Discourse Segmentation by Human and Automated Means

Rebecca J. Passonneau*
Bellcore and Columbia University

Diane J. Litman†
AT&T Research

The need to model the relation between discourse segment structure and linguistic features of utterances is almost universally acknowledged in the literature on discourse. However, there is weak consensus on the nature of segments and the criteria for recognizing or generating them. We present quantitative results of a two part study using a corpus of spontaneous, narrative monologues. The first part of our paper presents a method for empirically validating discourse segments. We report highly significant results of segmentations performed by naive subjects, where a commonsense notion of speaker intention is the segmentation criterion. In the second part of our study, data abstracted from the subjects' segmentations serve as a target for evaluating two sets of segmentation algorithms. Both sets of algorithms identify segments from linguistic features of discourse. The first set allows us to evaluate and compare the correlation of discourse segmentation with three types of linguistic cues (referential noun phrases, cue words, and pauses). We then develop a second set, using two methods: error analysis and machine learning. Testing the new algorithms on a new data set shows that when multiple sources of linguistic knowledge are used concurrently, algorithm performance improves.

1. Introduction

Each utterance[1] of a discourse either bears a relation to a preceding utterance, or constitutes the onset of a new unit of meaning or action that subsequent utterances may add to. The need to model the relation between the structure of such units (referred to here as *discourse segment structure*) and linguistic features of utterances is almost universally acknowledged in the literature on discourse. However, natural language systems rarely exploit the relation between discourse segment structure and linguistic devices because there is very little data about how they constrain one another. We have been engaged in a two-part study addressing this gap. We report on a method for empirically validating dis-

* 445 South Street, Morristown, NJ 07960 (beck@bellcore.com); Department of Computer Science, New York, NY 10027 (becky@cs.columbia.edu)
† 600 Mountain Avenue, Murray Hill, NJ 07974 (diane@research.att.com)
1 By the term utterance we mean the spoken or written use of a sentence or other linguistic unit.

okay.
SEGMENT X *Meanwhile,*
there are three little boys,
up on the road a little bit,
and they see this little accident.
And u-h they come over,
and they help $\boxed{\text{him}}$,
and you know,
help $\boxed{\text{him}}$ pick up the pears and everything.
SEGMENT Y *A-nd* the one thing that struck me about the- three little boys that were there,
is that one had ay uh I don't know what you call them,
but it's a paddle,
and a ball-,
is attached to the paddle,
and you know you bounce it?
And that sound was really prominent.
SEGMENT Z *Well anyway,*
so- u-m tsk all the pears are picked up,
and $\boxed{\text{he}}$'s on his way again,

Figure 1
Discourse segment structure and linguistic devices.

course segments, and on our development and evaluation of algorithms to identify these segments from linguistic features of discourse. We show that human subjects can reliably perform discourse segmentation using speaker intention as a criterion. We also show that when multiple sources of linguistic knowledge are used (referential noun phrases, cue words, and pauses), algorithm performance approaches human performance.

The excerpt[2] in Figure 1 illustrates the two aspects of discourse that our study addresses. The first pertains to an abstract structure consisting of meaningful discourse segments and their interrelations that underlies a coherent linear sequence of utterances. The utterances in segments X and Z of Figure 1—which describe how three boys come to the aid of another boy who fell off a bike —are more closely related to one another than to those in the intervening segment Y—which describe the paddleball toy owned by one of the three boys. The second discourse feature of interest is that the usage of a wide range of lexicogrammatical devices seems to constrain or be constrained by this more abstract structure. The segmental structure in Figure 1 interacts with the interpretation of the referent of the boxed pronoun *he* in segment Z. The referent of the underlined noun phrase *one* in segment Y is an available referent: without the segmentation, the reasoning

2 This excerpt is taken from a corpus of spoken narratives (Chafe, 1980b) described below.

required to reject it in favor of the intended referent of *he* is quite complex. However, segment Z begins with certain features that indicate a resumption of the speaker goals associated with segment X, such as the use of the phrase *well anyway*, and the repeated mention of the event of picking up the pears. With this segmentation, the referents introduced in segment X are more relevant for interpreting the pronoun in segment Z. Note also that in Figure 1, cue words (italicized) explicitly mark the boundaries of all three segments.[3] Our work is motivated by the hypothesis that natural language technologies can more sensibly interpret discourse and can generate more comprehensible discourse, if they take advantage of this interplay between segmentation and linguistic devices.

In section 2, we give a brief overview of related work. In section 3, we present our analysis of segmentation data we collected from a population of naive subjects. Our results demonstrate an extremely significant pattern of agreement on segment boundaries. In section 4, we use boundaries abstracted from the data produced by our subjects to quantitatively evaluate algorithms for segmenting discourse. In section 4.1, we discuss the coding and evaluation methods. In section 4.2, we test an initial set of algorithms for computing segment boundaries from a particular type of linguistic feature, either referential noun phrases, cue phrases, or pauses. In section 4.3.1, we analyze the errors of our initial algorithms in order to identify a set of enriched input features, and to determine how to combine information from the three linguistic knowledge sources. In section 4.3.2, we use machine learning to automatically construct segmentation algorithms from large feature sets. Our results suggest that it is possible to approach human levels of performance, given multiple knowledge sources. In section 5, we discuss the significance of our results and briefly highlight our current directions.

3 Boundaries are not always marked by cue phrases, and cue phrases do not always signal boundaries.

2. Related Work

There is much debate regarding what to define discourse segments in terms of, and what kinds of relations to assign among segments. As a consequence, the nature of any hypothesized interaction between discourse structure and linguistic devices depends both on the model of discourse that is adopted, and on the types of linguistic devices that are investigated. Here we briefly review previous work on characterizing discourse segments, and on correlating discourse segments with linguistic devices. We conclude each review by summarizing the differences between our study and previous work.

2.1 Characterizing the notion of a segment

A number of alternative proposals have been presented which relate segments to intentions (Grosz and Sidner, 1986), rhetorical structure theory relations (Mann, Matthiessen, and Thompson, 1992) or other semantic relations (Polanyi, 1988; Hobbs, 1979). The linguistic structure of Grosz and Sidner's (1986) discourse model consists of multi-utterance segments and structural relations among them, yielding a discourse tree structure. The hierarchical relations of their linguistic structure are isomorphic with the two other levels of their model, intentional structure and attentional state. Rhetorical relations do not play a role in their model. In Hobbs (1979) and Polanyi (1988), segmental structure is an artifact of coherence relations among utterances, such as elaboration, evaluation, cause, and so on. Their coherence relations are similar to those posited in rhetorical structure theory (RST) (Mann and Thompson, 1988), which informs much work in generation. Polanyi (1988) distinguishes among four types of Discourse Constituent Units (DCUs) based on different types of structural relations (e.g., sequence). As in Grosz and Sidner's (1986) model, Polanyi (1988) proposes that DCUs (analogous to segments) are structured as a tree, and in both models, the tree structure of discourse constrains how the discourse evolves, and how referring expressions are processed. Recent work (Moore and Paris, 1993; Moore and Pollack, 1992) has argued that to account for explanation dialogues, it is necessary to independently model both RST relations and intentions.

Researchers have begun to investigate the ability of humans to agree with one another on segmentation, and to propose methodologies for quantifying their findings. The types of discourse units being coded and the relations among them vary. Several studies have used trained coders to locally and globally structure spontaneous or read speech using the model of Grosz and Sidner (1986), including (Grosz and Hirschberg, 1992; Hirschberg and Grosz, 1992; Nakatani, Hirschberg, and Grosz, 1995; Stifleman, 1995; Hirschberg and Nakatani, 1996). In (Grosz and Hirschberg, 1992; Hirschberg and Grosz, 1992), percent agreement (cf. section 3.2) among 7 coders on 3 texts under two conditions—text plus speech or text alone—is reported at levels ranging from 74.3%-95.1%. In (Hirschberg and Nakatani, 1996), average reliability (measured using the *kappa* coefficient discussed in (Carletta, 1996)) of segment-initial labels among 3 coders on 9 monologues produced by the same speaker, labeled using text and speech, is .8 or above for both read and spontaneous speech; values of at least .8 are typically viewed as representing high reliability (cf. section 3.2). Reliability labeling from text alone is .56 for read and .63 for spontaneous speech.

Other notions of segment have also been used in evaluating naive or trained coders. Hearst (1993; 1994) asked naive subjects to place boundaries between paragraphs of running text, to indicate topic changes. Hearst reports agreement of greater than 80%, and indicates that significance results were found that were similar to those reported in (Passonneau and Litman, 1993). Flammia and Zue (1995) asked subjects to segment textual transcriptions of telephone task-oriented dialogues, using minimal segmentation instructions based on a notion of topic. 18 dialogues were segmented by 5 coders (with varying levels of expertise in discourse), with an average pairwise kappa coefficient of .45. To evaluate hierarchical aspects of segmentation, Flammia and Zue also developed a new measure derived from the kappa coefficient. Swerts (Swerts, 1995) asked 38 subjects to mark "paragraph boundaries" in transcriptions of 12 spontaneous spoken monologues; half of the subjects segmented from text alone and half from text plus speech. However, no quantitative evaluation of the results were reported. Swerts and Ostendorf (Swerts and Ostendorf, 1995) also empirically derived discourse structure, using a spoken corpus of database query interactions. Although the labelers had high levels of agreement, the segmentations were fairly trivial.

Isard and Carletta (1995) presented 4 naive subjects and 1 expert coder with transcripts of task-oriented dialogues from the HCRC Map Task Corpus (Anderson and others, 1991). Utterance-like units referred to as moves were identified in the transcripts, and subjects were asked to identify *transaction* boundaries. Since reliability was lower than the .80 threshold, they concluded that their coding scheme and instructions required improvement.

Moser and Moore (1995b) investigated the reliability of various features defined in Relational Discourse Analysis (Moser, Moore, and Glendening, 1995), based in part on RST. Their corpus consisted of written interactions between tutors and students, using 3 different tutors. Two coders were asked to identify segments, the *core* utterance of each segment, and certain intentional and informational relations between the core and the other *contributor* utterances. As reported in their talk (not in the paper), reliability on segment structure and core identification was well over the .80 threshold. Reliability on intentional and informational relations was around .75, high enough to support tentative conclusions.

Finally, a method for segmenting dialogues based on a notion of control was used in (Whittaker and Stenton, 1988; Walker and Whittaker, 1990). Utterances were classified into 4 types, each of which was associated with a rule that assigned a controller; the discourse was then divided into segments corresponding to which speaker has control.

Neither study presented any quantitative analysis of the ability to reliably perform the initial utterance classification. However, in (Whittaker and Stenton, 1988), a higher level of discourse structure based on topic shifts was agreed upon by at least 4 of 5 judges for 46 of the 56 control shifts.

In sum, relatively few quantitative empirical studies have been made of how to annotate discourse corpora with features of discourse structure, and those that exist use various models such as the Grosz and Sidner model (1986), an informal notion of topic (Hearst, 1993; Flammia and Zue, 1995), transactions (Isard and Carletta, 1995), Relational Discourse Analysis (Moser and Moore, 1995b), or control (Whittaker and Stenton, 1988; Walker and Whittaker, 1990). The modalities of the corpora investigated include dialogic or monologic, written, spontaneous or read, and the genres also vary. Quantitative evaluations of subjects' annotations using notions of agreement, inter-rater reliability, and/or significance show that good results can be difficult to achieve. As discussed in section 3, our aim is to answer basic questions about segmentation, thus we use naive subjects on a highly unstructured task. We use an informal notion of communicative intention as the segmentation criterion, motivated by Grosz and Sidner (1986) and Polanyi (1988); they argue that defining a segment as having a coherent goal, rather than being oriented towards a particular goal, is more general than establishing a repertoire of specific types of segment goals. However, we do not ask coders to identify hierarchical relations among segments. The hypothesis of a tree structure to discourse has frequently been questioned (Dale, 1992; Moore and Pollack, 1992; Hearst, 1994; Walker, 1995), and the magnitude of our segmentation task precludes asking subjects to specify hierarchical relations. Our corpus consists of transcripts of spontaneous speech, produced by 20 different speakers. Finally, we quantify our results using a significance test, a reliability measure, and, for purposes of comparison with other work, percent agreement.

