter estimates of rare or unseen events to improve syntactic or semantic disambiguation (see, for example, (Hindle, 1990; Dagan et al., 1994)). The procedure presented below for learning PSEs with distributional similarity involves both.

Many variants of distributional similarity have been used in NLP (see (Lee, 1999; Lee and Pereira, 1999) for comparisons of a number of methods). Dekang Lin's (1998) method is used in this work. In contrast to many implementations, which focus exclusively on verb-noun relationships, Lin's method incorporates a variety of syntactic relations. This is important for subjectivity recognition, because PSEs are not limited to verb-noun relationships. In addition, Lin's results are freely available.

Using his broad-coverage parser (Lin, 1994), Lin (1998) extracts dependency triples from text which consist of two words and the grammatical relationship between them: $(w1, relation, w2)$. To measure similarity between two words $w1$ and $w2$, $T(w1)$ and $T(w2)$ are identified, where $T(w)$ is the set of relation-word pairs correlated with w. The similarity $sim(w1, w2)$ between

Pattern	Instances
U-*adj* amount-*noun:*	copious amount; target amount
U-*adj* as-*IN:*	drastic as; perverse as; predatory as
U-*adj* debate-*noun:*	feverish debate; semantic debate
U-*adj* in-*IN:*	perk in; unsatisfying in; unwise in
U-*adj* political-*adj:*	humble political; repressive political
U-*adverb* U-*verb:*	adroitly dodge; crossly butter; unceasingly fascinate
U-*noun* amongst-*IN:*	dissenter amongst; interaction amongst; merriment amongst
U-*noun* back-*adverb:*	cutting back; hearken back
U-*noun* that-*IN:*	baloney that; revelation that; tijd that
U-*verb* U-*adverb:*	coexist harmoniously; flouncing tiresomely
U-*verb* his-*pronoun:*	badger his; permeate his; underprice his
ad-*noun* U-*noun:*	ad hoc; ad valorem
any-*DT* U-*adj:*	any navigational; any noxious; any unspent
any-*DT* U-*noun:*	any over-payment; any tapings; any write-off
are-*verb* U-*adj:*	are groundless; are invalid; are spirited
are-*verb* U-*noun:*	are escapist; are lowbrow; are resonance
but-*CC* U-*adj:*	but archaic; but fervent; but pathetic
but-*CC* U-*noun:*	but belch; but cirrus; but ssa
by-*IN* U-*verb:*	by overwork; by scattering; by utilize
different-*adj* U-*noun:*	different ambience; different subconferences
his-*pronoun* U-*noun:*	his fu; his pin; his tangiers
like-*IN* U-*noun:*	like hoffmann; like manute; like woodchuck
major-*adj* U-*noun:*	major contraction; major histo; major resellers
national-*adj* U-*noun:*	national commonplace; national yonhap
often-*adverb* U-*verb:*	often freak; often incite; often lapse
particularly-*adverb* U-*adj:*	particularly galling; particularly noteworthy
so-*adverb* U-*adj:*	so monochromatic; so overbroad; so permissive
this-*DT* U-*adj:*	this biennial; this inexcusable; this scurrilous
to-*TO* U-*adj:*	to judgmental; to preservative; to unconfirmed
your-*pronoun* U-*noun:*	your forehead; your manuscript; your popcorn

Table 10: Random sample of Ugen-2-Gram Collocations with at least 2 instances in OP1. For ugen-2-grams with more than 3 instances, the instances shown were randomly selected.

Pattern	Instances
U-*adj* and-*CC* U-*adj:*	arduous and raucous; obstreperous and abstemious
U-*noun* and-*CC* a-*DT:*	codification and a; loyalist and a
U-*noun* be-*verb* a-*DT:*	acyclovir be a; siberia be a
U-*noun* of-*IN* U-*noun:*	agglutination of lao; cradle of edita
U-*noun* of-*IN* its-*pronoun:*	outgrowth of its; repulsion of its
U-*noun* with-*IN* the-*DT:*	maestro with the; navajo with the
U-*verb* and-*CC* U-*verb:*	wax and brushed; womanize and booze
U-*verb* the-*DT* U-*noun:*	befall the coyote; impoverish the populace
U-*verb* to-*TO* a-*DT:*	cling to a; trek to a
are-*verb* U-*adj* of-*IN:*	are leery of; are mindful of
are-*verb* U-*adj* to-*TO:*	are opaque to; are suject to
a-*DT* U-*noun* U-*noun:*	a blast furnace; a companion jetty
a-*DT* U-*noun* and-*CC:*	a blindfold and; a rhododendron and
a-*DT* U-*noun* for-*IN:*	a propensity for; a watchword for
a-*DT* U-*verb* U-*noun:*	a jaundice ipo; a smoulder sofa
be-*verb* an-*DT* U-*noun:*	be an anachronism; be an uptick
it-*pronoun* be-*verb* U-*adverb:*	it be humanly; it be sooo
of-*IN* its-*pronoun* U-*noun:*	of its cyber; of its glacier
than-*IN* a-*DT* U-*noun:*	than a boob; than a menace
they-*pronoun* are-*verb* U-*noun:*	they are escapist; they are noncontenders
the-*DT* U-*adj* and-*CC:*	the convoluted and; the secretive and
the-*DT* U-*noun* on-*IN:*	the podium on; the trumpeter on
the-*DT* U-*noun* that-*IN:*	the baloney that; the cachet that
to-*TO* U-*verb* his-*pronoun:*	to detach his; to reclaim his
to-*TO* a-*DT* U-*adj:*	to a gory; to a trappist
to-*TO* his-*pronoun* U-*noun:*	to his transylvania; to his waver
to-*TO* their-*pronoun* U-*noun:*	to their arsenal; to their subsistence
trying-*verb* to-*TO* U-*verb:*	trying to sack; trying to whip
with-*IN* an-*DT* U-*noun:*	with an alias; with an avalanche
with-*IN* a-*DT* U-*noun:*	with a gosbank; with a plume

Table 11: Random sample of Ugen-3-Gram Collocations with at least two instances in OP1. The instances shown for each ugen-3-gram were randomly selected.

two words $w1$ and $w2$ is then defined as (where $I(x, r, y)$ is equal to the mutual information between words x and y):

$$\frac{\sum_{(r,w)\in T(w1)\cap T(w2)}(I(w1, r, w) + I(w2, r, w))}{\sum_{(r,w)\in T(w1)} I(w1, r, w) + \sum_{(r,w)\in T(w2)} I(w2, r, w)}$$

Lin processed a 64-million word corpus of news articles, creating a thesaurus entry for each word consisting of the 200 words of the same part of speech that are most similar to it.

As mentioned above, distributional similarity is typically used for one of two purposes: (1) creating dictionaries and (2) smoothing parameter estimates.

Consider (1), which is Lin's focus. The intuition behind his method is that words correlated with many of the same words are more similar. We hypothesized in this work that these words might be distributionally similar because they share pragmatic usages, such as expressing subjectivity, even if they are not close synonyms. For example, consider the 20 most similar words to the adjective **bizarre**: *strange*, *similar*, *scary*, *unusual*, *fascinating*, *interesting*, *curious*, *tragic*, *different*, *contradictory*, *peculiar*, *silly*, *sad*, *absurd*, *poignant*, *crazy*, *funny*, *comic*, *compelling*, *odd*. Some of these are relatively close synonyms, e.g., *strange*, *unusual*, *curious*, *peculiar*, *absurd*, *crazy*, *odd*. Others, while not close synonyms, are also subjective, e.g., *tragic*, *sad*, *poignant*, *compelling*. We would like to identify those as well. Thus, we attempt to extend the set of candidate PSEs beyond those in the training data, by considering words similar to those in the training data.

Now consider (2), smoothing parameter estimates. Evidence is given above in Sections 5.3.2 and 6.2 that low-frequency and unique words appear more often in subjective texts than expected. Thus, we do not want to discard low-frequency words from consideration, but cannot effectively judge the suitability of individual words. To decide whether to retain a word as a PSE, we consider the precision not of the individual word, but of the word together with its cluster of similar words. A set of seed words begins the process. For each seed s_i, the precision of the set $\{s_i\} \cup C_{i,n}$ in the training data is calculated, where $C_{i,n}$ is the set of the n words that are most similar to s_i. If the precision of $\{s_i\} \cup C_{i,n}$ is greater than a threshold T, then the words in this set are retained as PSEs. If it is not, neither s_i nor the words in $C_{i,n}$ are retained. The union of the retained sets will be notated $R_{T,n}$, that is, the union of all sets $\{s_i\} \cup C_{i,n}$ with precision on the training set $> T$.

In (Wiebe, 2000), the seeds (the s_i's) were extracted from the subjective-element annotations in corpus WSJ-SE. Specifically, the seeds were the adjectives that appear at least once in a subjective element in WSJ-SE. In 10-fold cross-validation experiments, where only 1/10 of the data is used for training, and 9/10 is used for testing, we achieved an average increase of more than 13 percentage points over the baseline precision of the entire set of words in the test data. A small amount of training data was used to explore the idea that the process is appropriate even when little training data is available.

In this work, the opinion-piece corpus is used to move beyond the manual annotations and small corpus of the earlier work. The process is performed separately for adjectives and verbs (other parts of speech will be tested in future work). In addition, a much looser criterion is used to choose the initial seeds: all of the adjectives (verbs) in the training data are used.

```
$trainingPrec(s)$ is the precision of $s$ in the training data
$validationPrec(s)$ is the precision of $s$ in the validation data
$testPrec(s)$ is the precision of $s$ in the test data
(similarly for $trainingFreq$,$validationFreq$, and $testFreq$)
$S =$ the set of all adjectives in the training data
for $T$ in [0.01,0.04,...,0.70]:
    for $n$ in [2,3,...,40]:
        $retained = \{\}$
        For $s_i$ in $S$:
            if $trainingPrec(\{s_i\} \cup C_{i,n}) > T$:
                $retained = retained \cup \{s_i\} \cup C_{i,n}$
        $R_{T,n} = retained$
$ADJ_{pses} = \{\}$
for $T$ in [0.01,0.04,...,0.70]:
    for $n$ in [2,3,...,40]:
        if $validationPrec(R_{T,n}) \geq 0.28$ (0.23 for verbs)
        and $validationFreq(R_{T,n}) \geq 100$:
                $ADJ_{pses} = ADJ_{pses} \cup R_{T,n}$
Results in Table 12 show $testPrec(ADJ_{pses})$ and $testFreq(ADJ_{pses})$.
```

Figure 3: Algorithm for selecting adjective and verb features using distributional similarity

				ADJ_{pses}		$VERB_{pses}$	
Training	Validation	Test	Baseline Prec	freq	+prec	freq	+prec
W9-10	W9-22	W9-33	.19	1576	+.12	1490	+.11
W9-22	W9-10						
W9-10	W9-33	W9-22	.14	859	+.15	535	+.11
W9-33	W9-10						
W9-22	W9-33	W9-10	.18	249	+.22	224	+.10
W9-33	W9-22						
All pairings of W9-10, W9-22,W9-33		W9-4	.13	1872	+.17	1777	+.15

Table 12: Frequencies and increases in precision for adjective and verb features identified using distributional similarity with filtering

The process for adjectives is given in algorithmic form in Figure 3. (The process is the same for verbs, with one small difference noted in the figure.) Seeds and their clusters are assessed on a training set for many parameter settings (cluster size n from 2 through 40, and precision threshold T from 0.01 through 0.70 by 3). As mentioned above, each n, T parameter pair yields a set of adjectives $R_{T,n}$, that is, the union of all sets $\{s_i\} \cup C_{i,n}$ with precision on the training set $> T$. A subset, ADJ_{pses}, of those sets is chosen based on precision and frequency in a validation set. Finally, the ADJ_{pses} are tested on the test set.