2.2 Correlation of segmentation with utterance features

The segmental structure of discourse has been claimed to constrain and be constrained by disparate phenomena, e.g., cue phrases (Hirschberg and Litman, 1993; Grosz and Sidner, 1986; Reichman, 1985; Cohen, 1984), plans and intentions (Carberry, 1990; Litman and Allen, 1990; Grosz and Sidner, 1986), prosody (Hirschberg and Pierrehumbert, 1986; Chafe, 1980a; Butterworth, 1980), reference (Webber, 1991; Grosz and Sidner, 1986; Linde, 1979), and tense (Webber, 1988; Hwang and Schubert, 1992; Song and Cohen, 1991). However, just as with the early proposals regarding segmentation, many of these proposals are based on fairly informal studies. It is only recently that attempts have been made to quantitatively evaluate how utterance features correlate with independently justified segmentations, thereby avoiding circularity. Many of the studies discussed in the preceding subsection take this approach. The types of linguistic features investigated include prosody (Grosz and Hirschberg, 1992; Hirschberg and Grosz, 1992; Nakatani, Hirschberg, and Grosz, 1995; Hirschberg and Nakatani, 1996; Swerts, 1995; Swerts and Ostendorf, 1995), term repetition (Hearst, 1993; Hearst, 1994), cue words (Moser and Moore, 1995b; Whittaker and Stenton, 1988), and anaphora (Walker and Whittaker, 1990).

The studies of (Grosz and Hirschberg, 1992; Hirschberg and Grosz, 1992) investigate the prosodic structuring of discourse. The correlation of various prosodic features with their independently obtained consensus codings of segmental structure (codings on which *all* labelers agreed) is analyzed using t-tests; the results support the hypothesis that discourse structure is marked intonationally in read speech. For example, pauses tended to precede phrases that initiated segments (independent of hierarchical structure) and to follow phrases that ended segments. Similar results are reported in (Nakatani, Hirschberg, and Grosz, 1995; Hirschberg and Nakatani, 1996) for spontaneous speech as well. Grosz and Hirschberg (1992; 1992) also use the classification and regression tree system CART (Brieman et al., 1984) to automatically construct and evaluate decision trees for classifying aspects of discourse structure from intonational feature values.

The studies of (Swerts, 1995; Swerts and Ostendorf, 1995) also investigate the prosodic structuring of discourse. In (Swerts, 1995), "paragraph boundaries" are empirically obtained as described above. The prosodic features pitch range, pause duration, and number of low boundary tones are then shown to increase continuously with *boundary strength* (the proportion of subjects identifying a boundary). However, there is no statistical analysis of the significance of these correlations. In (Swerts and Ostendorf, 1995), prosodic as well as textual features are shown to be correlated with their independently obtained (but fairly trivial) discourse segmentations of travel-planning interactions, with statistical significance.

Hearst's (1993; 1994) Text Tiling algorithm structures expository text into sequential segments based on term repetition. Hearst (1994) uses information retrieval metrics (cf. section 4.1) to evaluate two versions of Text Tiling against independently derived segmentations produced by at least 3 of 7 human judges. Precision was .66 for the best version, compared with .81 for humans; recall was .61 compared with .71 for humans. The use of term repetition (and a related notion of lexical cohesion) is not unique to Hearst's work, e.g. related studies include (Morris and Hirst, 1991; Youmans, 1991; Kozima, 1993; Reynar, 1994). However, unlike Hearst's work, these studies either use segmentations that are not empirically justified, or present only qualitative analyses of the correlation with linguistic devices.

After identifying segments, and core and contributor relations within segments, Moser and Moore (1995b) investigate whether cue words occur, where they occur, and what word occurs. In their talk, they presented results showing that the occurrence and placement of a discourse usage of a cue word correlates with relative order of core versus contributor utterances. For example, a discourse cue is more likely to occur when the contributor precedes the core utterance ($p \leq .001$).

Finally, Whittaker and Stenton (1988) examined a wide variety of ways that speakers signal discourse structure. They showed that prompts, repetitions and summaries rather than cue words more often signaled control-based discourse segment boundaries. No statistical analysis of the significance of the differences was presented, however. By statistically analyzing distributions of anaphora with respect to control-based discourse segments, Walker and Whittaker (1990) showed that shifts of attentional state (Grosz and Sidner, 1986) occurred when shifts in control were accepted by *all* dialogue participants.

In sum, relatively few studies correlate linguistic devices with empirically justified discourse segmentations. Quantitative evaluations of the correlations include the use of statistical measures and information retrieval metrics. As discussed in section 4, we derive discourse segmentations based on the statistical significance of the agreement among our subjects. In contrast to studies investigating a single feature, we investigate three types of linguistic devices—referential noun phrases, prosody, and cue phrases. In addition, we are concerned with the development of classification algorithms rather than with the demonstration of statistical correlations. We first develop algorithms using each type of linguistic device in isolation, motivated by existing hypotheses in the literature. Then we propose and evaluate methods for combining them. We use measures from information retrieval to quantify and evaluate our results.

3. Intention-Based Segmentation

Here we present the results of a study investigating the ability of naive subjects to identify the same segments in a corpus of spoken narrative discourse. Our first goal is purely exploratory. Despite the wide agreement that discourse structure and linguistic form are mutually constraining, there is little agreement on how to determine the structure of any particular discourse. Thus we do not assume that there are "correct" segmentations against which to judge subjects' responses. Also, as discussed in our previous work (Passonneau and Litman, To appear), the subjects' performance suggests that segmentation is a fuzzy phenomenon. Because our study is exploratory, we took the conservative approach of defining a very open-ended segmentation task that allowed subjects great freedom in the number and size of the segments they identify. Our statistical results indicate that, despite the freedom of the task, subjects independently perform surprisingly similar segmentations of the same discourse. We also show by example that subjects' segmentations reflect the presumed episode structure of the narrative.

We ask subjects to segment discourse using a non-linguistic criterion in order to avoid circularity when we later investigate the correlation of linguistic devices with segments derived from the segmentation task results. Abstracting statistically significant results from the subjects' responses is thus the second goal of our study of the segmentation task. Here we briefly review our statistical results and summarize the motivation for our method of abstracting a single segmentation for a given narrative from a set of subjects' responses. As noted below, more detailed discussion of the statistic we use is presented elsewhere. What we also discuss here, which has not been presented in previous work, is a preliminary evaluation of the reliability of our method where we give a conservative lower bound suggesting that the method is indeed reliable.

3.1 Methodology: Empirically derived segmentation

The claim has been made that different people (investigators or subjects) are likely to assign similar segment boundaries or segment relations to a discourse (Grosz and Sidner, 1986; Reichman, 1985; Mann and Thompson, 1988), but it has also been observed that discourse structure can be ambiguous (Pierrehumbert and Hirschberg, 1987). Studies asking subjects to assign topical units to sample discourses have shown that the resulting segments vary widely in both size and location (Rotondo, 1984). Yet until recently, there has been little attempt to quantify the degree of variability among subjects in performing such a task. Here we present the results of our study of naive subjects performing a relatively unstructured segmentation task on a corpus of similar discourses. Full details are presented in (Passonneau and Litman, To appear).

The corpus consists of 20 spoken narrative monologues known as the Pear stories, originally collected by Chafe (1980b). Chafe recorded and transcribed subjects who had been asked to view the same movie and describe it to a second person. The movie contained 7 sequential episodes about a man picking pears. Chafe identified three types of prosodic phrases from graphic displays of intonation contours, as described in section 4.1. The corpus contains just over 2,000 prosodic phrases with roughly 13,500 words.

For our study, each narrative was segmented by 7 "naive" subjects (as opposed to trained researchers or trained coders), using communicative intention as the segmentation criterion. Except in rare cases, no subject segmented more than one narrative. As discussed above, a variety of criteria for identifying discourse units have been proposed. Our decision to use a commonsense notion of intention as the criterion is aimed at giving the subjects the freedom essentially to choose their own segmentation criteria, and to modify their criteria to fit the evolving discourse.

Two structural constraints were also imposed on the content units that subjects were asked to identify. First, subjects were asked to perform a linear rather than a hierarchical segmentation, where a linear segmentation simply consists of dividing a narrative into sequential units. Second, subjects were restricted to placing boundaries between the prosodic phrases identified by Chafe (1980b). Subjects were presented with transcripts of the narratives formatted so that each non-indented new line was the beginning of a new prosodic phrase. The pause locations and durations transcribed by Chafe (cf. section 4.1.2) were omitted, but otherwise all lexical and non-lexical articulations were retained. The instructions given to the subjects (cf. Appendix) were designed so as have

...
21.2 okay.
$\boxed{\text{6 SUBJECTS (a, b, c, e, f, g)}}$
22.1 Meanwhile,
22.2 there are three little boys,
22.3 up on the road a little bit,
22.4 and they see this little accident.
23.1 And u-h they come over,
23.2 and they help him,
23.3 and you know,
$\boxed{\text{1 SUBJECT (c)}}$
23.4 help him pick up the pears and everything.
$\boxed{\text{5 SUBJECTS (a, d, e, f, g)}}$
24.1 A-nd the one thing that struck me about the- three little boys that were there,
$\boxed{\text{1 SUBJECT (b)}}$
24.2 is that one had ay uh I don't know what you call them,
24.3 but it's a paddle,
24.4 and a ball-,
24.5 is attached to the paddle,
24.6 and you know you bounce it?
$\boxed{\text{2 SUBJECTS (a, d)}}$
25.1 And that sound was really prominent.
$\boxed{\text{4 SUBJECTS (d, e, f, g)}}$
26.1 Well anyway,
$\boxed{\text{2 SUBJECTS (b, c)}}$
26.2 so- u-m tsk all the pears are picked up,
26.3 and he's on his way again,
...

Figure 2
Sample of subjects' responses.

as little bias as possible regarding segment size, and total number of segments. As we discuss further below, both the rate at which subjects assigned boundaries and the size of segments varied widely.

Figure 2 shows the subjects' responses for the excerpt corresponding to Figure 1. The potential boundary sites are between the text lines corresponding to prosodic phrases. The left column shows the prosodic phrase numbers, which are explained later. There are 19 phrases hence 18 boundary sites. The 7 subjects are differentiated by distinct letters of the alphabet. Note that a majority of subjects agreed on only 3 of the 18 possible boundary sites, corresponding to the segmentation illustrated in Figure 1. In general, subjects assigned boundaries at quite distinct rates, thus agreement among subjects is necessarily imperfect. All subjects assigned boundaries relatively infrequently. On average, subjects assigned boundaries at only 16.1% of the potential boundary sites

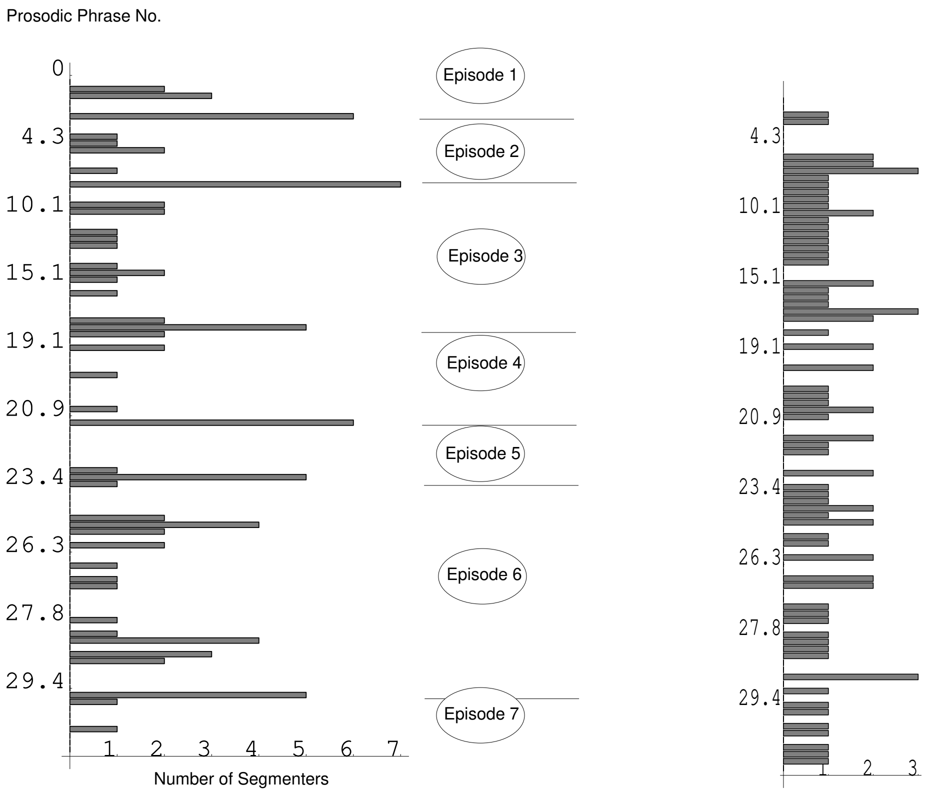

Figure 3
Bar chart of subjects' responses on 1 narrative (showing narrative episodes), compared to random distribution.

(min=5.5%; max=41.3%) in any one narrative. Boundary locations were relatively independent from one another, as shown by the the fact that segments varied in size from 1 to 49 phrases in length (Avg.=5.9). The assumption of independence is important for using statistical analyses of how probable the observed distributions are.

Figure 3 shows two bar charts. The one on the left gives the results for the full narrative excerpted in Figure 1. The x-axis is the number of subjects, from 0 to 7. The y-axis, from top to bottom, corresponds to the potential boundary locations, with prosodic phrase locations numbered as in Figure 2. Each horizontal bar thus represents the number of subjects assigning a boundary at a particular inter-phrase location. Interestingly, there were 6 segment boundaries identified by at least 5 subjects, yielding 7 segments that

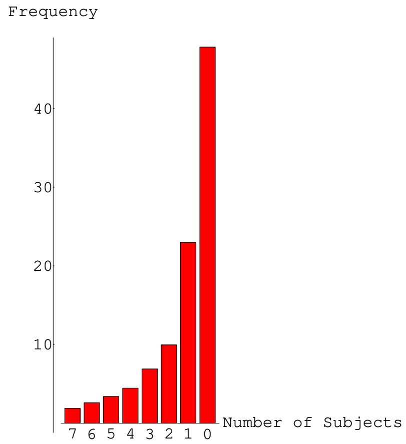

Figure 4
Frequency that N subjects identify any boundary slot as a boundary

correspond closely to the 7 sequential episodes that Chafe (Chafe, 1980b) used to describe the movie. The first 5 segments correspond to the first 5 episodes. The 6th segment corresponds to the 6th episode plus the beginning of the 7th, while the 7th segment corresponds to the end of the 7th episode.

The large proportion of white space to black space in the left bar chart of Figure 3 illustrates graphically that subjects assign boundaries relatively infrequently. The large regions of white space separated by very wide bars shows a striking consensus on certain segments (white space) and segment boundaries (wide black bars). To illustrate graphically the improbability of the occurrence of wide bars (high consensus boundaries), we also show a typical random distribution for a parallel data set in the second bar chart of Figure 3. To create this data, we repeatedly performed the following experiment, and randomly selected one result. First we created seven hypothetical subjects, each of whom assigns the same number of boundaries as one of the real subjects from the same number of potential boundary slots. The hypothetical subjects assign boundaries randomly (but with no repetition). In the random distribution, there are few bars of width 3, and none of any greater width.

We show below that, given the loosely structured task, the probability of the observed distribution depicted in Figure 3 is extremely low, hence highly significant. The statistical test we use identifies $x \geq 3$ as the threshold separating insignificant boundaries from significant ones. The large scattering of narrow bars ($1 \leq x \leq 2$) illustrates the inherent noisiness of the data arising from the fact that subjects assign boundaries at varying rates. The histogram in Figure 4 gives a different view of the same point, showing the relative frequency of cases where N subjects place a boundary at a given location, for N from 0 to 7. The y-axis is normalized to represent the average narrative length of 100 phrases, thus the bar at N=0 indicates that on average, 47.8 of the 100 phrases were not classified as boundaries.[4] The large majority of responses (80%) fall within the bars for N=0 (47.8%), N=1 (23.0%), and N=2 (10.0%), forming a rapidly descending curve. For N=3 and above, the slope of the curve suddenly becomes linear, and much less steep, corresponding to a much more gradual decrease in frequency as values of N go to 7. That there should be any cases where 6 or 7 subjects identify the same boundary is highly improbable, but on average, this happens 4.5 times per narrative. Summing the heights of the bars for N=3 through N=7 indicates that for an average narrative whose length is 100 phrases, there will be about 20 boundaries identified by 3 or more subjects. Below we show how improbable this distribution is.

3.2 Results

3.2.1 Evaluation metrics. Again, our first goal in evaluating the segmentation data from our subjects is to explore the possibility that subjects given as little guidance as possible might yet recognize rather similar segments in the narrative corpus. To make this evaluation, we first use a significance test of the null hypothesis that the distributions

4 Since the narratives vary in length and in relative frequency of boundaries placed by subjects, we normalized the data before averaging across narratives. Where L is the length of a narrative i, the actual frequency of cases where N subjects agree in narrative is multiplied by 100/L, where 100 is the average narrative length.

could have arisen by chance. We then analyze the distributions in more detail to determine what aspects of the distribution are significant, and thereby to abstract significant data for use in defining segmentations for each narrative. The results indicate that the observed distributions are highly significant, i.e., unlikely to have arisen by chance. In section 3.2.2, we briefly review Cochran's Q, the statistic that we use, and the test of the null hypothesis. We then partition Cochran's Q to determine the lowest value on the x-axis in Figure 3 at which agreements on boundaries become statistically significant. The results indicate significance arises when at least 3 subjects agree on a boundary.

Reliability metrics (Krippendorff, 1980; Carletta, 1996) are designed to give a robust measure of how well distinct sets of data agree with, or replicate, one another. They are sensitive to the relative proportion of the different data types (e.g., boundaries versus non-boundaries), but insensitive to the statistical likelihood that agreements will occur. We have already discussed how variable the subjects' responses are, both in number and placement of segment boundaries, so we know that our subjects are not replicating the same behavior. However, all 20 narratives show the same pattern of responses as illustrated in Figure 3: certain boundaries are identified by large numbers of subjects. For any one narrative, we should expect that a new set of 7 subjects should yield roughly the same set of segment boundaries. In other words, our method for abstracting a single set of boundaries from the responses of multiple subjects should be reproducible. In section 3.2.3, we evaluate our method by using Krippendorff's α to evaluate the reliability of boundaries derived from one set of subjects compared with those derived from another set of subjects on the same narrative.

Finally, for purposes of comparison with other studies of segmentation, we report percent agreement. Percent agreement is high, but as argued in Krippendorff (1980), percent agreement is relatively uninformative because it fails to take into account the response rate of individual subjects, a factor built into both Cochran's Q and Krippendorff's α.

Table 1
Matrix representation of boundary data.

Subject	Potential Boundary Sites										
	21.2 22.1	22.1 22.2	22.2 22.3	22.3 22.4	22.4 23.1	23.1 23.2	23.2 23.3	23.3 23.4	23.4 24.1	24.1 24.2	24.2 24.3
a	1								1		
b	1									1	
c	1							1			
d									1		
e	1								1		
f	1								1		
g	1								1		
Total	6	0	0	0	0	0	0	1	5	1	0

3.2.2 Significance. The segmentation data from the 20 narratives can be represented as a set of 20 $i \times j$ matrices, each of the form shown in Table 1. Each matrix has a height of i=7 subjects and width of j=n prosodic phrases less 1. (Table 1 is a partial matrix of width j=11.) The value in a cell $c_{i,j}$ is a one if the *i*th subject assigned a boundary at site *j*, and blank if they did not. We use Cochran's $\mathcal{Q}$ (1950) to evaluate the significance of the distributions in the matrices.[5] Cochran's test evaluates the null hypothesis that the sums of 1s in the columns, representing the total number of subjects assigning a boundary at the *j*th site (T_j), are randomly distributed. It does so by evaluating the significance of the differences in column totals (T_j) across the matrix, with each row total u_i (or total number of boundaries assigned by subject i) assumed to be fixed. Where the average column total is $\overline{T}$, the statistic is given by:

$$\mathcal{Q} = \frac{j\,(j-1) \sum (T_j - \overline{T})^2}{j\,(\sum u_i) - (\sum u_i^2)}$$

Our results indicate that the agreement among subjects is extremely significant. For the 20 narratives, the probabilities that the observed distributions could have arisen by chance range from $p = .1 \times 10^{-6}$ to $p = .6 \times 10^{-9}$.

We now turn to the second question addressed in the segmentation study, how to abstract a set of empirically justified boundaries from the data. We do this by selecting

5 We thank Julia Hirschberg for suggesting this test.

the statistically significant response data. Recall the large amounts of white space in Figure 3, contrasting with the few sharp peaks where many subjects identify the same boundary, which suggests that the significance of $\mathcal{Q}$ owes most to the cases where columns have many 1's. The question is, how many 1's is significant? We address this question by partitioning $\mathcal{Q}$ into distinct components for each possible value of T_j (0 to 7), based on partitioning the sum of squares in the numerator of $\mathcal{Q}$ into distinct samples (Cochran, 1950). Partitioning $\mathcal{Q}$ by the 8 values of T_j shows that $\mathcal{Q}_j$ is significant at the $p = .0001$ level for each distinct $T_j \geq 4$ across all narratives. Probabilities become more significant for higher levels of T_j, and the converse. At T_j=3, p is significant at the .01 level on 19 narratives, and for the remaining narrative $p = .0192$. When we look at correlation of segment boundaries with linguistic features, we use both thresholds $T_j \geq 4$, and $T_j \geq 3$ to select a set of empirically justified boundaries. On average, this gives us 12 ($T_j \geq 4$) or 20 ($T_j \geq 3$) boundaries for a 100-phrase narrative.

3.2.3 Reliability. Reliability metrics provide a measure of the reproducibility of a data set, for example, across conditions or across subjects. Recently, discourse studies have used reliability metrics designed for evaluating classification tasks to determine whether coders can classify various phenomena in discourse corpora, as discussed in section 2.1. The segmentation task reported here is not properly a classification task, in that we do not presume that there is a given set of segment boundaries that subjects are likely to identify. Given the freedom of the task and the use of untrained subjects, a reliability test would be relatively uninformative: it can be expected to range from very low to very high. In fact, sorting the 140 subjects into comparable pairs (i.e., subjects assigning a similar number of boundaries), a reliability metric that ranges between 1 for perfect reliability and -1 for perfect unreliability (Krippendorff's α, discussed below) gives a wide spread of reliability values (from -.3 to .9; average = .34). Our method aims at abstracting away from the absolute differences across multiple subjects per narrative (N=7) to derive a statistically significant set of segment boundaries. Thus, an appropriate test of whether our method is statistically reliable would be to compare two repetitions of the method on the same narratives to see if the the method produces reproducible data.

Although we do not have enough subjects on any single narrative to compare two distinct sets of 7 subjects, we do have 4 narratives with data from 8 distinct subjects. For each set of 8 subjects, we created two randomly selected partitions (A and B) with distinct 4 subjects in each. Then we assessed reliability by comparing the boundaries produced by partitions A and B on the 4 narratives (using a boundary threshold of at least 3 subjects). Because we only have 4 subjects within each partition, this necessarily produces fewer significant boundaries than our method. In other words, this test can only give us a conservative lower bound for reliability. (Recall that significance of a boundary increases exponentially with the number of subjects who agree on a boundary.) But even with this conservative evaluation, reliability is fairly good on two narratives, and promising on average.

A reliability measure indicates how reproducible a data set is by quantifying similarity across subjects in terms of the proportion of times that each response category occurs. This differs from a significance test of the null hypothesis (e.g., our use of Cochran's $\mathcal{Q}$), where observed data is compared to random distribution. We use Krippendorff's α (1980) to evaluate the "reliability" of the two data sets from partitions A and B. The general formula for α is 1 - $\frac{D_O}{D_E}$, where D_O and D_E are observed disagreements and expected disagreements. Computation of α is described below.

Krippendorff's α reports to what degree the observed number of matches could be expected to arise by chance. Again in contrast with Cochran's $\mathcal{Q}$, it is simply a ratio rather than a point on a distribution curve with known probabilities. Values range from 1 to -1, with 0 representing that there are no more agreements observed in the data than would

Table 2
Krippendorff's α comparing boundaries derived from two sets of 4 subjects on 4 narratives.