Table 12 shows the results for four test datasets. The training and validation data are identified for each test set. Multiple training-validation dataset pairs are used for each test set. The results are for the union of the adjectives (verbs) chosen for each pair. The *freq* columns give total frequencies, and the *+prec* columns show the improvements in precision from the baseline.

Recently, Lin and Pantel (2001) have applied Lin's distributional similarity method to finding similar phrases, not just similar words. In the future, we plan to integrate the work in section 6.3 and the current section, by testing phrases that are distributionally similar to the collocations identified in section 6.3 (as well as other phrases known to be potentially subjective).

In previous subsections, we generated features from the subjective-element data and tested them on the opinion-piece data. In this section, we performed both training and and testing using opinion-piece data. In the interests of testing consistency, we now assess the precision of the adjective and verb features (i.e., ADJ_{pses} and $VERB_{pses}$ in Figure 3 and Table 12), generated from opinion-piece data, on the subjective-element data (the NG-SE, WSJ-SE, and NG-FE datasets).

Recall that the precision of a set S with respect to subjective elements is:

$$prec(S) = \frac{\text{number of instances of members of S in subjective elements}}{\text{total number of instances of members of S}}$$

For comparison, we use the sets of adjectives and verbs that appear at least once in a

	Adj Baseline		Verb Baseline		ADJ_{pses}			$VERB_{pses}$		
	freq	prec	freq	prec	freq	prec	improve	freq	prec	improve
WSJ-SE-D	1632	.126	2980	.147	136	.286	127%	151	.251	71%
WSJ-SE-M	1632	.194	2980	.119	136	.431	122%	151	.252	112%
NG-SE	1104	.367	2629	.154	185	.616	68%	275	.220	43%
NG-FE-MM	6771	.025	14669	.028	1047	.039	56%	1495	.037	32%
NG-FE-R	6771	.022	14669	.019	1047	.034	55%	1495	.021	11%

Table 13: Average frequencies and percent improvements in subjective-element data of the sets tested in Table 12. The baselines are the precisions of adjectives/verbs that appear in subjective elements in the subjective-element data.

subjective element. Precisions for these sets, which can be found in the left columns in Table 13, make for challenging baselines. Note that there are four sets of each of ADJ_{pses} and $VERB_{pses}$ represented in Table 13, namely those tested on each of W9-33, W9-22, W9-10, and W9-4 in Table 12. Each of these sets is tested on each of the subjective-element datasets, and the averages are given in Table 13. In particular, for each subjective-element dataset (e.g., WSJ-SE-D), the average frequency and the average percentage improvement of the test sets of Table 12 are given.

Precisions of the adjectives PSEs are higher than the verb PSEs in all datasets. This shows consistency with the higher precision of the ADJ_{pses} sets in the opinion-piece datasets. Precisions of the ADJ_{pses} sets and $VERB_{pses}$ sets are low in the flame element data (NG-FE), although improvements over baselines are realized. Precisions of the verb PSEs on the SE datasets are comparable to their precisions in the opinion-piece data, and results for adjective PSEs are better still, achieving a precision over 60% on the NG-SE dataset.

7 Features Used in Concert

7.1 Consistency in Precision Among Datasets

Table 14 summarizes the results from previous sections in which the opinion-piece data is used for testing. The performance of the various features are consistently good or bad on the same datasets: the performance is better for all features on W9-10 and W9-04 than on W9-22 and W9-33 (except for the ugen-4-grams, which are very low frequency, and the verbs, which are low on W9-10). This is so despite the fact that the features were generated using different procedures and data: the adjectives and verbs were generated from Wall Street Journal data annotated with document-level opinion-piece classifications; the n-gram features were generated from a corpus of newsgroup and WSJ data annotated with expressions-level subjective-element classifications; and the unique unigram feature requires no training. This consistency in performance suggests that the results are not brittle. It will be useful in future work to attempt to identify the differences among the datasets that are responsible for the

	W9-10		W9-04		W9-22		W9-33	
	freq	+prec	freq	+prec	freq	+prec	freq	+prec
unique words	4763	+.16	4794	+.15	4274	+.11	4567	+.11
fixed-2-grams	1972	+.07	1840	+.07	1933	+.04	1839	+.05
ugen-2-grams	256	+.26	281	+.21	261	+.17	254	+.17
fixed-3-grams	243	+.09	213	+.08	214	+.05	262	+.05
ugen-3-grams	133	+.27	148	+.29	147	+.16	133	+.15
fixed-4-grams	17	+.06	18	+.15	12	+.18	14	-.07
ugen-4-grams	3	+.82	13	+.12	15	+.27	13	+.25
adjectives	249	+.22	1872	+.17	859	+.15	1576	+.12
verbs	224	+.10	1777	+.15	535	+.11	1490	+.11
baseline precision		.18		.13		.14		.19

freq: total frequency. **+prec:** increase in precision over baseline.

Table 14: Frequencies and Increases in Precision for All Features

differences in performance.

7.2 Choosing Density Parameters from Subjective Element Data

In (Wiebe, 1994), whether a PSE is interpreted to be subjective depends, in part, on how subjective the surrounding context is. We explore this idea in the current work, assessing whether PSEs are more likely to be subjective if they are surrounded by subjective elements. In particular, we experiment with a density feature to decide whether or not a PSE instance is subjective: if a sufficient number of subjective elements are nearby, then the PSE instance is considered to be subjective; otherwise, it is discarded. The density parameters are a window size W and a frequency threshold T.

In this section, we explore density in the manually-annotated subjective-element (SE) data, and choose density parameters for later use in automatic disambiguation (in Section 7.3).

The process for calculating density in the subjective-element data is given in Figure 4. The PSEs are defined to be all adjectives, verbs, modals, nouns, and adverbs that appear at least once in a subjective element, with the exception of some stop words (line 0 of Figure 4). Note that these PSEs depend only on the subjective-element manual annotations, not on the automatically identified features used elsewhere in the paper, nor on the document-level opinion-piece classes. *PSEinsts* is the set of PSE instances to be disambiguated (line 1). *HiDensity* (initialized on line 2) will be the subset of *PSEinsts* that are retained. In the loop, the density of each PSE instance P is calculated, which is the number of subjective elements that begin or end in the W words preceding or following P (line 6). P is retained if its density is at least T (line 7).

The precision of a set S with respect to subjective-element classifications is the number of members of S that appear in subjective elements over the total number of members of

```
0. $PSEs$ = all adjs, verbs, modals, nouns, and
     adverbs that appear at least once in an $SE$
     (except not, will, be, have).
1. $PSEinsts$ = the set of all instances of $PSEs$
2. $HiDensity = \{\}$
3. For $P$ in $PSEinsts$:
     4. leftWin($P$) = the $W$ words before $P$
     5. rightWin($P$) = the $W$ words after $P$
     6. density($P$) = # of $SEs$ whose first or last
          word is in leftWin($P$) or rightWin($P$)
     7. if density($P$) $\geq T$:
          $HiDensity = HiDensity \cup \{P\}$

8. $prec(PSEinsts) = \frac{\# \text{ of } PSEinsts \text{ in } SEs}{|PSEinsts|}$

9. $prec(HiDensity) = \frac{\# \text{ of } HiDensity \text{ in } SEs}{|HiDensity|}$
```

Figure 4: Algorithm for calculating density in subjective element (SE) data

S. Lines 8-9 assess the precision of the original (*PSEinsts*) and new (*HiDensity*) sets of PSE instances. If *prec*(*HiDensity*) is greater than *prec*(*PSEinsts*), then there is evidence that the number of subjective elements near a PSE instance is related to its subjectivity in context.

The process in Figure 4 was repeated for different parameter settings (T in $[1, 2, 4, \ldots, 48]$ and W in $[1, 10, 20, \ldots, 490]$) on each of the five SE datasets (WSJ-SE1-D, WSJ-SE2-D, WSJ-SE1-M, WSJ-SE2-M, and NG-SE). To find good parameter settings, the results for each dataset are sorted into 5-point precision intervals, and then sorted by frequency within each interval. For example, the top three precision intervals for WSJ-SE1-M are 0.77-0.82, 0.82-0.87, and 0.87-0.92 (no parameter values yield higher precision than 0.92) while the top three intervals for WSJ-SE2-D are 0.46-0.51, 0.51-0.56, and 0.95-1.0 (no parameter values yield precisions between 0.57 and 0.94). The top six precision intervals for each dataset are shown in Table 15, with the parameter values (i.e., T and W), frequency, and precision of the most frequent result in each interval indicated.

The top of Table 15 gives baseline frequencies and precisions, which are $|PSEinsts|$ and *prec*($PSEinsts$), respectively, in line 8 of Figure 4.

Note that set $PSEinsts$ includes words that appear just once, and their single appearance is in an SE. The use of density can only harm precision with respect to this set, since their precision is 100%, and some are removed when the density criterion is applied.

The parameter values exhibit a range of frequencies and precisions, with the expected

	WSJ-SE1-M	WSJ-SE1-D	WSJ-SE2-M	WSJ-SE2-D	NG-SE
Baseline freq	1566	1245	1167	1108	3303
Baseline prec	.49	.47	.41	.36	.51
Range	.87-.92	.95-1.0	.95-1.0	.95-1.0	.95-1.0
T,W	10,20	12,50	20,50	14,100	10,10
freq	76	12	1	1	3
prec	.89	1.0	1.0	1.0	1.0
Range	.82-.87	.90-.95	.73-.78	.51-.56	.67-.72
T,W	6,10	12,60	46,190	22,370	26,90
freq	63	22	53	221	664
prec	.84	.91	.78	.51	.67
Range	.77-.82	.84-.89	.66-.71	.46-.51	.63-.67
T,W	12,40	12,80	18,60	16,310	8,30
freq	292	42	53	358	1504
prec	.78	.88	.68	.47	.63
Range	.72-.77	.78-.83	.61-.66		
T,W	12,50	10,70	14,50		
freq	403	74	88		
prec	.73	.78	.61		
Range	.67-.72	.73-.78	.56-.61		
T,W	20,110	14,110	1,10		
freq	540	85	774		
prec	.68	.73	.57		
Range	.62-.67	.68-.73	.56-.61		
T,W	6,40	12,100			
freq	807	133			
prec	.62	.71			

Table 15: Most frequent entry in the top 6 precision intervals for each subjective element (SE) dataset

```
0. $PSEinsts$ = the set of instances in the test
     data of all PSEs described in Section 6
1. $HiDensity = \{\}$
2. For $P$ in $PSEinsts$:
     3. leftWin($P$) = the $W$ words before $P$
     4. rightWin($P$) = the $W$ words after $P$
     5. density($P$) = # of $PSEinsts$ whose first or last
          word is in leftWin($P$) or rightWin($P$)
     6. if density(P) $\geq$ T:
          $HiDensity = HiDensity \cup \{P\}$

7. $prec(PSEinsts) = \frac{\text{\# of } PSEinsts \text{ in } OPs}{|PSEinsts|}$

8. $prec(HiDensity) = \frac{\text{\# of } HiDensity \text{ in } OPs}{|HiDensity|}$
```

Figure 5: Algorithm for calculating density in opinion piece (OP) data

tradeoff between precision and frequency. We choose the following parameters to test in the experiments described below in section 7.3: for each dataset (e.g., WSJ-SE1-D, WSJ-SE1-M, etc), for each precision interval whose lower bound is at least 10 percentage points higher than the baseline for that dataset, the top two T,W pairs yielding the highest frequencies in that interval are chosen. Among the five datasets, a total of 45 parameter pairs were so selected.

This experiment was done once. We did not experiment with different ranges for T and W, the size of the precision ranges, the lower bound of 10 points higher than baseline, etc.

7.3 Density for Disambiguation

In this section, density is exploited as an informative feature for PSE disambiguation. The process is shown in Figure 5. There are only two differences between the algorithms in Figures 4 and 5. First, in Figure 4, density was defined in terms of the number of subjective elements nearby. However, subjective-element annotations are not available in test data. In Figure 5, density is defined in terms of the number of other PSE instances nearby, where $PSEinsts$ consists of all instances of the automatically identified PSEs described in Section 6 and for which results are given in Table 14. The test data used in the current section is OP1. As specified above in Section 4, all training to define the PSE instances in OP1 was performed on data separate from OP1.

The second difference from the algorithm presented in Figure 4 is that we now assess precision with respect to the document-level classes: the precision of a set is now the number of set members appearing in documents that are classified as opinion pieces (OPs) divided

	WSJ-SE1-M	WSJ-SE1-D	WSJ-SE2-M	WSJ-SE2-D	NG-SE
T,W	10,20	12,50	20,50	14,100	10,10
freq	237	3176	170	10510	8
prec	.87	.72	.97	.57	1.0
T,W	6,10	12,60	46,190	22,370	26,90
freq	459	5289	1323	21916	787
prec	.68	.68	.95	.37	.92
T,W	12,40	12,80	18,60	16,310	8,30
freq	1398	9662	906	24454	3239
prec	.79	.58	.87	.34	.67
T,W	12,50	10,70	14,50		
freq	3176	10995	1581		
prec	.73	.55	.81		
T,W	20,110	14,110	1,10		
freq	5330	12206	21221		
prec	.73	.53	.34		
T,W	6,40	12,100			
freq	11426	13637			
prec	.50	.50			
PSE Baseline: Freq=30938, Prec=.28					

Table 16: Results for high-density PSEs in test data $OP1$ using parameters chosen from subjective-element data

by the cardinality of the set (see lines 7-8 of Figure 5).

An interesting question arose when defining the PSE instances: what should be done with words that are identified to be PSEs (or parts of PSEs) according to multiple criteria? For example, *sunny*, *radiant*, and *exhilarating* are all unique in corpus OP1, and are all in the adjective PSE feature defined for testing on OP1. Collocations add additional complexity. For example, consider the sequence *and splendidly*, which appears in the test data. The sequence *and splendidly* matches the ugen-2-gram (and-*conj* U-*adj*), and the word *splendidly* is unique (all instances of ugen-n-grams result in at least two matches: the ugen-n-gram, and a unique). In addition, more than one n-gram feature may be matched by a sequence. For example, *is it that* matches three fixed-n-gram features: *is it*, *is it that*, and *it that*.

In the current experiments, the more PSEs a word matches, the more weight it is given. The hypothesis behind this treatment is that additional matches represent additional evidence that a PSE instance is subjective. It is realized as follows: each match of each member of each type of PSE is considered to be a PSE instance. Thus, among them, there are 11 members in $PSEinsts$ for the 5 phrases *sunny*, *radiant*, *exhilarating*, *and splendidly*, and *is it that*, one for each of the matches mentioned above.

The process in Figure 5 was performed with the 45 parameter-pair values (T and W) chosen from the subjective-element data as described in Section 7.2. Table 16 shows results for a subset of the 45 parameters, namely the most frequent parameter pair chosen from the top six precision intervals for each training set. The bottom of the table gives a baseline frequency and a baseline precision in OP1, defined as $|PSEinsts|$ and $prec(PSEinsts)$, respectively, on line 7 of Figure 5.

As can be seen, the density features result in substantial increases in precision. Among all 45 parameter pairs, the minimum percentage increase over baseline is 21%. In addition, 24% of the 45 parameter pairs yield increases of 200% or more; 38% yield increases between 100% and 199%, and 38% yield increases between 21%-99%. Interestingly, the results for NG-SE are strong, even though that data is newsgroup rather than Wall Street Journal data.

Notice that, except for one blip ($T, W = 6, 10$ under WSJ-SE1-M), the precisions decrease and the frequencies increase as we go down each column in Table 16. The same pattern can be observed with all 45 parameter pairs. *But the parameter pairs are ordered in Table 16 based on performance in the manually-annotated subjective-element data*, not based on performance in the test data OP1! For example, the entry in the first row, first column ($T, W = 10, 20$) is the parameter pair giving the highest frequency in the top precision interval of WSJ-SE1-M (frequency and precision in WSJ-SE1-M, using the process of Figure 4). Thus, the relative precisions and frequencies of the parameter pairs are carried over from the training to the test data.

7.4 High-Density Sentence Annotations

To assess the subjectivity of sentences with high-density PSEs, we extracted the sentences in corpus OP2 that contain at least one high-density PSE, and manually annotated them. We chose the density parameter pair T,W=12,30, based on its precision and frequency in OP1. This parameter setting yields results that are relatively high precision and low frequency. We chose a low-frequency setting to make the annotation study feasible.

133 sentences were so identified. They are referred to below as the *system-identified* sentences. The sentences and their tags are given in Appendix 1.

The extracted sentences were independently annotated by two judges. One is a co-author of this paper (judge 1), and the other has performed subjectivity annotation before, but is not otherwise involved in this research (judge 2). Sentences were annotated according to the coding instructions of (Wiebe et al., 1999), which, recall, are to classify a sentence as subjective if there is a significant expression of subjectivity in the sentence, of either the writer or someone mentioned in the text. In addition to the subjective and objective classes, a judge could tag a sentence "unsure" if he or she is unsure of his or her rating or considers the sentence to be borderline.

An equal number (133) of other sentences were randomly selected from the corpus to serve as controls. The 133 system-identified sentences and the 133 control sentences were randomly mixed together. The judges were asked to annotate all 266 sentences, not knowing which are system-identified and which are control (nor which are in opinion pieces and which are not).

(1)	The outburst of shooting came nearly two weeks after clashes between Moslem worshippers and Somali soldiers.	oo
(2.a)	But now the refugees are streaming across the border and alarming the world.	ss
(2.b)	In the middle of the crisis, Erich Honecker was hospitalized with a gall stone operation.	oo
(2.c)	It is becoming more and more obvious that his gallstone-age communism is dying with him: ...	ss
(3.a)	Not brilliantly, because, after all, this was a performer who was collecting paychecks from lounges at Hiltons and Holiday Inns, but creditably and with the air of someone for whom "Ten Cents a Dance" was more than a bit autobiographical.	ss
(3.b)	"It was an exercise of blending Michelle's singing with Susie's singing," explained Ms. Stevens.	oo
(4)	Enlisted men and lower-grade officers were meat thrown into a grinder.	ss
(5)	"If you believe in God and you believe in miracles, there's nothing particularly crazy about that."	ss

Table 17: Examples of system-identified sentences

	S	O	U
S	98	2	3
O	2	14	0
U	2	11	1

Table 18: Sentence annotation contingency table; judge 1 counts are in rows and judge 2 counts are in columns.

Each sentence was presented with the sentence that precedes it and the sentence that follows it in the corpus, to provide some context for interpretation.

Table 17 shows examples of the system-identified sentences. Sentences classified by both judges as objective are marked "oo" and those classified by both judges as subjective are marked "ss".

Judge 1 classified 103 of the system-identified sentences as subjective; 16 as objective; and 14 as unsure.

Judge 2 classified 102 of the system-identified sentences as subjective; 27 as objective; and 4 as unsure. The contingency table is given in Table 18.

Given the highly skewed distribution in favor of subjective sentences, and the disagreement on the lower frequency classes ("unsure" and "objective"), the κ value using all three classes is not high: 0.60. However, consistent with the findings in (Wiebe et al., 1999), the κ value for agreement on the sentences for which neither judge is unsure *is* high: 0.86.

A different breakdown of the sentences is illuminating. For 98 of the sentences (set SS), judges 1 and 2 tag the sentence as subjective. Among the other 35 sentences (those tagged objective, those upon which the judges disagree, etc), 20 (set *inBlock*) appear in a block of contiguous system-identified sentences that includes a member of SS. For example, in Table

17, (2.a) and (2.c) are in *SS* while (2.b) is in *inBlock*, and (3.a) is in *SS* while (3.b) is in *InBlock*. Thus, fully 89% of all sentences are either in *SS* or *inBlock*. Among the 15 other sentences, 6 are adjacent to subjective sentences that were not identified by our system (so were not annotated by the judges). These other subjective sentences are the following (they are labeled – in Appendix 1). All contain significant expressions of subjectivity of the writer or someone mentioned in the text, the criterion used in this work for classifying a sentence as subjective.