Boundary Threshhold	Narrative				
	2	4	7	15	Avg.
$T_j \geq 3$	.50	.60	.73	.50	.58

happen by chance. A value of .5 would indicate that the observed number of agreements is halfway between chance and perfect agreement. Negative values indicate the degree to which observed disagreements differ from chance. In principle, α is computed from the same type of matrix shown in Table 1, and can be applied to multi-valued variables that are quantitative or qualitative. Here we summarize computation of a simplified formula for α used for comparing two data sets with a single, dichotomous variable. To exemplify the computation, we use the first two rows of Table 1, giving a matrix of size i=2 $\times$ j=11. The value of D_O (proportion of observed disagreements) is then simply $\frac{M}{j}$, where M is the total number of mismatches (j being the potential number of matches). In our example, D_O has a value of $\frac{2}{11}$ (.18). Where n_1 is the total number of 1's and n_0 is the total number of blanks, D_E is given by $\frac{n_0 n_1}{j(2j-1)}$. In our example, D_E is $\frac{18\times 4}{11\times 21}$ (.31). The formula for α then simplifies to:

$$\alpha = 1 - \frac{(2j-1)(M)}{n_0 n_1}$$

This gives $\alpha = .42$, meaning that the observed case of one agreement out of two potential agreements on boundaries in our example is not quite halfway between chance and perfect agreement. Consider a case where two subjects had twelve responses each (j=12), each subject responded with '1' half the time ($n_1 = n_0 = 12$), and wherever one put a 1 the other did not (M=12). Though the data contains the maximum number of disagreements, α =-0.92, or somewhat less than -1, meaning that a small proportion of the observed disagreement would have arisen by chance.

Table 2 presents the reliability results from a comparison of boundaries found by two distinct partitions of subjects' responses on 4 narratives. An α of .80 using two partitions of 7 subjects would represent very good reproducibility, with values above .67 being somewhat good (Krippendorff, 1980). Note that reliability on narrative 7 (.73) is good despite the small number of subjects. Since, as noted above, we would expect reliability to be much higher if there were 7 subjects, we believe that values above .5 for N=4 subjects indicate potentially good reproducibility. On average α=.58 and the spread is low (σ = .09).

3.2.4 Percent Agreement. Both significance and reliability can stand alone as evaluation metrics, unlike percent agreement. However, we also report percent agreement in order to compare results with other studies. As defined in (Gale, Church, and Yarowsky, 1992), *percent agreement* is the ratio of observed agreements with the majority opinion to possible agreements with the majority opinion. As detailed in (Passonneau and Litman, To appear), the average percent agreement for our subjects on all 20 narratives is 89% (max.=92% ; min.=82%). On average, percent agreement is highest on non-boundaries (91%; max.=95%; min.=84%) and lowest on boundaries (73%; max.=80%; min.=60%), reflecting the fact that non-boundaries greatly outnumber boundaries. These figures compare with other studies (74% to 95% in (Grosz and Hirschberg, 1992), depending upon discourse feature, and greater than 80% in (Hearst, 1993)).

3.2.5 Discussion. We have shown that an atheoretical notion of speaker intention is understood sufficiently uniformly by naive subjects to yield highly significant agreement across subjects on segment boundaries in a corpus of spoken narratives. Probabilities of the observed distributions range from $.6 \times 10^{-9}$ to $.1 \times 10^{-6}$ as given by Cochran's Q. The result is all the more striking given that we used naive coders on a loosely defined task. Subjects were free to assign any number of boundaries, and to label their segments with anything they judged to be the narrator's communicative intention. Partitioning

Cochran's Q shows that the proportion of boundaries identified by at least 3 subjects was significant across all 20 narratives ($p \leq .02$). Significance increases exponentially as the number of subjects agreeing on a boundary increases. A conservative means for estimating a lower bound for the reliability of our method, using Krippendorff's α as a metric, indicates that the method is reliable. The reliability evaluation is conservative in part because it uses far fewer subjects to derive boundaries. Note however, that it is also conservative because it is based on the proportion of identical matches between two data sets. This type of metric ignores the inherent fuzziness of segment location, as discussed in (Passonneau and Litman, To appear). We conclude that boundaries identified by at least 3 of 7 subjects most likely reflect the validity of the underlying notion that utterances in discourse can be grouped into more or less coherent segments. What remains is the question of whether linguistic features correlate at all well with these segments.

4. Algorithmic Identification of Segment Boundaries using Linguistic Cues

As discussed in section 2, there has been little work on examining the use of linguistic cues for recognizing or generating segment boundaries,[6] much less on evaluating the comparative utility of different types of information. In this section we present and evaluate a collection of algorithms that identify discourse segment boundaries, where each relies on a different type of linguistic information. In section 4.1 we introduce our methodology. In section 4.2 we evaluate three initial algorithms, each based on the use of a single linguistic device frequently proposed in the literature: pauses, cue words and referential noun phrases, respectively.[7] Each algorithm was developed prior to any acquaintance with the narratives in our corpus. We evaluate each algorithm by examining its performance in segmenting an initial test set of 10 of our 20 narratives. We also evaluate a

6 A notable exception is the literature on pauses.
7 This section draws from (Passonneau and Litman, 1993).

simple method for combining algorithms. These initial evaluations allow us to quantify the performance of existing hypotheses, to compare the utility of three very different types of linguistic knowledge, and to begin investigating the utility of combining knowledge sources. In section 4.3, we then present two methods for enhancing performance: error analysis, and machine learning.[8] Here we use the 10 narratives previously used for testing as training data. The resulting algorithms are then tested on 5 new narratives. By using enriched linguistic information and by allowing more complex interactions among linguistic devices, both methods achieve results that approach human performance.

4.1 Methodology

4.1.1 Algorithm input and output. Each algorithm is designed to replicate the subjects' segmentation task (break up a narrative into contiguous segments, with segment breaks falling between prosodic phrases). The input to each algorithm is a set of potential boundary sites, coded with respect to a wide variety of linguistic features. The output is a classification of each potential boundary site as either *boundary* or *non-boundary*. In the target output, we classify a potential boundary site as *boundary* if it was identified as such by at least i of the 7 subjects in our empirical study, where we use two values of i. Otherwise it is classified as *non-boundary*. In our experiments, we investigate the correlation of linguistic cues with boundaries identified by both i=3 and i=4 subjects.

Figure 5 is a modified version of Figure 2, showing the classification of the statistically validated boundaries in the same narrative excerpt. (The bracketed numbers represent pauses, as explained below.) The boxes in the figure show the subjects' responses at each potential boundary site; if no box is shown, none of the 7 subjects place a boundary at the site. The italicized parentheticals at each potential boundary site show the resulting boundary classification. Only 3 of the 18 possible boundary sites are classified as

8 This section draws from (Litman and Passonneau, 1995a).

...
21.2 okay.
$\boxed{\text{6 SUBJECTS}}$ *(boundary)*
22.1 [.5 [.2] Meanwhile],
(non-boundary)
22.2 there are three little boys,
(non-boundary)
22.3 [.15] up on the road a little bit,
(non-boundary)
22.4 and they see this little accident.
(non-boundary)
23.1 [1.6 [.55] And u-h] they come over,
(non-boundary)
23.2 and they help him,
(non-boundary)
23.3 [.4? and [.2]] you know,
$\boxed{\text{1 SUBJECT}}$ *(non-boundary)*
23.4 help him pick up the pears and everything.
$\boxed{\text{5 SUBJECTS}}$ *(boundary)*
24.1 [2.7 [1.0] A-nd [1.15]] the one thing that struck me about the- [.3] three little boys that were there,
$\boxed{\text{1 SUBJECT}}$ *(non-boundary)*
24.2 is that one had ay uh [.4] I don't know what you call them,
(non-boundary)
24.3 but it's a paddle,
(non-boundary)
24.4 and a ball-,
(non-boundary)
24.5 [.2] is attached to the paddle,
(non-boundary)
24.6 and you know you bounce it?
$\boxed{\text{2 SUBJECTS}}$ *(non-boundary)*
25.1 .. And that sound was really prominent.
$\boxed{\text{4 SUBJECTS}}$ *(boundary)*
26.1 [4.55 Well anyway,
$\boxed{\text{2 SUBJECTS}}$ *(non-boundary)*
26.2 [.45] so- u-m [.1] throat clearing [.45] tsk [1.15]] all the pears are picked up,
(non-boundary)
26.3 and.. he's on his way again,
...

Figure 5
Excerpt from narrative 9, with boundaries.

boundary, for both i=3 and i=4.

4.1.2 Coding of linguistic features. Given a narrative of n prosodic phrases, there are n-1 potential boundary sites between each pair of prosodic phrases P_i and P_{i+1}, i from 1 to n-1. Each potential boundary site in our corpus is coded for features representing the three different sources of linguistic information of interest: prosody, cue phrases, and referential noun phrases. The linguistic features used in our two sets of experiments are shown in Figure 6. Our initial experiments use only the features marked as '$\circ$,' while our later experiments use the full feature set, along with modifications to the noun phrase features.

Values for the prosodic features are obtained by automatic analysis of the transcripts,

Prosodic Features

- before:+sentence.final.contour,-sentence.final.contour.
- after: +sentence.final.contour,-sentence.final.contour.
- ∘ pause: true, false.
- duration: continuous.

Cue Phrase Features

- ∘ cue_1: true, false.
- $word_1$: also, and, anyway, basically, because, but, finally, first, like, meanwhile, no, now, oh, okay, only, or, see, so, then, well, where, NA.
- cue_2: true, false.
- $word_2$: and, anyway, because, boy, but, now, okay, or, right, so, still, then, NA.

Noun Phrase Features

- ∘ coref: +coref, -coref, NA.
- ∘ infer: +infer, -infer, NA.
- ∘ global.pro: +global.pro, -global.pro, NA.

Combined Feature

- cue-prosody: complex, true, false.

Figure 6
Features and their range of values.

whose conventions are defined in (Chafe, 1980b) and illustrated in Figure 5:

1. "." and "?" indicate sentence-final intonational contours
2. "," indicates phrase-final but not sentence final intonation
3. "[X]" indicates a pause lasting X seconds (measured to an accuracy of about .05 seconds)
4. "[W [Y] *lexical material* [Z]] indicates a sequence lasting W seconds where a Y second pause is followed by lexical material then a pause of Z seconds
5. ".." indicates a break in timing too short to be measured as a pause

(The values in the transcripts are based on an analysis of displays of fundamental frequency contours.) The features *before* and *after* depend on the final punctuation of the phrases P_i and P_{i+1}, respectively. The value is '+sentence.final.contour' if "." or "?", '-sentence.final.contour' if ",". *Pause* is assigned 'true' if P_{i+1} begins with [X] (convention 3) (or with [W [Y] for convention 4), 'false' otherwise. *Duration* is assigned X (convention 3) (or Y for convention 4) if *pause* is 'true', 0 otherwise. The prosodic features were motivated by previous results in the literature. For example, phrases beginning discourse segments were correlated with preceding pause duration in (Grosz and Hirschberg, 1992; Hirschberg and Grosz, 1992). These and other studies (e.g., (Hirschberg and Litman, 1993)) also found it useful to distinguish between sentence and non-sentence final intonational contours.

The cue phrase features are also obtained by automatic analysis of the transcripts. Cue_1 is assigned 'true' if the first lexical item in P_{i+1} is a member of the set of cue words summarized in (Hirschberg and Litman, 1993). $Word_1$ is assigned this lexical item if cue_1 is true, 'NA' (not applicable) otherwise.[9] Cue_2 is assigned 'true' if cue_1 is true and the second lexical item is also a cue word. $Word_2$ is assigned the second lexical item if cue_2 is true, 'NA' otherwise. As with the pause features, the cue phrase features were motivated by previous results in the literature. Initial phrase position (cue_1) was correlated with discourse signaling uses of cue words in (Hirschberg and Litman, 1993). A potential correlation between discourse signaling uses of cue words and adjacency patterns between cue words (cue_2) was also suggested. Finally, (Litman, 1994) found that treating cue phrases individually rather than as a class ($word_1$, $word_2$) enhanced the results of (Hirschberg and Litman, 1993).

Two of the noun phrase (NP) features are hand coded, following (Passonneau, 1994). The two authors coded independently and merged their results. Coding was performed on automatically created coding sheets for each narrative, consulting transcripts that were specially formatted to show prosodic phrase boundaries and numbers, but which were otherwise identical to Chafe's (Chafe, 1980b) original transcriptions. Boundary data, which had been collected but not analyzed, was not available. Comprehensive operational definitions for recognition of coreference (the *coref* feature) were extensively documented

9 The cue phrases that occur in the corpus are shown as potential values in Figure 6.

in (Passonneau, 1994). When coding involves either relatively objective phenomena or a well-defined decision procedure, one can expect a high inter-rater reliability, or a high degree of consensus among different coders (cf. (Duncan and Fiske, 1977; Mokros, 1984)). In addition, preliminary data from a third coder provides good evidence that *coref* can be coded very reliably. Subjectivity in coding the *infer* feature was eliminated by providing operational definitions of a small set of types of inferential links, also fully documented in (Passonneau, 1994), where *infer* occurs only if one or more of the bridging inferences occurs. The last feature, *global.pro*, is computed from the coding of other features and of previously occurring boundaries.