- When Wieslaw Kielar, one of the first inmate-slaves in Auschwitz-Birkenau, a Catholic, and later a Polish journalist, saw the crematoria begin to operate, he later wrote: "Although we had experienced a great deal during our more than three years in the camp, it was still a shock so great that one lost one's belief in everything, even in God."
- Bathed in cold sweat, I watched these Dantesque scenes, holding tightly the damp hand of Edek or Waldeck who, like me, were convinced that there was no God.
- "The Japanese are amazed that a company like this exists in Japan," says Kimindo Kusaka, head of the Softnomics Center, a Japanese management-research organization.
- A visit to a grimy supermarket in the center of Volzhsky, an industrial town near Volgograd, shows why disgruntlement is on the rise.
- And even if drugs were legal, what evidence do you have that the habitual drug user wouldn't continue to rob and steal to get money for clothes, food or shelter?
- The moral cost of legalizing drugs is great, but it is a cost that apparently lies outside the narrow scope of libertarian policy prescriptions.
- I doubt that one exists.
- Members returned from the August recess after being badgered and bruised by angry seniors.
- They were upset at his committee's attempt to pacify the program critics by cutting the surtax paid by the more affluent elderly and making up the loss by shifting more of the burden to the elderly poor and by delaying some benefits by a year.

Thus, 93% of the sentences identified by the system are subjective or are near subjective sentences. As mentioned above, the sentences together with their tags are given in Appendix 1.

7.5 Using Features for Opinion-piece Recognition

In previous sections, we presented methods for identifying PSEs. In this section, we apply these PSEs to the task of document classification. Specifically, we use PSEs to perform document-level subjectivity recognition. As discussed earlier, this is a difficult task because all documents, whether primarily opinionated or factual, are composed of a mix of both subjective and objective language. An additional obstacle is the typical distribution of documents, heavily skewed toward non-opinion pieces. Despite these hurdles, we achieve positive results in opinion-piece classification, results that are statistically significantly better than the baseline classification of choosing the more frequent class.

Each document is characterized by one feature: the total count of PSE instances in the document, normalized by document length (in words). As in section 7.3 ("Density for Disambiguation"), the PSEs used in this section are all instances of the automatically identified PSEs described in Section 6 and for which results are given in table 14.

We use the k-nearest-neighbor (KNN) algorithm with leave-one-out cross validation. KNN is an instance-based learning algorithm (Mitchell, 1997). Given an instance, for us a document, the KNN algorithm classifies the document according to the majority classification of the document's k closest neighbors. In our experiments, the distance from one document to another is the difference in normalized PSE count. With leave-one-out cross-validation, the set of n documents to be classified is divided into a training set of size n-1 and a validation set of size 1. To classify the document using KNN, we search through the n-1 documents to find the k documents whose normalized PSE counts are closest to the normalized PSE count of the document being classified. The majority classification among the k documents is assigned to the document in the validation set. This process is repeated until every document is classified.

Which value of k to use is chosen during a preprocessing phase, when KNN is run on a separate training set, using leave-one-out cross-validation, for multiple values of k, 1-15. The value of k that attains the best classification during the preprocessing phase is the one used during the later phase described in the previous paragraph. Three values of k resulted in equally good classifications: 11, 13 and 15.

The data used during the preprocessing phase to chose the value of k is W9-4 (one of the four datasets composing OP1). The datasets used during the later stages are W9-10, W9-22, and W9-33 (the other three datasets composing OP1). The manually refined opinion-piece annotations are used.

The classification results for the three test sets are given in Table 19, using k equal to 11 (results for k equal to 13 and 15 are similar). The first column in the table is the baseline accuracy that results from choosing the more frequent class, i.e. non-opinion pieces. The second column is the percent decrease in classification error. The remaining two columns are the number of correctly classified opinion pieces (TP; true positives) and the number of incorrectly classified non-opinion pieces (FP; false positives). Although we experimented with subsets of PSE features (e.g., only collocations), nothing approached significance until all types of PSE features were included.

When testing individual PSEs types (see Table 14), their precisions were consistently higher on W9-10 and lower on W9-22 and W9-33. This differing performance is reflected in the classification results as well. Classification on W9-10 is significantly better than baseline (using McNemar's test (Everitt, 1977)), and much better than the results using W9-22 and W9-33.

Dataset	baseline	-error	TP	FP
W9-10	.889	41%	19	5
w9-22	.915	8%	5	3
w9-33	.921	12%	9	6

Table 19: Classification Results for Three WSJ Datasets

The classification experiment was also repeated on a much larger test set (over 1000 documents), and the results were also significant, with a classification accuracy of 0.93 and a 30% reduction in error from the baseline.

Several studies that investigate classifying documents by genre or style (Karlgren and Cutting, 1994; Kessler et al., 1997; Argamon et al., 1998) include editorials as one of many targeted classes. In contrast, we use the opinionated classes in order to learn and test subjective language. The positive results from the opinion-piece classification in this paper show the usefulness of the various PSE features when used together.

8 Relations to Other Work

There has been much work in other fields involving subjective language, including linguistics, literary theory, psychology, philosophy, and content analysis. Such work has informed our work to date and will continue to be important for future work. Our earlier work (Wiebe, 1994; Wiebe, 1990), which laid the groundwork for the research reported here, drew ideas from research in linguistics and literary theory for its basic conceptualization of subjectivity as well as its manually developed catalogue of PSEs especially (Doležel, 1973; Hamburger, 1973; Kuroda, 1973; Uspensky, 1973; Fillmore, 1974; Fillmore, 1975; Chatman, 1978; Cohn, 1978; Fodor, 1979; Brinton, 1980; Banfield, 1982; Chafe, 1986)). More recent work which is proving useful in planning our future experiments includes (Fludernik, 1993; Stein and Wright, 1995) on subjectivity, (Haiman, 1998) on irony, and (Partington, 1998) on semantic prosody. In addition, there are some knowledge resources that may be useful in future work. For example, in social psychology, David Heise has compiled dictionaries of affective language based on subject ratings www.indiana.edu/˜socpsy/ACT/data.html. It would be interesting to use entries in such dictionaries to seed the distributional similarity process used in this paper, as well as to seed the mutual bootstrapping procedure planned for future work (see Section 9).

Unfortunately, we were not able to find a sufficiently large, freely available corpus annotated with subjectivity distinctions for our current work. For example, while newsgroup messages were annotated for a number of variables related to subjectivity under the *Project H* (Rafaeli et al., 1997) interdisciplinary study of electronic discussions, the amount of data for any single variable is too small to use. Any sufficiently large such corpus that becomes available would be welcome training and testing data for the procedures presented in this paper.

There is also work in fields such as content analysis and psychology on statistically characterizing texts in terms of word lists manually developed for distinctions related to subjectivity. For example, (Hart, 1984) performs counts of a manually-developed list of words and rhetorical devices (e.g., "sacred" terms such as *freedom*) in political speeches to explore potential reasons for public reactions. Anderson and McMasters (1989) use fixed sets of high-frequency words to assign connotative scores to documents and sections of documents along dimensions such as how pleasant, acrimonious, pious, confident, etc. the text is.

What distinguishes our work from work on subjectivity in other fields is that we focus on (1) automatically learning knowledge from corpora, (2) automatically performing contextual disambiguation, and (3) using knowledge of subjectivity in NLP applications.

Other researchers in NLP have performed similar types of classifications and used similar types of features. Many were cited in earlier sections of this paper. Samuel et al. (1998) use a similar procedure to identify collocational features for dialog act recognition. Low-frequency words have been used as features in information extraction (Weeber et al., 2000), and features identified using distributional similarity have been used for syntactic and semantic disambiguation (Hindle, 1990; Dagan et al., 1994) and to develop lexical resources from corpora (Lin, 1998; Riloff and Jones, 1999). In this paper, we develop variations of these features for subjectivity. We are not aware of other work in NLP that uses "ugen" collocational features (i.e., collocations with positions filled by unique words) or density features as developed in this paper.

Spertus (1997) developed a system for recognizing inflammatory messages. As mentioned earlier in the paper, inflammatory language is a type of subjective language, so the task she addresses is closely related to ours. She uses machine learning to select among manually developed features. Though it would be useful to add her features (as well as other manually developed features proposed in the literature) to ours to potentially improve our classification results, the focus in our work is automatically identifying features from the data.

A number of projects investigating genre detection include editorials as one of the targeted genres. For example, in Karlgren and Cutting (1994), editorials are one of 15 categories, and in Kessler et al. (1997), editorials are one of six. Given their goal to perform genre detection in general, they use low-level features that are not specific to editorials. Neither shows significant improvements for editorial recognition. Argamon et al. (1998) address a slightly different task, though it does involve editorials. Their goal is to distinguish not only, e.g., news from editorials, but also these categories in different publications. Their best results are distinguishing among the news categories of different publications; their lowest results involved editorials. In contrast to the above studies, the focus of our work is learning features of subjectivity. We perform opinion-piece recognition in order to assess the usefulness of the various features when used together.

As discussed above in Section 2, we anticipate that knowledge of subjective language may be usefully exploited in a number of current NLP application areas and believe that the work presented in this paper will support experimentation with subjective language in such applications.

Of course, there are many types of AI systems for which state-of-affairs types such as beliefs and desires are central, including systems that perform plan recognition to understand narratives (Dyer, 1982; Lehnert et al., 1983), argument understanding (Alvarado et al., 1986), understanding stories from different perspectives (Carbonell, 1979), and generating language under different pragmatic constraints (Hovy, 1987). Knowledge of linguistic subjectivity could enhance the abilities of such systems to recognize and generate expressions referring to such states of affairs in natural text.

9 Conclusions and Future Work

Knowledge of subjective language promises to be beneficial for many NLP applications including information extraction, question answering, text categorization, and summarization. This paper presents the results of an empirical study in learning knowledge of subjective language from corpora, in which a number of feature types are learned from different kinds of data annotated at multiple levels.

From our study of expression-level annotations, we discovered that uniques words (i.e., *hapax legomena*) are more often subjective than expected (Table 4 on page 12). We explored this correlation between low-frequency words and subjectivity more fully in the opinion-piece data and found that unique words are valuable clues of subjectivity in this data too (Table 6). In addition, the precision of unique words and other low-frequency words increases with corpus size, with the increases tapering off at the largest corpus size tested (Figure 1 on page 16).

A method is presented for automatically identifying potentially subjective collocations. We use this method first to identify collocations composed of fixed sequences of words that tend to be subjective when they appear together. These include expressions such as *of the century* and *get out of here*. (Smajda, 1993) argues for the importance of identifying more general forms of collocations. This paper addresses one unusual type of generalization: one or more positions in the collocation may be filled by any unique word. Interestingly, many of these fixed and generalized collocations include non-content words, such as those typically found on the stop lists of NLP systems (e.g., *of, the, get, out, here* in the above examples).