All three features are applied in the context of functionally independent clauses (FICs), which are also hand coded. Comprehensive operational definitions for recognition of FICs were extensively documented in (Passonneau, 1994). FICs are tensed clauses that are neither verb arguments nor restrictive relatives. A feasibility study of the parsers CASS (Abney, 1990) and FIDDITCH (Hindle, 1983) showed that coding FIC units on our data could be automated.[10] If a new FIC (C_j) begins in prosodic phrase P_{i+1}, then NPs in C_j are compared with NPs in previous clauses and the feature values assigned as follows[11]:

1. *coref* = '+coref' if any NPs in C_j and C_{j-1} corefer; else *coref* = '-coref'
2. *infer* = '+infer' if the referent of an NP in C_j can be inferred from C_{j-1} on the basis of a pre-defined set of inference relations; else *infer* = '-infer'
3. *global.pro* = '+global.pro' if the referent of a definite pronoun in C_j is mentioned in a previous utterance, but not prior to the last time a

10 Specifications for adapting the parser and incorporating it into coding software were formulated, but never implemented.

11 The NP algorithm can assign multiple boundaries within one prosodic phrase if the phrase contains multiple clauses; these very rare cases are normalized (Passonneau and Litman, 1993). A total of 5 boundaries are eliminated in 3 of the 10 test narratives (out of 213 in all 10).

boundary was assigned; else *global.pro* = '-global.pro'

Note that the *global.pro* feature is defined in a manner that depends on incremental assignment of boundaries and coding of features. To evaluate *global.pro* for an utterance C_j requires that all boundaries occurring prior to C_j have been assigned. If a new FIC is not initiated in P_{i+1}, values for all three features are 'NA'. The NP features reflect Passonneau's hypotheses that adjacent utterances are more likely to contain expressions that corefer, or that are inferentially linked, if they occur within the same segment; and that a definite pronoun is more likely than a full NP to refer to an entity that was mentioned in the current segment, if not in the previous utterance. These hypotheses are inspired by centering theory (Grosz, Joshi, and Weinstein, 1995), psycholinguistic research (Marslen-Wilson, Levy, and Tyler, 1982; Levy, 1984), and pilot studies on data from corpora (Schiffman, 1985) or published excerpts (Grosz, 1977; Grosz and Sidner, 1986). Unlike the cue and pause features, the NP features were thus not directly based on simplifications of existing results.

Cue-prosody, which encodes a combination of prosodic and cue word features, was motivated by an analysis of errors on our training data, as described in section 4.3.1. *Cue-prosody* is 'complex' if:

1. *before* = '+sentence.final.contour'
2. *pause* = 'true'
3. And either:
   (a) cue_1 = 'true', $word_1 \neq$ 'and'
   (b) cue_1 = 'true', $word_1$ = 'and', cue_2 = 'true', $word_2 \neq$ 'and'

Else, *cue-prosody* has the same values as *pause*.

22.4 and they$_i$ see this little accident.
23.1 [1.6 [.55] And u-h] they$_i$ come over,
23.2 and they$_i$ help him$_j$,
23.3 [.4? and [.2]] you know,
23.4 (ZERO$_i$) help him$_j$ pick up the pears and everything.

Site	before	after	pause	duration	cue_1	$word_1$	cue_2	$word_2$	coref	infer	global.pro	cue-prosody
(22.4,23.1)	+	-	t	0.55	t	and	f	NA	+	-	+	t
(23.1,23.2)	-	-	f	0	t	and	f	NA	+	-	+	f
(23.2,23.3)	-	-	t	0	t	and	f	NA	NA	NA	NA	t
(23.3,23.4)	-	+	f	0	f	NA	f	NA	+	-	+	f

Figure 7
Example feature coding of potential boundary sites.

Figure 7 illustrates how four example boundary sites in Figure 5 would be coded using the features in Figure 6. The subscripting on noun phrases indicates coreference.

The ability of humans to reliably code linguistic features similar to those coded in Figure 7 has been demonstrated in various studies. Evaluation of prosodic labeling using TOBI, a prosodic transcription system somewhat similar to that used in the Pear corpus, has been found to be quite reliable between transcribers (Pitrelli, Beckman, and Hirschberg, 1994). The results of a study of 953 spoken cue phrases showed that 2 judges agreed on whether cue phrases illustrated a discourse signaling usage or not in 878 (92.1%) cases (Hirschberg and Litman, 1993). For these 878 cases, an algorithm that assigned cues discourse signaling usages if they were the first lexical item in their intermediate intonational phrase performed with 75% accuracy (Litman, 1994), which is analogous to the method used here to assign the value 'true' to the feature cue_1.

4.1.3 Evaluation. The segmentation algorithms presented in section 4.2 are evaluated by quantifying their performance in segmenting a test set of 10 narratives from our corpus. As discussed above, there is no training data for the algorithms in this section, which are derived from the literature. These initial results provide us with a baseline for quantifying improvements resulting from distinct modifications to the algorithms.

In contrast, the algorithms presented in section 4.3 are developed using the 10 narratives previously used for testing as a *training* set of narratives. The algorithms in this

	Subjects	
Algorithm	Boundary	Non-Boundary
Boundary	a	b
Non-Boundary	c	d

Recall $= \frac{a}{(a+c)}$ **Fallout** $= \frac{b}{(b+d)}$

Precision $= \frac{a}{(a+b)}$ **Error** $= \frac{(b+c)}{(a+b+c+d)}$

Figure 8
Information retrieval metrics.

section are developed by tuning the previous algorithms (e.g., by considering both new and modified linguistic features) such that performance on the training set is increased. The resulting algorithms are then evaluated by examining their performance on a separate *test* set of 5 more narratives. (The remaining 5 of the 20 narratives in the corpus are reserved for future research.) The 10 training narratives range in length from 51 to 162 phrases (Avg.=101.4), or from 38 to 121 clauses (Avg.=76.8). The 5 test narratives range in length from 47 to 113 phrases (Avg.=87.4), or from 37 to 101 clauses (Avg.=69.0). The ratios of test to training data measured in narratives, prosodic phrases and clauses, respectively, are 50.0%, 43.1% and 44.9%. For the machine learning algorithm we also estimate performance using *cross-validation* (Weiss and Kulikowski, 1991), as detailed in section 4.3.2. The evaluations in this section allow us to compare the utility of two tuning methods: error analysis, and machine learning.

To quantify algorithm performance, we use the information retrieval metrics shown in Figure 8. Recall is the ratio of correctly hypothesized boundaries to target boundaries. Precision is the ratio of hypothesized boundaries that are correct to the total hypothesized boundaries. (Cf. Figure 8 for fallout and error.) These metrics assume that ideal behavior would be to identify all and only the target boundaries: the values for b and c in Figure 8 would thus both equal 0, representing no errors.[12] The ideal values for recall, precision, fallout, and error are 1, 1, 0, and 0, while the worst values are 0, 0, 1, and 1. To get an intuitive summary of overall performance, we also sum the deviation of the observed

12 Elsewhere we have discussed problems with the use of IR metrics, given that segmentation is a fuzzy phenomenon. However, they provide a rough (lower bound) measure of performance.

value from the ideal value for each metric: (1-recall) + (1-precision) + fallout + error. The summed deviation for perfect performance is thus 0.

Finally, to interpret our quantitative results, we use the performance of our human subjects as a target for the performance of our algorithms (Gale, Church, and Yarowsky, 1992). Table 3 shows the average human performance for both the training and test sets of narratives, for both boundaries identified by at least 3 and 4 subjects. Note that human performance is basically the same for both sets of narratives. However, two factors prevent this performance from being closer to ideal (e.g., recall and precision of 1). The first is the wide variation in the number of boundaries that subjects used, as discussed above. The second is the inherently fuzzy nature of boundary location. We discuss this second issue at length in (Passonneau and Litman, To appear). We also present relaxed IR metrics that penalize near misses (cases where an algorithm does not place a boundary at a statistically validated boundary location, but does place one within one phrase of the validated boundary) less heavily in (Litman and Passonneau, 1995b).

4.2 Initial hypotheses

In principle, the process of determining whether the statistically validated segment boundaries correlate with linguistic devices requires a complex search through a large space of possibilities, depending on what set of linguistic devices one examines, and what

Table 3
Average human performance.

	Recall	Precision	Fallout	Error	Summed Deviation
Boundary Threshold = 3					
Training Set	.63	.72	.06	.12	.83
Std. Dev.	.17	.17	.05	.05	.31
Test Set	.64	.68	.07	.11	.86
Std. Dev.	.19	.20	.06	.06	.42
Boundary Threshold = 4					
Training Set	.74	.55	.09	.11	.91
Std. Dev.	.19	.16	.06	.05	.33
Test Set	.73	.55	.09	.10	.91
Std. Dev.	.20	.21	.06	.06	.43

features are used to recognize and classify them. Rather than developing a method to search blindly through the space of possibilities, we first provide an initial evaluation of three linguistic devices whose distribution or surface form has frequently been hypothesized to be conditioned by segmental structure: referential noun phrases, cue words and pauses. We evaluate three algorithms, each of which uses features pertaining to only one of these linguistic devices, in order to see whether linguistic associations proposed in the literature can be used by natural language processing systems to perform segmentation, and to compare the utility of different knowledge sources. Unlike most previous work, which typically considers each linguistic device in isolation, we also evaluate a simple additive method for combining linguistic devices, in which a boundary is proposed if each separate algorithm proposes a boundary. As we will see, the performance of our algorithms improves with the amount of knowledge exploited. The recall of the three algorithms is comparable to human performance, the precision much lower, and the fallout and error of only the noun phrase algorithm comparable. Furthermore, the results on combining algorithms suggests that with more sophisticated methods, results approaching human performance can be achieved.

4.2.1 Pauses. Several studies have demonstrated various correlations between pauses and discourse segment boundaries (Grosz and Hirschberg, 1992; Hirschberg and Grosz, 1992; Hirschberg and Nakatani, 1996; Swerts, 1995), e.g., segment initial phrases have been correlated with longer preceding pause durations. As shown in Figure 9, we used a simplification of these results[13] to develop an algorithm for identifying boundaries in our corpus using pauses. If a pause occurs at the beginning of the prosodic phrase after the potential boundary site, the potential boundary site is classified as *boundary* and the

13 Our initial algorithm does not take the duration of the pause into account; pause duration is considered in the algorithms presented in section 4.3.2, however. In addition, since our segmentation task is not hierarchical, we do not note whether phrases begin, end, suspend, or resume segments.

if pause = true **then** *boundary*
else *non-boundary*

Figure 9
Pause algorithm.

Site	before	after	pause	duration	cue_1	$word_1$	cue_2	$word_2$	coref	infer	global.pro	cue-prosody
(22.4,23.1)	+	-	t	0.55	t	and	f	NA	+	-	+	t
(23.1,23.2)	-	-	f	0	t	and	f	NA	+	-	+	f
(23.2,23.3)	-	-	t	0	t	and	f	NA	NA	NA	NA	t
(23.3,23.4)	-	+	f	0	f	NA	f	NA	+	-	+	f

Site	BOUNDARY	PAUSE	CUE	NP	EA	ML
(22.4,23.1)	-	+	+	-	-	-
(23.1,23.2)	-	-	+	-	-	-
(23.2,23.3)	-	+	+	-	-	-
(23.3,23.4)	-	-	-	-	-	-

Figure 10
Statistically-validated versus algorithmically derived boundaries.

phrase is taken to be the beginning of a new segment.

Figure 10 shows boundaries assigned by the pause algorithm (PAUSE) for the boundary slot codings from Figure 7, repeated at the top of the figure. For example, the pause algorithm assigns a boundary between prosodic phrases 22.4 and 23.1, but not between phrases 23.1 and 23.2, because the values of the feature pause are true and false.

Table 4 shows the average performance of the pause algorithm for statistically validated boundaries at the .0001 level (those boundaries proposed by at least 4 subjects). Recall is 92% (σ=.008; max=1; min=.73), precision is 18% (σ=.002; max=.25; min=.09), fallout is 54% (σ=.004; max=.65; min=.45), and error is 49% (σ=.004 max=.61; min=.41). Our algorithm thus performs with recall higher than human performance.[14] However, precision is low, both fallout and error are quite high. The summed deviation metric, which takes all the metrics into account, shows that on the whole performance is considerably worse than humans.

4.2.2 Cue words. Cue words (e.g., "now") are words that are sometimes used to explicitly signal the structure of a discourse. Hirschberg and Litman (1993) examined a

14 Note that the humans did not have access to pause information. Other studies have shown that when both speech and text are available to labelers, segmentation is clearer (Swerts, 1995) and reliability improves (Hirschberg and Nakatani, 1996).

if cue_1 = true **then** *boundary*
else *non-boundary*

Figure 11
Cue word algorithm.

large set of cue words proposed in the literature and showed that certain prosodic and structural features, including a position of first in prosodic phrase, are highly correlated with the discourse uses of these words. As shown in Figure 11, we developed a baseline segmentation algorithm based on a simplification of these results, using the value of the single cue phrase feature cue_1. That is, if a cue word occurs at the beginning of the prosodic phrase after the potential boundary site, the usage is assumed to be discourse. Thus the potential boundary site is classified as *boundary* and the phrase is taken to be the beginning of a new segment. Figure 10 shows boundaries (CUE) assigned by the algorithm.

Table 4 shows the average performance of the cue word algorithm. Recall is 72% (σ=.027; max=.88; min=.40), precision is 15% (σ=.003; max=.23; min=.04), fallout is 53% (σ=.006 max=.60; min=.42) and error is 50% (σ=.005 max=.60; min=.40). While recall is quite comparable to human performance (row 4), the precision is low while fallout and error are quite high.

4.2.3 Referential noun phrases. The last segmentation algorithm we describe takes as input information about referential NPs. We refer to this algorithm as NP. Unlike the previous algorithms, in NP the potential boundaries are first computed as ordered pairs

Table 4
Evaluation for $T_j \geq 4$.

	Recall	Precision	Fallout	Error	Summed Deviation
PAUSE	.92	.18	.54	.49	1.93
CUE	.72	.15	.53	.50	2.16
NP	.50	.31	.15	.19	1.53
Humans	.74	.55	.09	.11	.91

For (FIC_{i-1}, FIC_i):
if (coref = -coref **and** infer = -infer **and** global.pro = -global.pro) **then** *boundary*
else *non-boundary*

Figure 12
Referential NP algorithm.

of adjacent functionally independent clauses (FIC_i,FIC_{i+1}; cf. section 4.1.2) then normalized to ordered pairs of prosodic phrases (cf. note 11). NP operates on the principle that if an NP in the current FIC provides a referential link to the current segment, the current segment continues. However, NPs and pronouns are treated differently based on the assumption that the referent of a third person definite pronoun is more prominently in focus (cf. (Passonneau, 1994)). A third person definite pronoun provides a referential link if its index occurs anywhere in the *current segment*. Any other NP type provides a referential link if its index occurs in the *immediately preceding FIC*. Figure 12 illustrates the two decisions made by NP for each pair of adjacent FICs. As described in section 4.1.2, the `coref` feature is `-coref` if no NP in FIC_i corefers with an NP in the FIC_{i-1}; the `infer` feature is `-infer` if no NP in FIC_i is inferentially linked to an NP in FIC_{i-1}; the `global.pro` feature is `-global.pro` if FIC_i contains no third person definite pronoun co-referring with an NP in any prior FIC up to the last boundary assigned by the algorithm. If any feature has a positive value, no boundary is assigned; if all have negative values, (FIC_{i-1},FIC_i) is classified as a boundary.

The column headed NP in Figure 10 indicates boundaries assigned by the NP algorithm. No boundaries are assigned by NP. The first three phrases in Figure 7 correspond directly to three consecutive FICs, and each FIC has an NP co-referring with an NP in the next; likewise the global.pro feature is present. However, phrase 23.3 is the onset of an FIC that continues through 23.4, so phrase 23.3 is not coded for NP features. The coref and global.pro features are present in the FIC that ends in 23.4, due to coreference of a pronominal NP with an NP in the preceding FIC (from phrase 23.2).

Table 4 shows the average performance of the referring expression algorithm (row labeled NP) on the 4 measures we use here. Recall is .50 (σ=.17; max=.71; min=.18), precision is .31 (σ=.097; max=.50; min=.20), fallout is .15 (σ=.06; max=.27;min=.07) and error rate is 0.19 (σ=.06; max=.31; min=.12). Recall is worse than PAUSE, CUE and human performance, and precision is better than PAUSE and CUE but worse than human performance. Note that the error rate and fallout, which in a sense are more robust measures of inaccuracy than precision, are both much better than CUE and PAUSE.

4.2.4 Additive algorithms. We report here evaluation of a simple additive method for combining the three algorithms described above. That is, a boundary is proposed if some combination of the algorithms proposed a boundary. We tested all pairwise combinations, and the combination of all three algorithms, as shown in Table 5. Precision is the most likely metric to be improved. For a composite algorithm, recall cannot be increased: if neither NP, PAUSE nor CUE found a boundary, then no combination of them can. However, the composite algorithms use narrower criteria for boundaries, which should reduce the number of false positives. The precision of the additive algorithms is indeed higher than any of the algorithms alone. PAUSE/NP has the best additive algorithm performance as measured by the summed deviation.

4.2.5 Discussion. By using average human performance as a baseline against which to evaluate algorithms, we are asking whether algorithms perform in a manner that reflects an abstraction over a population of humans, rather than whether they perform

Table 5
Additive algorithms.

	Recall	Precision	Fallout	Error	Summed Deviation
PAUSE/CUE	.69	.29	.29	.29	1.66
PAUSE/NP	.47	.42	.08	.13	1.42
CUE/NP	.36	.34	.09	.15	1.59
PAUSE/CUE/NP	.34	.47	.05	.12	1.43

like a typical human. No algorithm or combination of algorithms performs as well as this baseline. The referring expression algorithm (NP) performs better than the other unimodal algorithms (PAUSE and CUE), and a combination of PAUSE and NP performs best. Our results thus suggest that accurately predicting discourse segmentation involves far more than directly using known linguistic differences between discourse boundaries and non-boundaries.[15] Here we analyze some of the likely reasons for our results, to motivate the methodologies for algorithm improvement presented in the next section.

First, we must take into account the dimensions along which the three algorithms differ apart from the different types of linguistic information used. As shown in Figs. 9, 11, and 12, NP uses more knowledge than PAUSE and CUE. PAUSE and CUE each depend on only a single feature, while NP relies on 3 features. Unsurprisingly, NP performs most like humans. For both PAUSE and CUE, the recall is relatively high, but the precision is very low, and the fallout and error rate are both poor. For NP, recall and precision are not as different, precision is higher than PAUSE and CUE, and fallout and error rate are both relatively low. These results, as well as the improved performance of the additive algorithms, suggest that performance can be improved by considering more features. The algorithms presented in section 4.3 indeed use more features, as shown in Figure 6.

A second dimension to consider in comparing performance is that humans and NP assign boundaries based on a global criterion, in contrast to PAUSE and CUE. Our subjects typically use a relatively gross level of speaker intention. By default, NP assumes that the current segment continues, and assigns a boundary under relatively narrow criteria. However, PAUSE and CUE rely on cues that are relevant at the local as well as the global

15 Whittaker and Stenton (1988) also show that cue phrases are not reliable for predicting segment boundaries, and similarly argue for the use of multiple linguistic devices. In addition, our training data (like previous research) shows that pauses preceding boundaries have average longer durations. For $T_j \geq 3$, the average pause duration is .64 (σ=.65) before boundaries, and .39 (σ=1.70) before non-boundaries; for $T_j \geq 4$, the average durations are .72 (σ=.67) and .39 (σ=1.64), respectively. As will be seen in Section 4.3.2, this correlation does not translate into any high performing algorithm based primarily on pause duration.

level, and consequently assign boundaries more often. This leads to a preponderance of cases where PAUSE and CUE propose a boundary but where a majority of humans did not. However, when both PAUSE and CUE are combined with the more global NP, as in PAUSE/NP and CUE/NP, we see that performance improves. These results suggest that another way to improve performance is to consider more sophisticated methods for combining features across the three types of linguistic devices.

4.3 Developing new hypotheses by combining multiple knowledge sources

In this section we present two methods for developing segmentation algorithms that combine the features of multiple linguistic devices in more complex ways than simply combining the outputs of independent algorithms. Our first method relies on an analysis of the errors made by the best performing algorithm. Our second method uses machine learning tools to automatically construct segmentation algorithms from a large set of input features: features used in our previous experiments, enhancements to hand-coded features, and new features obtainable automatically from our transcripts. Both methods consider much more knowledge than previously considered by ourselves or others, and result in algorithms that exhibit marked improvements in performance. We present our results using two sets of statistically-validated boundaries: those derived using a significance level of .0001 (corresponding to $T_j \geq 4$ subjects, as in the previous section), and those derived using a less conservative level of .02 (corresponding to $T_j \geq 3$ subjects).

4.3.1 Error analysis. To improve performance, we analyzed the two types of IR errors, defined in Figure 8 above, made by the original NP algorithm on the training data (Passonneau and Litman, 1993). Type "b" errors, mis-classification of non-boundaries, were reduced by redefining the coding features pertaining to clauses and NPs. Most "b" errors correlated with one of two kinds of the information used in the NP algorithm, identification of clauses (FICs) and of inferential links. The redefinition of FIC motivated by error analysis led to fewer clauses. For example, FIC assignment depends in part on filtering out clausal interjections, utterances that have the syntactic form of clauses but that function as interjections. These include phrases like *let's see, let me see, I don't know* when they occur with no overt or implied verb phrase argument. The extensional definition of clausal interjections was expanded, thus certain utterances were no longer classed as FICs under the revised coding. Other changes to the definition of FICs pertained to sentence fragments, unexpected clausal arguments, and embedded speech. Because the algorithm assigns boundaries between units that contain at most one FIC, reducing the number of FICs in a narrative can reduce the number of proposed boundaries.

Error analysis also led to a redefinition of *infer*, and to the inclusion of new types of inferential relations that an NP referent might have to prior discourse. Previously, *infer* was a relation between the referent of an NP in one utterance, and the referent of an NP in a previous utterance. This was loosened to include referential links between an NP referent and referents mentioned in or inferable from any part of the previous utterance. For example, discourse deixis (demonstrative reference to a referent derivable from prior discourse (Schiffman, 1985; Webber, 1991)) was added to the types of inferential links to code for. In the second utterance of, "*The storm is still raging, and that's why the plane is grounded*," the demonstrative pronoun illustrates an example of discourse deixis. Expanding the definition of *infer* also reduces the number of proposed boundaries: recall that the algorithm does not assign a boundary if there is an inferential link between an NP in the current utterance unit and the prior utterance unit.

Three types of inference relations linking successive clauses (C_{i-1}, C_i) were added (originally there were 5 types (Passonneau, 1994)). Now, a pronoun (e.g., *it, that, this*) in C_i referring to an action, event or fact inferable from C_{i-1} provides an inferential link. So does an implicit argument, as in Figure 13, where the missing argument of *notice* is inferred to be the event of the pears falling. The third case is where an NP in C_i is

Cl.	Phr.	
6	3.01	[1.1 [.7] A-nd] he's not really.. doesn't seem to be paying all that much attention
7		[.55? because [.45]] you know *the pears fall*$_i$,
8	3.02	and.. he doesn't really notice ($\emptyset_i$),

Figure 13
Inferential link due to implicit argument.

Table 6
Performance on training set.

Average	Recall	Precision	Fallout	Error	Summed Deviation
Boundary Threshold = 3					
NP	.42	.40	.14	.22	1.54
Std. Dev.	.18	.12	.06	.07	.34
EA	.58	.62	.08	.14	1.02
Std. Dev.	.14	.10	.04	.05	.18
Boundary Threshold = 4					
NP	.50	.31	.15	.19	1.53
Std. Dev.	.17	.10	.06	.06	.23
EA	.70	.47	.10	.12	1.05
Std. Dev.	.16	.06	.04	.03	.15

described as part of an event that results directly from an event mentioned in C_{i-1}.

"C" type errors (cf. Figure 8), mis-classification of boundaries, often occurred where prosodic and cue features conflicted with NP features. The original NP algorithm assigned boundaries wherever the three values '-coref', '-infer', '-global.pro' co-occurred. Experiments led to the hypothesis that the most improvement came by assigning a boundary if the *cue-prosody* feature had the value 'complex', even if the algorithm would not otherwise assign a boundary, as shown in Figure 14. Figure 10 shows boundaries assigned by the resulting algorithm (EA, for error analysis).