We used the results of a method for clustering words according to distributional similarity (Lin, 1998) to identify adjectival and verbal clues of subjectivity (Figure 3 on page 25). It is well known that distributional similarity finds both synonyms and words related to each other in other ways (e.g., "nurse" and "doctor"). The hypotheses behind our use of distributional similarity are that synonyms of subjective words are often also subjective, and even words that are not close synonyms may be distributionally similar because they can be similarly used to express subjectivity. Our adjective and verb features are trained on document-level opinion-piece data, and show large increases in precision on both document-level opinion-piece test data (Figure 12 on 25) and on data manually annotated at the expression level with subjective elements (Table 13 on page 27).

Table 14 on page 27 summarizes the results of testing all of the above types of potentially subjective expressions (*PSEs*). All show increased precision in the evaluations. Together, they show an important consistency in performance, in that in almost all cases they perform better or worse on the same datasets. This is so despite the fact that different kinds of data and procedures are used to learn them. Another type of consistency is also evident. PSEs learned from the manual annotations (namely, the collocational features) have precisions higher than baseline in opinion-piece data. Similarly, when the adjective and verb features, learned from opinion-piece data, are tested on the subjective-element annotations, they too have precisions that beat the baseline (Table 13 on page 27).

Having a large stable of PSEs, it was important to disambiguate whether or not PSE instances are subjective in the contexts in which they appear. We discovered that the density of other potentially subjective expressions in the surrounding context is important: if a sufficient number of other clues are nearby, clues are more likely to be subjective than if there are not. Two parameters define the density features, a frequency threshold and a window size. Parameter values are selected using training data manually annotated at the expression level for subjective elements, and then tested on data annotated at the document level for opinion pieces. The PSEs in the training data are defined in terms of the manual annotations, while the PSEs in the test data are automatically identified from text. All of the selected parameters lead to increases in precision on the test data, the majority leading to increases over 100%. Once again we found consistency between expression-level and document-level annotations. PSE sets defined by density have high precision in both the subjective-element data and the opinion-piece data (see Table 15 on page 30 and Table 16 on page 32). The large differences between training and testing suggest that our results are not brittle.

Using a density feature selected from a training set, sentences containing high-density PSEs were extracted from a separate test set, and manually annotated by two judges. Fully 93% of the 133 sentences are subjective sentences or are near subjective sentences. Admittedly, the chosen density feature is a high-precision, low-frequency one (133 sentences is not many). But since the process is fully automatic, it can be applied to more unannotated text to identify regions containing subjective sentences. In addition, because the precision and frequency of the density features is stable across datasets, lower precision but higher frequency options are available.

Clearly, there are many PSEs in the system-identified sentences that the system does not know (e.g., "gallstone-age communism" and "meat thrown into a grinder" in Table 17 on page 34). This paper demonstrates that density can help a system use the subjectivity clues it does know to recognize subjective sentences. The reader is encouraged to look at the sentences identified by the system in Appendix 1.

Finally, the value of the various types of PSEs was demonstrated with the task of opinion-piece classification. Using the k-nearest neighbor classification algorithm with leave-one-out cross-validation, a classification accuracy of 93% was achieved on a large test set, with a reduction in error of 30% from the baseline.

There are many avenues for future work. Our immediate plans include applying the system to large amounts of data, and then applying information extraction and bootstrapping techniques (Riloff and Jones, 1999) to the results to identify subjective language in those sentences which the system does not yet know.

Future work is required to determine how to exploit density features to improve the performance of text categorization (genre detection) algorithms. Another thing to explore is searching for clues of objectivity, such as the politeness features used by (Spertus, 1997). Another is identifying the type of a subjective expression (e.g., positive or negative evaluative), extending work such as (Hatzivassiloglou and McKeown, 1997) on classifying lexemes to the classification of instances in context (compare, e.g., "great!" and "oh great.")

In addition, it would be illuminating to apply our system to data annotated with discourse trees (Carlson et al., 2001). We hypothesize that most objective sentences identified by our system are dominated in the discourse by subjective sentences, and that we are moving toward identifying subjective discourse segments.

Turning to applications, with colleagues, we are starting a project to use subjectivity analysis to improve the accuracy of information extraction technology, focusing on speculative and evaluative sentences. Speculative sentences may contain important information, but an IE system should be able to distinguish between objective facts and speculative information. We hypothesize that evaluative statements are rarely the primary sources of relevant information and are often a source of erroneous extractions due to hyperbole, metaphor, and hypotheticals. We want to proactively identify evaluative sentences so that they can be discarded before information extraction begins, thereby preventing many false hits.

We have also begun an exploration of identifying and characterizing perspectives in texts, to support question answering from multiple perspectives. An important part of this will be recognizing opinionated segments of text.

Finally, we will make our features available to other researchers, so they can experiment with subjective language in their applications.

References

S. J. Alvarado, M. G. Dyer, and M. Flowers. 1986. Editorial comprehension in oped through argument units. In *Proc. of the 1986 National Conference on Artificial Intelligence (AAAI-86)*, pages 250–256.

C.W. Anderson and G.C. McMasters. 1989. Quantification of rewriting by the brothers grimm: A comparison of successive versions of three tales. *Computers and the Humanities*, 23:41–246.

C. Aone, M. Ramos-Santacruz, and W. Niehaus. 2000. Assentor: An nlp-based solution to e-mail monitoring. In *Proc. IAAI-2000*, pages 945–950.

S. Argamon, M. Koppel, and G. Avneri. 1998. Routing documents according to style. In *CAISE-98*.

A. Banfield. 1982. *Unspeakable Sentences*. Routledge and Kegan Paul, Boston.

R. Barzilay, M. Collins, J. Hirschberg, and S. Whittaker. 2000. The rules behind roles: Identifying speaker role in radio broadcasts. In *Proc. AAAI*.

S. Bergler. 1992. *Evidential Analysis of Reported Speech*. Ph.D. thesis, Brandeis University.

D. Biber. 1993. Co-occurrrence patterns among collocations: A tool for corpus-based lexical knowledge acquisition. *Computational Linguistics*, 19(3):531–538.

R. Bod. 1995. *Performance Models of Natural Language*. Ph.D. thesis, ILLC, Universiteit van Amsterdam.

E. Brill. 1992. A simple rule-based part of speech tagger. In *Proc. of the 3rd Conference on Applied Natural Language Processing (ANLP-92)*, pages 152–155.

Laurel Brinton. 1980. ‘represented perception’: A study in narrative style. *Poetics*, 9:363–381.

R. Bruce and J. Wiebe. 1999. Recognizing subjectivity: A case study of manual tagging. *Natural Language Engineering*, 5(2).

J. G. Carbonell. 1979. *Subjective Understanding: Computer Models of Belief Systems.* Ph.D. thesis, Tech. Rept. 150, Department of Computer Science, Yale University, New Haven, CT.

L. Carlson, D. Marcu, and M. E. Okurowski. 2001. Building a discourse-tagged corpus in the framework of rhetorical structure theory. In *Proceedings of the 2nd SIGDIAL Workshop on Discourse and Dialogue, Eurospeech 2001.*

W. Chafe. 1986. Evidentiality in English conversation and academic writing. In Wallace Chafe and Johanna Nichols, editors, *Evidentiality: The Linguistic Coding of Epistemology*, pages 261–272. Ablex, Norwood, NJ.

Seymour Chatman. 1978. *Story and Discourse: Narrative Structure in Fiction and Film.* Cornell University Press, Ithaca, NY.

K. W. Church and D. Yarowsky. 1990. Word association norms, mutual information, and lexicography. *Computational Linguistics*, 1:22–29.

J. Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Meas.*, 20:37–46.

D. Cohn. 1978. *Transparent Minds: Narrative Modes for Representing Consciousness in Fiction.* Princeton University Press, Princeton, NJ.

M.J. Collins and J. Brooks. 1995. Prepositional phrase attachment through a backed-off model. In *Proc. 3rd Workshop on Very Large Corpora (WVLC-93)*, Cambridge. ACL SIGDAT.

T. Copeck, K. Barker, S. Delisle, and S. Szpakowicz. 2000. Automating the measurement of linguistic features to help classify texts as technical. In *Proc. TALN Conference 2000.*

W. Daelemans, A. van den Bosch, and J. Zavrel. 1999. Forgetting exceptions is harmful in language learning. *Machine Learning*, 34:11–41.

I. Dagan, S. Pereira, and Lee. 1994. Similarity-based estimation of word cooccurrence probabilities. In *32th Annual Meeting of the ACL (ACL-94)*, pages 272–278.

I. Dagan, L. Lee, and F. Pereira. 1997. Similarity-based estimation of word cooccurrence probabilities. In *Proc. ACL-EACL 1997*, pages 56–63, Madrid, Spain.

A. P. Dawid and A. M. Skene. 1979. Maximum likelihood estimation of observer error-rates using the EM algorithm. *Applied Statistics*, 28:20–28.

Barbara DiEugenio. 2000. On the usage of kappa to evaluate agreement on coding tasks. In *Proc. 2nd International Conference on Language Resources and Evaluation (LREC2000)*, Athens, Greece.

L. Doležel. 1973. *Narrative Modes in Czech Literature*. University of Toronto Press, Toronto, Canada.

M. G. Dyer. 1982. Affect processing for narratives. In *AAAI 1982*, pages 265–268.

B. S. Everitt. 1977. *The Analysis of Contingency Tables.* Chapman and Hall, London.

Charles Fillmore. 1974. Pragmatics and the description of discourse. In C. Fillmore, G. Lakoff, and R. Lakoff, editors, *Berkeley Studies in Syntax and Semantics I.* Dept. of Linguistics and Institute of Human Learning, University of California Berkeley.

C. J. Fillmore. 1975. *Santa Cruz lectures on deixis.* Indiana University Linguistics Club, Bloomington, IN.

M. Fludernik. 1993. *The Fictions of Language and the Languages of Fiction.* Routledge, London.

Janet Dean Fodor. 1979. *The Linguistic Description of Opaque Contexts.* Outstanding dissertations in linguistics 13. Garland, New York & London.

L. Goodman. 1974. Exploratory latent structure analysis using both identifiable and unidentifiable models. *Biometrika*, 61:2:215–231.

J. Haiman. 1998. *Talk is Cheap: Sarcasm, Alienation, and the Evolution of Language.* Oxford University Press, New York.

Käte Hamburger. 1973. *The Logic of Literature.* Indiana University Press, Bloomington,Indiana.

R. Hart. 1984. Systematic analysis of political discourse: The development of diction. In K. Sanders et al., editor, *Political Communication Yearbook: 1984*, pages 97–134. Southern Illinois University Press.

V. Hatzivassiloglou and K. McKeown. 1997. Predicting the semantic orientation of adjectives. In *ACL-EACL 1997*, pages 174–181, Madrid, Spain, July.

V. Hatzivassiloglou and J. Wiebe. 2000. Effects of adjective orientation and gradability on sentence subjectivity. In *18th International Conference on Computational Linguistics (COLING-2000).*

D. Hindle. 1990. Noun classification from predicate-argument structures. In *28th Annual Meeting of the ACL (ACL-90)*, pages 268–275.

E. Hovy. 1987. *Generating Natural Language under Pragmatic Constraints.* Ph.D. thesis, Yale University.

J. Karlgren and D. Cutting. 1994. Recognizing text genres with simple metrics using discriminant analysis. In *COLING-94.*

D. Karp, Y. Schabes, M. Zaidel, and D. Egedi. 1992. A freely available wide coverage morphological analyzer for English. In *Proc. of the 14th International Conference on Computational Linguistics (COLING-92).*

D. Kaufer. 2000. *Flaming: A White Paper.* www.eudora.com.

B. Kessler, G. Nunberg, and H. Schutze. 1997. Automatic detection of text genre. In *Proc. ACL-EACL-97.*

K. Krippendorf. 1980. *Content Analysis: An Introduction to its Methodology.* Sage Publications, Beverly Hills.

S.-Y. Kuroda. 1973. Where epistemology, style and grammar meet: A case study from the japanese. In P. Kiparsky and S. Anderson, editors, *A Festschrift for Morris Halle*, pages 377–391. Holt, Rinehart & Winston, New York, NY.

L. Lee and F. Pereira. 1999. Distributional similarity models: Clustering vs. nearest neighbors. In *Proc. ACL '99.*

L. Lee. 1999. Measures of distributional similarity. In *Proc. ACL '99*, pages 25–32.

W. G. Lehnert, M. Dyer, P. Johnson, C. Yang, and S. Harley. 1983. BORIS: An Experiment in In-Depth Understanding of Narratives. *Artificial Intelligence*, 20:15–62.

D. Lin and P. Pantel. 2001. Dirt - discovery of inference rules from text. In *Proceedings of ACM SIGKDD Conference on Knowledge Discovery and Data Mining 2001*, pages 323–328.

D. Lin. 1994. Principar–an efficient, broad-coverage, principle-based parser. In *Proc. COLING 94*, pages 482–488.

D. Lin. 1998. Automatic retrieval and clustering of similar words. In *Proc. COLING-ACL '98*, pages 768–773.

D. Lin. 1999. Automatic identification of non-compositional phrases. In *Proc. ACL-99*, pages 317–324.

Diane J. Litman and R. J. Passonneau. 1995. Combining multiple knowledge sources for discourse segmentation. In *Proc. 33rd Annual Meeting of the Association for Computational Linguistics (ACL-95)*, pages 108–115. Association for Computational Linguistics, june.

D. Marcu, M. Romera, and E. Amorrortu. 1999. Experiments in constructing a corpus of discourse trees: Problems, annotation choices, issues. In *The Workshop on Levels of Representation in Discourse*, pages 71–78.

M. Marcus, Santorini, B., and M. Marcinkiewicz. 1993. Building a large annotated corpus of English: The penn treebank. *Computational Linguistics*, 19(2):313–330.

T. Mitchell. 1997. *Machine Learning*. McGraw-Hill.

G. Nunberg, I. Sag, and T. Wasow. 1994. Idioms. *Language*, 70:491–538.

A. Partington. 1998. *Patterns and Meanings*. John Benjamins, Amsterdam.

S. Rafaeli, M. McLaughlin, and F. Sudweeks. 1997. Editors' introduction. *Special Issue on Network and Netplay, Journal of Computer Mediated Communication*, 2(4).

E. Riloff and R. Jones. 1999. Learning Dictionaries for Information Extraction by Multi-Level Bootstrapping. In *Proceedings of the Sixteenth National Conference on Artificial Intelligence*.

W. Sack. 1995. Representing and recognizing point of view. In *Proc. AAAI Fall Symposium on AI Applications in Knowledge Navigation and Retrieval*.

K. Samuel, S. Carberry, and K. Vijay-Shanker. 1998. Dialogue act tagging with transformation-based learning. In *Proc. COLING-ACL 1998*, pages 1150–1156, Montreal, Canada, August.

F. Smajda. 1993. Retrieving collocations from text: Xtract. *Computational Linguistics*, 19:143–177.

E. Spertus. 1997. Smokey: Automatic recognition of hostile messages. In *Proc. IAAI*.

D. Stein and S. Wright, editors. 1995. *Subjectivity and Subjectivisation*. Cambridge University Press, Cambridge.

L. Terveen, W. Hill, B. Amento, D. McDonald, and J. Creter. 1997. Building task-specific interfaces to high volume conversational data. In *Proc. CHI 97*, pages 226–233.

S. Teufel and M. Moens. 2000. What's yours and what's mine: Determining intellectual attribution in scientific texts. In *Proc. Joint SIGDAT Converence on EMNLP and VLC*.

Boris Uspensky. 1973. *A Poetics of Composition*. University of California Press, Berkeley, CA.

T. van der Wouden. 2001. Collocational behaviour in non-content words. In *Proc. ACL-01 Workshop on Collocation: Computational Extraction, Analysis, and Exploitation*, July.

M. Weeber, R. Vos, and R. H. Baayen. 2000. Extracting the lowest-frequency words: Pitfalls and possibilities. *Computational Linguistics*, 26(3).

J. Wiebe, K. McKeever, and R. Bruce. 1998. Mapping collocational properties into machine learning features. In *Proc. 6th Workshop on Very Large Corpora (WVLC-98)*, pages 225–233, Montreal, Canada, August. ACL SIGDAT.

J. Wiebe, R. Bruce, and T. O'Hara. 1999. Development and use of a gold standard data set for subjectivity classifications. In *Proc. 37th Annual Meeting of the Assoc. for Computational Linguistics (ACL-99)*, pages 246–253, University of Maryland, June. ACL.

J. Wiebe, R. Bruce, M. Bell, M. Martin, and T. Wilson. 2001a. A corpus study of evaluative and speculative language. In *Proc. 2nd ACL SIGdial Workshop on Discourse and Dialogue*, Aalborg, Denmark, September.

J. Wiebe, T. Wilson, and M. Bell. 2001b. Identifying collocations for recognizing opinions. In *Proc. ACL-01 Workshop on Collocation: Computational Extraction, Analysis, and Exploitation*, July.

J. Wiebe. 1990. *Recognizing Subjective Sentences: A Computational Investigation of Narrative Text.* Ph.D. thesis, State University of New York at Buffalo.

J. Wiebe. 1994. Tracking point of view in narrative. *Computational Linguistics*, 20(2):233–287.

J. Wiebe. 2000. Learning subjective adjectives from corpora. In *17th National Conference on Artificial Intelligence (AAAI-2000).*

G. K. Zipf. 1935. *The Psycho-Biology of Literary Vocabulary.* Houghton Mifflin, Boston.

Appendix: System-Identified Sentences from Section 7.4

Tags are given at the beginning of the sentence in boldface, judge 1's tag followed by judge 2's. **s** is for subjective; **o** is for objective; and **u** is for uncertain or borderline. A tag of – indicates one of the 9 subjective sentences not identified by the system listed in Section 7.4.

The sentences are printed out per document. Sentences printed one after the other are contiguous in the original document unless otherwise indicated with, e.g., [· · · 4 sentences · · ·]. Consider the first document shown, Document=460, which is an opinion piece (i.e., Opinion-Piece=yes). There are 10 system-identified sentences in Document 460. The first 6 are contiguous in the original and are listed first. They are followed by 4 contiguous sentences that were not identified by the system; those are followed in the original by the other 4 system-identified sentences.

The documents are printed in the following order. The documents whose system-identified sentences are all tagged **ss** are printed first. Among those, the ones with more system-identified sentences are printed before those with fewer. Next, documents with – sentences are printed. Again, the documents with more sentences of interest are printed before those with fewer. Among the remaining documents, those with more **ss** sentences are printed before those with fewer **ss** sentence.

Document=460 Opinion-Piece=yes

ss: Not that the man who so ostentatiously exulted in his literary prestidigitation and who so gleefully declared his hatred of Freud ("the Viennese Quack"), Henry James ("that pale porpoise")–as well as Dostoevski, Faulkner, Eliot and Thomas Mann (the last two simply were "big fakes")–will ever congeal into a safe classic.
ss: He was much too eager to create "something very weird and dynamic," "catastrophic and jolly" like "this great and coily thing" "Lolita."
ss: But in Nabokov's work nothing is simply as it seems.
ss: Beauty and pity, the grotesque and the heartbreaking, freedom and tyranny, the aesthetic and the moral, time lost and time regained–these themes run throughout his books despite disclaimers that he cared not a whit for the social, political or moral aspects of literature.
ss: His notorious high Mandarin disdain was his protective coloration against political or moralistic predators.
ss: Nabokov knew more than most about the bite of history, politics and moralism, about noble causes and the sense of loss.

[· · · 4 sentences · · ·]

ss: In his masterpiece, "Lolita," he portrayed the evolution of lust into love and condemned the cruel perversion that–trying to stop time in its tracks–created pornography and death.
ss: Despite his later evasions, Nabokov was "Conradically conservative" (to steal his pun) and did believe in the moral importance of art: "I never meant to deny the moral impact of art which is certainly inherent in every genuine work of art," he wrote a correspondent in 1945.
ss: "What I do deny and am prepared to fight to the last drop of my ink is the deliberate moralizing which to me kills every vestige of art in a work however skillfully written....
ss: In my opinion, the fact that Tolstoy's 'The Kreutzer Sonata' and 'The Power of Darkness' were written with a deliberate moral purpose largely defeats their purpose, killing the inherent morality of uninhibited art."

Document=1137 Opinion-Piece=yes

ss: Here everything fits together–the edge in Jagger's voice, the wavelike question-and-answer pattern of both vocals and instrumentation, the repeated rhythmic figures that create the kind of seductive groove that characterizes the Stones' best work.

[· · · 1 sentence · · ·]

ss: The rocker "Hold on to Your Hat" sounds like second-rate Aerosmith.

[· · · 3 sentences · · ·]

ss: It's hard to believe that the same man who penned "Sympathy for the Devil" or "Street Fighting Man" wrote: "Now you're sad sad sad Mad mad mad Sad sad sad But you're gonna be fine."