Table 6 presents the average IR scores across the narratives in the *training* set for the NP and EA algorithms. The top half of the table reports results for boundaries that at least 3 subjects agreed upon (T=3), and the lower half for boundaries using a threshold value of 4 (T=4), where NP duplicates the figures from Table 4. Going by the summed

if (coref = -coref **and** infer = -infer **and** global.pro = -global.pro) **then** *boundary*
  elseif cue-prosody = complex **then** *boundary*
else *non-boundary*

Figure 14
EA algorithm.

deviations, the overall performance is about the same, although variation around the mean is lower for T=4. The figures illustrate a typical tradeoff between precision and recall; where one goes up, the other goes down. All scores are better for EA.

Table 7 shows the results of the tuned algorithm on the 5 randomly selected test narratives for NP and EA. Performance on the test set is slightly better overall for T=4, as shown by lower summed deviations. The NP results are very similar to the training set except for worse precision. Thus, despite the high standard deviations, 10 narratives seems to have been a sufficient sample size for evaluating the initial NP algorithm. EA results are better than NP in Table 7 or Table 6. This is strong evidence that the tuned algorithm is a better predictor of segment boundaries than the original NP algorithm. The test results of EA are of course worse than the corresponding training results, particularly for precision (.44 versus .62). This confirms that the tuned algorithm is over calibrated to the training set. Using summed deviations as a summary metric, EA's improvement is about 1/3 of the distance between NP and human performance.

The standard deviations in Tables 6-7 are often close to 1/4 or 1/3 of the reported averages. This indicates a large amount of variability in the data, reflecting wide differences across narratives (speakers) in the training set with respect to the distinctions recognized by the algorithm. Although the high standard deviations show that the tuned

Table 7
Performance on test set.

Average	Recall	Precision	Fallout	Error	Summed Deviation
Boundary Threshold = 3					
NP	.44	.29	.16	.21	1.64
Std. Dev.	.18	.17	.07	.05	.32
EA	.50	.44	.11	.17	1.34
Std. Dev.	.21	.06	.03	.04	.29
Boundary Threshold = 4					
NP	.56	.25	.16	.20	1.55
Std. Dev.	.29	.15	.08	.05	.23
EA	.60	.37	.11	.15	1.30
Std. Dev.	.20	.05	.03	.02	.17

algorithm is not well fitted to each narrative, it is likely that it is over specialized to the training sample in the sense that test narratives are likely to exhibit further variation.

4.3.2 Machine learning. While error analysis is a useful method for refining an existing feature representation, it does not facilitate experimentation with large sets of multiple features simultaneously. To address this, we turned to machine learning to automatically develop algorithms from large numbers of both training examples and features.

We use the machine learning program C4.5 (Quinlan, 1993) to automatically develop segmentation algorithms from our corpus of coded narratives, where each potential boundary site has been classified and represented as a set of linguistic features. The first input to C4.5 specifies the names of the classes to be learned (*boundary* and *non-boundary*), and the names and potential values of a fixed set of coding features (Figure 6). The second input is the training data, i.e., a set of examples for which the class and feature values (as in Figure 7) are specified. Our training set of 10 narratives provides 1004 examples of potential boundary sites. The output of C4.5 is a classification algorithm expressed as a decision tree, which predicts the class of a potential boundary given its set of feature values.

Because machine learning makes it convenient to induce decision trees under various conditions, we have performed numerous experiments varying the number of features used, the definitions used for classifying a potential boundary site as *boundary* or *non-boundary* and the options available for running the C4.5 program. Figure 15 shows one of the highest-performing learned decision trees from our experiments. This decision tree was learned under the following conditions: all of the features shown in Figure 6 were used to code the training data, boundaries were classified using a threshold of 3 subjects, and C4.5 was run using only the default options.[16] The decision tree predicts the class

16 The manually derived segmentation algorithm evaluates boundary assignment incrementally, i.e.,

```
if before = -sentence.final.contour then non-boundary
elseif before = +sentence.final.contour then
  if coref = NA then non-boundary
  elseif coref = +coref then
    if after = +sentence.final.contour then
      if duration ≤ 1.3 then non-boundary
      elseif duration > 1.3 then boundary
    elseif after = -sentence.final.contour then
      if word1 ∈ {also,basically,because,finally,first,like,
        meanwhile,no,oh,okay,only,see,so,well,where,NA}
        then non-boundary
      elseif word1 ∈ {anyway,but,now,or,then} then boundary
      elseif word1 = and then
        if duration ≤ 0.6 then non-boundary
        elseif duration > 0.6 then boundary
  elseif coref = -coref then
    if infer = +infer then non-boundary
    elseif infer = NA then boundary
    elseif infer = -infer then
      if after = -sentence.final.contour then boundary
      elseif after = +sentence.final.contour then
        if cue1 = true then
          if global.pro = NA then boundary
          elseif global.pro = -global.pro then boundary
          elseif global.pro = +global.pro then
            if duration ≤ 0.65 then non-boundary
            elseif duration > 0.65 then boundary
        elseif cue1 = false then
          if duration > 0.5 then non-boundary
          elseif duration ≤ 0.5 then
            if duration ≤ 0.35 then non-boundary
            elseif duration > 0.35 then boundary
```

Figure 15
Learned decision tree for segmentation.

of a potential boundary site based on the features *before*, *after*, *duration*, cue_1, $word_1$, *coref*, *infer*, and *global.pro*. Note that although not all available features are used in the tree, the included features represent 3 of the 4 general types of knowledge (prosody, cue phrases and noun phrases). Each level of the tree specifies a test on a single feature, with a branch for every possible outcome of the test.[17] A branch can either lead to the assignment of a class, or to another test. For example, the tree initially branches based on the value of the feature *before*. If the value is '-sentence.final.contour' then the first

utterance-by-utterance, after computing the features for the current utterance (or FIC unit). This allows relative information about previous boundaries to be used in deriving the *global.pro* feature. By allowing machine learning to use *global.pro*, we are testing whether characterizing the use of referring expressions (certain pronouns) in terms of relative knowledge about segments (whether the current referent was already mentioned in the current segment) is useful for classifying the current boundary site. Although none of the other features are derived using classification knowledge of any other potential boundary sites, note that *global.pro* does not encode the boundary/non-boundary classification of the particular site in question. Furthermore, even when machine learning does not use *global.pro* (as with the "Learning 2" algorithm discussed below), performance does not suffer.

17 The actual tree branches on every value of $word_1$; the figure merges these branches for clarity.

branch is taken and the potential boundary site is assigned the class *non-boundary*. If the value of *before* is '+sentence.final.contour' then the second branch is taken and the feature *coref* is tested. Figure 10 shows boundaries (ML) assigned by this algorithm.

The performance of this learned decision tree averaged over the 10 training narratives is shown in Table 8, on the line labeled "Learning 1". The line labeled "Learning 2" shows the results from another machine learning experiment, in which one of the default C4.5 options used in "Learning 1" is overridden. The default C4.5 approach creates a separate subtree for each possible feature value; as detailed in (Quinlan, 1993), this approach might not be appropriate when there are many values for a feature, which is true for features such as $word_1$ and $word_2$. In "Learning 2" C4.5 allows feature values to be grouped into one branch of the decision tree. While the "Learning 2" tree is more complex than the tree of Figure 15, it does have slightly better performance. The "Learning 2" decision tree predicts the class of a potential boundary site based on the features *before*, *duration*, cue_1, $word_1$, $word_2$, *coref*, *infer*, and *cue-prosody*. At T=3, "Learning 1" performance is comparable to human performance (Table 3), and "Learning 2" is slightly better than humans; at T=4, both learning conditions are superior to human performance. The results obtained via machine learning are also better than the results obtained using error analysis ("EA"in Table 6), primarily due to better precision. In general, the machine learning results have slightly greater variation around the average.

The performance of the learned decision trees averaged over the 5 test narratives is shown in Table 9. Comparison of Tables 8 and 9 shows that, as with the error analysis results (and as expected), average performance is worse when applied to the testing rather than the training data, particularly with respect to precision. However, the best machine learning performance is an improvement over our previous best results ("EA" in Table 7). For T=3, "Learning 1" is comparable to "EA" while "Learning 2" is better. For T=4, "EA" is better than "Learning 1", but "Learning 2" is better still. As with the training data, "EA" has somewhat less variation around the average.

We also use the resampling method of *cross-validation* (Weiss and Kulikowski, 1991) to estimate performance, which averages results over multiple partitions of a sample into test versus training data. We performed 10 runs of the learning program, each using 9 of the 10 training narratives for that run's training set (for learning the tree) and the remaining narrative for testing. Note that for *each* iteration of the cross-validation, the learning process begins from scratch and thus each training and testing set are still disjoint. While this method does not make sense for humans, computers can truly ignore previous iterations. For sample sizes in the hundreds (our 10 narratives provide 1004 examples) 10-fold cross-validation often provides a better performance estimate than the hold-out method (Weiss and Kulikowski, 1991). Results using cross-validation are shown in Table 10, and are better than the estimates obtained using the hold-out method (Table 9), with the major improvement coming from precision.

Finally, Table 11 shows the results from a set of additional machine learning experiments, in which more conservative definitions of boundary are used. For example, using a threshold of 7 subjects yields the set of consensus boundaries, as defined in (Hirschberg and Nakatani, 1996). Comparison with Table 9 shows that for T=5, "Learning 1" rather than "Learning 2" is the better performer. However, the more interesting result is that

Table 8
Performance on training set.

Average	Recall	Precision	Fallout	Error	Summed Deviation
Boundary Threshold = 3					
Learning 1	.54	.76	.04	.11	.85
Std. Dev.	.18	.12	.02	.04	.28
Learning 2	.59	.78	.03	.10	.76
Std. Dev.	.22	.12	.02	.04	.29
Boundary Threshold = 4					
Learning 1	.47	.84	.01	.07	.77
Std. Dev.	.26	.18	.02	.04	.42
Learning 2	.53	.77	.02	.07	.79
Std. Dev.	.23	.18	.02	.03	.35

Table 9
Performance on test set.

Average	Recall	Precision	Fallout	Error	Summed Deviation
Boundary Threshold = 3					
Learning 1	.43	.48	.08	.16	1.34
Std. Dev.	.21	.13	.03	.05	.36
Learning 2	.47	.50	.09	.16	1.27
Std. Dev.	.18	.16	.04	.07	.42
Boundary Threshold = 4					
Learning 1	.31	.41	.06	.13	1.47
Std. Dev.	.29	.15	.08	.05	.23
Learning 2	.39	.52	.05	.11	1.24
Std. Dev.	.20	.05	.03	.02	.17

for T=6 and T=7, the learning approach has an important limitation with respect to the boundary classification task. In particular, the way in which C4.5 minimizes error rate isn't an effective strategy when the distribution of the classes is highly skewed. For both T=6 and T=7, extremely few of the 1004 training examples are classified as *boundary* (40 and 19 examples, respectively). C4.5 minimizes the error rate by always predicting *non-boundary*. For example, for T=6, because only 4% of the training examples are boundaries, C4.5 achieves an error rate of 4% by always predicting *non-boundary*. However, this low error rate is achieved at the expense of the other metrics. Using the terminology of Figure 8, since the algorithm never predicts the class *boundary*, it is necessarily the case that a=0, b=0, recall=0, and precision is undefined ("-" in Table 11). In addition, for T=7, two of the 5 test sets happen to contain no boundaries; for these cases

Table 10
Using 10-fold cross-validation.

Average	Recall	Precision	Fallout	Error	Summed Deviation
Boundary Threshold = 3					
Learning 1	.43	.63	.05	.15	1.14
Std. Dev.	.19	.16	.03	.03	.24
Learning 2	.46	.61	.07	.15	1.15
Std. Dev.	.20	.14	.04	.03	.21
Boundary Threshold = 4					
Learning 1	.30	.71	.02	.10	1.11
Std. Dev.	.15	.19	.02	.03	.26
Learning 2	.35	.52	.04	.11	1.28
Std. Dev.	.19	.24	.02	.04	.40

Table 11
Performance on test set for higher boundary thresholds.

Average	Recall	Precision	Fallout	Error	Summed Deviation
Boundary Threshold = 5					
Learning 1	.31	.46	.03	.08	1.35
Std. Dev.	.11	.30	.02	.02	.43
Learning 2	.28	.39	.04	.08	1.46
Std. Dev.	.17	.24	.03	02	.45
Boundary Threshold = 6					
Learning 1	0	-	0	.04	-
Std. Dev.	0	-	0	.02	-
Learning 2	0	-	0	.04	-
Std. Dev.	0	-	0	.02	-
Boundary Threshold = 7					
Learning 1	-	-	0	.02	-
Std. Dev.	-	-	0	.02	-
Learning 2	-	-	0	.02	-
Std. Dev.	-	-	0	.02	-

Table 12
Paired comparison of EA and automated algorithm results, using Student's T (df=4).