[· · · 6 sentences · · ·]

ss: The set is gargantuan, four stories of corroded scaffolding, orange girders and steam vents representing an abandoned steel mill that might easily dwarf any group of less mythic proportions.
ss: Though the image is one of decay, the Stones make it clear by the ferocity with which they rip through the opener, "Start Me Up," that there's no rust on their wheels.
ss: At 46 Jagger, in Joker-evoking green tails and tight black pants, still gyrates enough to do his father the phys-ed instructor proud. (Because of the bad sightlines, even as close as the 26 th row you have to watch the two giant video screens flanking the stage to see.) While the stationary Stones, drummer Charlie Watts and statuelike bassist Bill Wyman, hold down the bottom, Richards is also a man in motion.

Document=593 Opinion-Piece=yes

ss: Like Hardy, Mr. Ackroyd writes a rather unpolished prose.
ss: Yet whereas Hardy's stylistic coarseness suggests a guileless man struggling to convey a difficult and enigmatic vision, Mr. Ackroyd's suggests sheer sloppiness.
ss: Too often, moreover, he seems condescending toward his characters and glib about the immense abstractions he addresses.

[· · · 2 sentences · · ·]

ss: Joey's wife, Floey, for example, is little more than a fountain of unamusing malapropisms: "a wild fruit chase," "The Hound of the D'Urbervilles."
ss: Evangeline, eternally and hyperbolically, professes the most obviously insincere emotions, proclaiming everything in sight the most "bizarre" or "wonderful" of its kind that she has ever seen.

Document=584 Opinion-Piece=yes

ss: Equally intriguing is the fiercely egotistical Nazi film maker Helga Bauer (Frances Barber), a fictionalized version of the film maker Leni Riefenstahl.
ss: Such impressive villains might seem a tough act to follow, but the portrayals of Shirer, his charming Austrian wife, Tess (Marthe Keller), and the other good guys are equally compelling.
ss: Pitted against Goebbels's overweening propaganda machine, a scrappy figure like Norman Ebbutt of the London Times (Peter Jeffrey) looks heroic even when shouting a drunken warning to an arriving delegation of befuddled Britons.
ss: The same is true for the rest of the ragged crew of American and British reporters who balk at being spoon-fed lies.
ss: This battle of individual wills succeeds in putting the global conflict between freedom and totalitarianism into a neat dramatic nutshell.

Document=1148 Opinion-Piece=yes

ss: During a postmortem on a televised concert in Germany, they complain about the glaring lights ("If I was in the audience I would have left," snorts cellist David Soyer).
ss: They save their big needles for each other.

[· · · 4 sentences · · ·]

ss: First violinist Arnold Steinhardt has the handsome high-strung quality of an overbred racehorse; one thinks of him as Prince Arnold.

[· · · 10 sentences · · ·]

ss: Don't think that the music plays second fiddle to the personality probe of these engaging and amusing fellows.

Document=214 Opinion-Piece=yes

ss: Her Pa (Howard Duff) is the kind of guy who, while saying grace at the supper table, pauses at the word "sin" and glares at the daughter he hasn't seen for two decades, because he knows in his heart that she enjoyed what happened in the cold-storage room, and has been indulging the same taste ever since in the fleshpots of Chicago.

[· · · 9 sentences · · ·]

ss: As for the women, they're pathetic.
ss: Kate's Ma (Louise Latham) is a moral coward.

[· · · 5 sentences · · ·]

ss: At this point, the truce between feminism and sensationalism gets mighty uneasy.

Document=975 Opinion-Piece=no

ss: "If you believe in God and you believe in miracles, there's nothing particularly crazy about that.

[· · · 11 sentence · · ·]

ss: MMPI's publishers say the test shouldn't be used alone to diagnose psychological problems or in hiring; it should be given in conjunction with other tests.
ss: But if a job candidate does poorly on MMPI, that will at the very least provide good grist for a job interview.

Document=70 Opinion-Piece=yes

ss: The narrator may be talking about the depredations of the Shining Path Maoists among the Indians of the Andes, or he may be referring to the plunging inti, Peru's rubber currency, or the corrupting effect of the cocaine trade.

[$\cdots$ 6 sentences $\cdots$]

ss: Saul knew about the Machiguengas from his studies, and through him the narrator became interested in this most recalcitrant and un-Westernizable of all the indigenous peoples who had come under the Spanish yoke.
ss: Saul, meanwhile, came to believe that anthropology, even at its most benign, was as insidious a form of cultural imperialism as the superficially more blatant activities of Christian missionaries.

Document=1222 Opinion-Piece=no

ss: The Bush approach of mixing confrontation with conciliation strikes some people as sensible, perhaps even inevitable, because Mr. Bush faces a Congress firmly in the hands of the opposition.
ss: "Bush, it seems to me, is playing a very smart game," says Charles Jones, a scholar at the Brookings Institution who is doing a study on presidential power.

Document=1172 Opinion-Piece=yes

ss: That statement makes me doubt Mr. Kramer ever served in a combat zone.
ss: Enlisted men and lower-grade officers were meat thrown into a grinder.

Document=1165 Opinion-Piece=yes

ss: Well, I guess it's time we owned up.
ss: We Woodstock "radicals" and "veterans" really were responsible for damaging the social fabric: We did cause racial, ethnic and sexual discrimination, poverty, affordable-housing shortages and hopelessness, environmental degradation, and, probably as well, insider trading, stock-price manipulation and S & L fraud.

Document=1058 Opinion-Piece=yes

ss: Liberal pundits have gone out of their way to blame the Christians' military commander, General Michel Aoun, for provoking a Syrian assault.
ss: It would be like blaming the Americans at Bastogne for the German encirclement.

Document=1053 Opinion-Piece=yes

ss: When after Hitler's war these same people discovered themselves led decade after decade by communist "leaders" named Gomulka, Kadar or Novotny, none forgot for a moment that their place in civilization's progress had been stolen from them.

Document=970 Opinion-Piece=yes

ss: Safes filled with fine shotguns hold as much excitement and promise for a firearms connoisseur as does a box of shiny new crayons for a child.

Document=732 Opinion-Piece=yes

ss: I was not surprised to learn that Ms. Drabble plans a sequel describing the experiences of another character who has been traveling in Cambodia.

Document=653 Opinion-Piece=no

ss: “Peaceable Kingdom,” about a gutsy woman zookeeper who lives right on the zoo grounds with her kids and pet seal, is another one of those warm-hearted family shows that make critics cringe.

Document=498 Opinion-Piece=no

ss: When microwaved, the soggy sandwich doesn’t remotely resemble the golden brown delectable that most Americans grew up on.

Document=445 Opinion-Piece=yes

ss: Their premier passing combo of Wade Wilson to squirmy Anthony Carter may be the game’s best if you don’t count San Fran’s Joe Montana to Jerry Rice.

Document=416 Opinion-Piece=yes

ss: Rep. Tom Lantos asked: “Would you like to rephrase your last few sentences, because they didn’t strike a very reasonable chord?”

Document=286 Opinion-Piece=yes

ss: He had been summoned to the Central Committee of the Soviet Communist Party, after he finished his lunch at the Savoy Hotel, an unlikely prelude to a bureaucratic brow-beating: Eight-foot-tall Rubenesquely naked ladies float on their canvases toward a ceiling teeming with cherubs, all surrounded by gilt laid on with a pastry chef’s trowel and supported by marble corinthian columns whose capitals are fluting fountains of gold.

Document=258 Opinion-Piece=no

ss: It is indeed hard to back away from a widely publicized forecast, and Mr. Straszheim is fidgeting with the handcuffs on this trip.

Document=239 Opinion-Piece=yes

ss: Still, despite their efforts to convince the world that we are indeed alone, the visitors do seem to keep coming and, like the recent sightings, there's often a detail or two that suggests they may actually be a little on the dumb side.

Document=134 Opinion-Piece=yes

ss: We do not know whether RU-486 will be as disastrous as some of the earlier fertility-control methods released to unblinking, uncritical cheers from educated people who should have known better. (Remember the Dalkon Shield and the early birth-control pills?) We will not know until a first generation of female guinea pigs–all of whom will be more than happy to volunteer for the job–has put the abortion pill through the clinical test of time.

Document=1183 Opinion-Piece=yes

–: When Wieslaw Kielar, one of the first inmate-slaves in Auschwitz-Birkenau, a Catholic, and later a Polish journalist, saw the crematoria begin to operate, he later wrote: "Although we had experienced a great deal during our more than three years in the camp, it was still a shock so great that one lost one's belief in everything, even in God."
oo: Kielar and two friends watched crowds of newly arrived innocents, while silent smoke arose above the crematoria.
–: "Bathed in cold sweat, I watched these Dantesque scenes, holding tightly the damp hand of Edek or Waldeck who, like me, were convinced that there was no God.

[· · · 45 sentences · · ·]

ss: St. Therese taught us that signs of God's absence are no less signs of Him: in the night, in the dark, when we seem to be entirely abandoned.
ss: Let Auschwitz-Birkenau remind us of the desolation proper to a place where God was hidden, His absence (and that of human civilization) so grievously felt, amid immeasurable human suffering.

Document=1153 Opinion-Piece=no

ss: ODS employees, most of them young job-hoppers who have left posts at mainstream companies to find greater stimulation and fulfillment, regularly crowd into a desk-packed room for marathon talkfests.

[· · · 4 sentences · · ·]

oo: One tired official, who made a sudden move after sitting stiffly through a lengthy meeting, had to be hospitalized with a pulled back muscle.
–: "The Japanese are amazed that a company like this exists in Japan," says Kimindo Kusaka, head of the Softnomics Center, a Japanese management-research organization.

Document=1017 Opinion-Piece=no

–: A visit to a grimy supermarket in the center of Volzhsky, an industrial town near Volgograd, shows why disgruntlement is on the rise.

uo: One morning, the wire baskets arranged along the sides of the dimly lit store are empty but for dirty jars of pickled green tomatoes and small cans of herring in a sickly tomato sauce.

[· · · 32 sentences · · ·]

ss: His statement came amid a fierce Brezhnev-style propaganda campaign in the Communist Party daily Pravda that has painted grass-roots political movements in the Baltics and elsewhere as antidemocratic forces bent on destroying the Soviet state.

Document=401 Opinion-Piece=yes

ss: Already the toll of drug use on American society-measured in lost productivity, in rising health insurance costs, in hospitals flooded with drug overdose emergencies, in drug caused accidents, and in premature death–is surely more than we would like to bear.

[· · · 9 sentences · · ·]

oo: Many former addicts who have received treatment continue to commit crimes during their recovery.

–: And even if drugs were legal, what evidence do you have that the habitual drug user wouldn't continue to rob and steal to get money for clothes, food or shelter?

[· · · 14 sentences · · ·]

–: The moral cost of legalizing drugs is great, but it is a cost that apparently lies outside the narrow scope of libertarian policy prescriptions.

so: I do not have a simple solution to the drug problem.

–: I doubt that one exists.

Document=1209 Opinion-Piece=yes

–: Members returned from the August recess after being badgered and bruised by angry seniors.

so: House Ways and Means Chairman Dan Rostenkowski was booed and had his car blocked by a pack of screaming senior citizens in his Chicago district.

–: They were upset at his committee's attempt to pacify the program critics by cutting the surtax paid by the more affluent elderly and making up the loss by shifting more of the burden to the elderly poor and by delaying some benefits by a year.

Document=982 Opinion-Piece=yes

ss: Swirling arcs and circular shapes (notably in the vast conoid window that offers a fabulous view of the Dallas skyline to the south and west) fancifully play off against the limestone, glass and marble building, which contains a shoebox hall placed catty-cornered within a larger rectangle.

[· · · 23 sentences · · ·]

ss: Other events also followed in the hall: the Kronos Quartet played before a wildly appreciative audience that had a different kind of blue hair from that sported by Dallas society's grandes dames.

[· · · 1 sentence · · ·]

ss: Now, Maestro Mata was perfectly capable of coaxing whispers from the chorus and bringing out the full operatic grandeur of both orchestral and vocal parts.
ss: Two Met veterans, Tatiana Troyanos and Paul Plishka, were in perfect form and ably encouraged two relative newcomers, Susan Dunn and Richard Leech; all of their sounds resonated.
ss: So did Mstislav Rostropovich's cello in his solo recital on Monday.
oo: Leontyne Price brought the celebration to a close last night.

Document=761 Opinion-Piece=yes

ss: In the same choreographer's "The Conception," to music by one Piart, a couple tussled inconclusively on the floor until the final fadeout.
uo: In Leonid Lebedev's "Aria From 'The Fifth Bachianas Brasileiras,'" set to music by Villa-Lobos, two women and a man, all dressed in shiny lilac leotards, stood in a spotlight in the middle of the stage and grappled with one another.
ss: By the end, the women seemed to have found consolation in each other's arms.

[· · · 1 sentence · · ·]

ss: In Boris Eifman's "Adagio," a cluster of people wearing rags swayed for a while to the sticky quasi-baroque music of that name attributed somewhat dubiously to Albinoni, before leaving the spotlight to a solitary man, who tore off his tattered cloak and emoted in tight bathing trunks before rejoining the group at the back of the stage.

[· · · 2 sentences · · ·]

ss: Lacking confidence in the effectiveness of his crudely propagandistic theme, Mr. Vinogradov introduced a skeletal figure of death, who, despite the encumbrance of a scythe, gamely danced a series of pas de deux with the more pathetic victims of naval oppression.

Document=46 Opinion-Piece=yes

ss: Crises larger and more dangerous to the quality of life than they were 10 years ago.

ss: If you are doubtful, consider for a moment that the Pomton Lakes Reservoirs in northern New Jersey, which supply the tristate area with drinking water, are riddled with toxic PCBs.

[··· 19 sentences ···]

uo: The topic never comes up in ozonedepletion "establishment" meetings, of which I have attended many.

ss: It seems to me that such measurements are a vital part of any intellectually honest evaluation of the threat posed by CFCs.

ss: While recognizing that professional environmentalists may feel threatened, I intend to urge that UV-B be monitored whenever I can.

Document=959 Opinion-Piece=yes

su: Half a century of studying computers has taught us a lot about how the brain doesn't work, but very little about how it does.

ss: So until recently, trying to design better computers by analyzing the brain has been about as fruitful as trying to design an airplane by analyzing feathers and flapping wings.

[··· 17 sentences ···]

ss: But such glitches fade amid the vivid stories Mr. Gilder tells of the oddballs who thrive in a field that favors those whose blood, as he writes, is rarely blue and whose money is rarely seasoned: "The United States did not enter the microcosm through the portals of the Ivy League, with Brooks Brothers suits, gentleman Cs, and warbling society wives....

ss: From immigrants and outcasts, street toughs and science wonks, nerds and boffins, the bearded and the beer-bellied, the tacky and uptight, and sometimes weird, the born again and born yesterday, with Adam's apples bobbing, psyches throbbing, and acne galore, the fraternity of the pizza breakfast, the Ferrari dream, the silicon truth, the midnight modem, and the seventy hour week... from the coarse fanaticism and desperation, ambition and hunger, genius and sweat of the outsider, the downtrodden, the banished and the bullied come most of the progress in the world and in Silicon Valley."

Document=627 Opinion-Piece=yes

ss: Election campaigns were fought in the free part of Germany on the theme of who gets the redder carpet and who is more loved by the oppressors.

ss: But now the refugees are streaming across the border and alarming the world.

oo: In the middle of the crisis, Erich Honecker was hospitalized with a gall stone operation.

ss: It is becoming more and more obvious that his gallstone age communism is dying with him: the Bonn intelligence service "Bundesnachrichtendienst" reported that Gorbachev adviser Valentin Falin has warned that "widespread dissatisfaction among the East German population is likely to lead to hardly controllable mass demonstrations within a relatively short time–next spring at the latest."

Document=597 Opinion-Piece=yes

ss: Sure they can be punished–kept from dessert, made to do pushups until their arms ache.
uo: They can be given short haircuts and early reveille, and they can be yelled at incessantly.
uo: But punishment has never been the purpose of boot camp.
uo: The 50 pushups for a dirty rifle are an incentive to keep the rifle clean henceforth.

[· · · 23 sentences · · ·]

ss: All to a good end: healthy people work better, fight better.
ss: Do we do the same for drug boot camp?
su: Give the recruits eye exams and make sure they become stronger and healthier?

Document=67 Opinion-Piece=yes

oo: She has also worked frequently as a lyricist and as a vocal contractor, hiring and conducting choruses to sing movie title tracks.

[· · · 1 sentence · · ·]

ss: But I was not brave enough to say I was willing to starve to death to be a recording artist.

[· · · 42 sentences · · ·]

ss: I tried not to do much "don't do this, don't do that."

[· · · 1 sentence · · ·]

ss: Not brilliantly, because, after all, this was a performer who was collecting paychecks from lounges at Hiltons and Holiday Inns, but creditably and with the air of someone for whom "Ten Cents a Dance" was more than a bit autobiographical.
oo: "It was an exercise of blending Michelle's singing with Susie's singing," explained Ms. Stevens.
us: "If 'Dangerous Liaisons' had been a musical she would have had to give a different vocal performance because she was a different character in that movie."
uo: Ms. Pfeiffer's vocal performance in "Baker Boys"–recorded for posterity on the soundtrack album (GRP)–is such that after her first number, "More Than You Know," viewers begin murmuring to each other "Is that really her singing?

Document=682 Opinion-Piece=no

ss: Their oversize heads are bald and shaped like eggplants.
oo: Their long, bony hands have only four digits.
ss: The story "has to be true," says Carol Renee Modrall, who runs the Granny's Opera House theater at nearby Carlsbad Caverns and who was hired by the show's producers to assist with everything from costuming to catering.

Document=103 Opinion-Piece=no

ss: "Depending on the president, we could either be a trillion-dollar economy by the end of the century or stay where we are," says political scientist Amaury de Souza.
ss: "And where we are is bad."
os: Despite valiant efforts by Finance Minister Mailson Ferreira da Nobrega, inflation came to 36% in September alone and is expected to top 1,000% for the year.

Document=1004 Opinion-Piece=yes

uo: They were trained to exploit these police tactics by contorting their own bodies, causing their own physical discomfort.
ss: They were trained as to the precise moment to shout, "Video, video" so their cameraman could film closeups of their contorted bodies and faces.

Document=554 Opinion-Piece=no

su: Earlier this month, a senior Soviet official visited Chernobyl, where he was berated by a resident: "Tell me, please, how are we supposed to live?
ss: We are afraid of the water.

Document=548 Opinion-Piece=no

os: Son of the late Marshal Ye Jianying, the man who paved the way for Deng Xiaoping's re-emergence to power in 1977, Mr. Ye maintains close ties to the military.
ss: It is because of this that Beijing feels the need to tread softly in removing the governor from his power base.

Document=397 Opinion-Piece=no

ss: Instead of grappling with pressing issues of state, such as ethnic unrest, Mr. Goncharov, like most of his reform-minded colleagues, spends 90% of his time trying to sort out the mundane and often nightmarish problems of his 275,000 constituents: chronic housing shortages, bad sanitation, petty corruption.
uu: Sometimes he can help.
uo: Often he can do little.
us: Most of the time, he is so overburdened, he barely has time to read the piles of draft laws sent from Moscow.

Document=101 Opinion-Piece=no

ss: "All you have to do is eat a big pizza, and then go to bed," he says.

[· · · 10 sentences · · ·]

uo: One of the busiest ghostbusters is Robert Baker, a 68-year-old semi-retired University of Kentucky psychology professor whose bushy gray eyebrows arch at the mere mention of a ghost.

Document=821 Opinion-Piece=no

oo: Gunfire erupted in Somalia's capital, near the home of an ousted defense minister who has been detained on charges of plotting against the Mogadishu government, sources said.
oo: The outburst of shooting came nearly two weeks after clashes between Moslem worshippers and Somali soldiers.

Document=735 Opinion-Piece=yes

oo: Stanley did not get a hit off Cavarretta.
oo: He served in the Army during World War II, mopping up after the Marines on Guadalcanal and storming the beaches of New Georgia in the Solomons.

Document=725 Opinion-Piece=yes

uo: The purchase was followed by the arrests and seizure of more art supposedly by Dali, Toulouse-Lautrec, Matisse, Kandinsky, Leger, Corot, Schiele, and Howard Chandler Christy, whose murals grace New York's Cafe des Artistes.

Document=533 Opinion-Piece=no

oo: The FDA approval for Losec only allows its use in treating inflammations of the esophagus and stomach lining, which cause a symptom similar to severe heartburn, and in treating a rare condition, known as Zollinger-Ellison Syndrome, caused by excessive secretion of gastric acids.

Document=454 Opinion-Piece=yes

oo: Carol Neblett gives a recital Oct. 1 featuring soprano arias from Korngold's "Die tote Stadt," Massenet's "Le Cid," Verdi's "I Masnadieri," and Bellini's "La Straniera"; songs by Scarlatti, Debussy, Brahms and Rachmaninoff; and a new song cycle written for Ms. Neblett by Thomas Pasatieri, who will accompany her at the piano.