Boundary Threshold = 4		
Comparison	Metric	Probability
EA with Learning 1	Recall	$p \leq .20$
EA with Learning 1	Fallout	$p \leq .10$
EA with Learning 2	Recall	$p \leq .25$
EA with Learning 2	Error	$p \leq .20$
Boundary Threshold = 3		
EA with Learning 1	Precision	$p \leq .0005$
EA with Learning 1	Error	$p \leq .10$

c=0 and thus the value of recall is also sometimes undefined. The problem of unbalanced data is not unique to the boundary classification task. Current work in machine learning is exploring ways to induce patterns relevant to the minority class, for example, by allowing users to explicitly specify different penalties for false positive and false negative errors (Lewis and Catlett, 1994). (In contrast, C4.5 assumes that both types of errors are penalized equally.) Other researchers (e.g., (Hirschberg, 1991)) have proposed sampling the majority class examples in a training set in order to produce a more balanced training sample.

4.3.3 Discussion. We have presented two methods for developing segmentation hypotheses using multiple linguistic features. The first method, error analysis, tunes features and algorithms based on analysis of training errors. The second method, machine learning, *automatically* induces decision trees from coded corpora. Both methods rely on an enriched set of input features compared to our previous work. With each method, we have achieved marked improvements in performance compared to our previous work and are approaching human performance. Quantitatively, the machine learning versus EA methods differ only on certain metrics, and bear a somewhat inverse relation to one another for boundaries defined by $T \geq 4$ versus $T \geq 3$. Table 12 of comparisons between EA and the two machine learning conditions indicates which differences are statistically significant by indicating the probability of a paired comparison on each of the five test narratives using Student's T test. For the T=4 boundaries, the superior recall of EA compared with conditions 1 and 2 of the automated algorithms is significant. Conversely, the superior fallout of condition 1 and superior error rate of condition 2 are significant. For the T=3 boundaries, the differences are not statistically significant for condition 2, but for condition 1, precision and error rate are both superior, and the difference as compared with EA is statistically significant. The difference that is both the largest and the most statistically significant is the higher precision of the condition 1 automated algorithm. Qualitatively, the algorithms produced by error analysis are more intuitive and easier to understand than those produced by machine learning. Furthermore, note that the machine learning algorithm used the changes to the coding features that resulted from the error analysis. This suggests that error analysis is a useful method for understanding how to best code the data, while machine learning provides a cost-effective (and automatic) way to produce an optimally performing algorithm given a good feature representation.

5. Conclusion and Future Directions

Our initial hypotheses regarding discourse segmentation were that multi-utterance segment units reflect discourse coherence, and that while the semantic dimensions of this coherence may vary, it arises partly from consistency in the speaker's communicative goals (Grosz and Sidner, 1986; Polanyi, 1988). The results from the first part of our study (section 3) support these hypotheses. On a loosely defined linear segmentation task, the number of times different naive subjects identify the same segment boundaries in a given narrative transcript is extremely significant. Across the 20 narratives, statistical significance arises where at least 3 or 4 out of 7 subjects agree on the same boundary location, depending on an arbitrary choice between probabilities of .02 versus .0001 as the significance threshold. We conclude that the segment boundaries identified by at least 3 or 4 of our subjects provide a statistically validated annotation to the narrative corpus corresponding to segments having relatively coherent communicative goals.

Before making concluding remarks on part two of our study, we mention a few questions for future work on segmentation. We believe our results both confirm the utility of abstracting from the responses of relatively many naive subjects (our method), and indicate a strong potential for developing coding protocols using smaller numbers of trained coders (as in (Hirschberg and Grosz, 1992; Nakatani, Hirschberg, and Grosz, 1995; Hirschberg and Nakatani, 1996)). The use of an even larger number of naive subjects might yield a finer-grained set of segments (cf. (Rotondo, 1984; Swerts, 1995)). This is an important dimension of difference between the two sets of segments we use: segments identified by a minimum of 4 subjects are larger and fewer in number than those identified by a minimum of 3. In addition, performance can be improved by taking into account that some segment boundary locations may be relatively fuzzy, as we discuss in (Passonneau and Litman, To appear). Finally, differences in segmentation may reflect different interpretations of the discourse. By looking at our subjects' segment descriptions, we were able to raise this issue in (Passonneau and Litman, To appear).

The second part of our study (section 4) concerned the algorithmic identification of segment boundaries based on various combinations of three types of linguistic input - referential noun phrases, cue phrases, and pauses. We first evaluated an initial set of three algorithms, each based on a single type of linguistic input, and their additive combinations. Our results showed that the algorithms performed quite differently from one another on boundaries identified by at least 4 subjects on a test set of 10 narratives from our corpus. In particular, the NP algorithm (which used 3 features) outperformed both the cue phrase and pause algorithms (each of which used only a single feature). While none of the algorithms approached human performance, the facts that performance improved with the number of features coded, and by combining algorithms in a simple additive way, suggested directions for improvement. We applied two training methods, error analysis and machine learning, to the previous test set of 10 narratives. Richer linguistic input and more sophisticated methods of combining linguistic data led to significant improvements in performance when the new algorithms were evaluated on a test set of 5 new narratives. The best performing algorithm resulted from the machine learning experiment where certain default options were overridden ("Learning 2" in Table 9). For the T=4 boundary set, "Learning 2" recall was 53% as good as humans, precision was 95% as good, fallout was better than humans, and error (11%) was almost as low as that of humans (10%). Thus the main need for improvement is in recall.

A comparison of results on two sets of boundaries, those identified by at least 3, versus those identified by at least 4 subjects, shows roughly comparable performance. The "Learning 1" algorithm performs better on the set defined by T=3 (Table 9); Error Analysis (Table 7) and "Learning 2" (Table 9) perform better on the T=4 set. We have not yet determined what causes these differences, although in an early paper on our pilot study, we reported that there is a strong tendency for recall to increase and precision to decrease as boundary strength increases (Passonneau and Litman, 1993). On the one hand, performance was consistently improved by enriching the linguistic input. On the other hand, there is wide performance variation around the mean. Despite this variation, as we pointed out in (Litman and Passonneau, 1995a), there are certain narratives that the "NP", "EA" and both machine learning algorithms perform similarly well or poorly on. These observations indicate a need for further research regarding the interaction among variation in speaker style, granularity of segmentation, and richness of the linguistic input.

Finally, while our results are quite promising, how generally applicable are they, and do results such as ours have any practical import? As discussed in section 2, the ability to both segment discourse and to correlate segmentation with linguistic devices has been demonstrated in dialogues and monologues, using both spoken and written corpora, across a wide variety of genres (e.g., task-oriented, advice-giving, information-query, expository, directions, and newspapers). Studies such as these suggest that our methodologies and/or results have the potential of being applicable to more than spontaneous narrative monologues.

As for the utility of our work, even though the algorithms in this paper were produced using some features that were manually coded, once developed, they could be used in reverse to enhance the comprehensibility of text generation systems or the naturalness of text-to-speech systems that already attempt to convey discourse structure (e.g., systems such as (Moore and Paris, 1993; Hirschberg, 1990)). For example, given the algorithm shown in Figure 14, a generation system could better convey its discourse boundaries by constructing associated utterances where the values of *coref*, *infer*, and *global.pro* are as shown in the first line of the figure, or (for only a spoken language system) where the value of *cue-prosody* is complex. In related work, we have tested the hypothesis that the use of a discourse focus structure based on the Pear segmentation data improves performance of a generation algorithm, thus providing a quantitative measure of the utility of the segmentation data (Passonneau, Forthcoming). There we presents results of an evaluation of an NP generation algorithm under various conditions. The input to the algorithm consisted of semantic information about utterances in a Pear narrative, such as the referents mentioned in the utterance. Output was evaluated against what the human narrator actually said. When the input to the algorithm included a grouping of discourse referents into focus spaces derived from discourse segments, performance improved by 50%.

In addition, if our results were fully automated, they could also be used to enhance the ability of *understanding* systems to recognize discourse structure, which in turn improves tasks such as information retrieval (Hearst, 1994) and plan recognition (Litman and Allen, 1990). Recent results suggest that many of our manually coded features have the promise of being automatically coded. Given features largely output by a speech recognition system, Wightman and Ostendorf (1994) automatically recognize prosodic phrasing with 85-86% accuracy; this accuracy is only slightly less than human-human accuracy. Similarly, although our spoken corpus was manually transcribed, this could have been automated using speech recognition (although this would introduce further sources of error). In (Aone and Bennett, 1995), machine learning is used to automatically derive anaphora resolution algorithms from automatically produced feature representations; the learned algorithms outperform a manually derived system (whose average recall and precision was 66.5% and 72.9%, respectively). Finally, the results of (Litman, 1994) show that there are comparably performing alternatives to the cue phrase algorithm used here, including some that use feature sets that can be fully coded automatically.

A. Instructions for Subjects Performing Linear Segmentation

The instructions given to the subjects in the segmentation study described the segmentation task in general terms, and illustrated the task by example. There were two sets of instructions, differing only in the sample discourse used to illustrate the task. Each subject received only one set of instructions. Here we show one of the two instruction sets.

ATTENTION: You are being asked to participate in a study of how ordinary spoken language is used and understood. You will be asked to read a transcript of a person describing a movie, and to mark down your impressions of how the movie description is structured. You will not be GRADED on your RESPONSES, because there are no CORRECT answers. However, the task requires that you concentrate, and follow the instructions precisely. Your responses can only be used in the study if you follow the directions exactly.

The upper right hand corner of each transcript has a speaker number. Please fill in the following information, including the speaker number, and then proceed to the instructions below.

Your Name:
Today's Date:
Speaker Number:

INSTRUCTIONS:
The movie narrations you are about to read were collected from twenty people who had all watched the same movie. The movie was made specifically for this task. Each speaker was asked to describe the movie to someone else who had never seen it before. The person listening to the movie narration didn't say anything. The transcription you have been given accurately records everything that each speaker said while describing the movie. As you will notice, the speakers often repeat themselves, or correct themselves, or make speech errors. This is the way people actually talk. When you read the narration, try to imagine it being told to you by someone you know.
Each speaker was asked to describe a movie he or she just saw in his or her own words. In order to describe the movie, the speaker had to make numerous choices and decisions about what to say next. Here are some examples of things the speaker had to do, or decided to do:

- The speaker had to remember how the story was structured, and to convey this structure to the listener.
- The speaker sometimes decided to inform the listener about the techniques used in making the movie. For example, the movie had a vivid sound track, but no one in the movie talked. Many of the speakers talk about the sound track, or other aspects of how the movie was made.
- The speaker sometimes informed the listener about his or her own responses to the movie. For example, some speakers made guesses about why the people in the movie acted the way they did.

To summarize, you should think of each movie narration as resulting from many decisions made by the speaker about what to do next. You will be asked to evaluate what the speaker was doing at each point in the narration. Write directly on the transcript, FOLLOWING THE INSTRUCTIONS BELOW.

Read through the transcript and draw a HORIZONTAL line ACROSS THE PAGE between COMPLETE text lines (utterances) where you think the speaker started DOING SOMETHING NEW. A COMPLETE text line is one that ends in a ".", "," or "?", or punctuation followed by a quote mark. NOTE: some COMPLETE text lines are continued over more than one printed line. IF YOU DRAW LINES IN THE WRONG PLACE, YOUR RESULT WILL BE DISCARDED!

CORRECT	'cause he's kind of romantic about the West. Well, he gets out to this small army post and it's abandoned.
	He's all alone in the wilderness, with lots of supplies but no people.

NOT CORRECT	'cause he's kind of romantic about the West. Well, he gets out to this small army post and
	it's abandoned. He's all alone in the wilderness, with lots of supplies but no people.

In the wide left hand margin, say in abbreviated form what the speaker is doing. On the next page is an example of how to proceed. Note that the example may be INCOMPLETE. You are free to use any criteria in deciding what the narrator of your transcript is *doing*.

`speaker recommends movie`	Well it's really a great movie, really beautiful scenery. You should see it, I recommend it, I really do.
`introduces main character, and where the movie takes place`	The first part just sets up how the Kevin Kostner character goes out West. He's a soldier in the Civil War, and he was a hero so he can be posted where he wants, and he asks to go out to the frontier, out to North Dakota, 'cause he's kind of romantic about the West.
`describes the fort and the countryside`	Well, he gets to this army post and it's abandoned. He's all alone in the wilderness, with lots of supplies but no people. It's beautiful country, really wide open, hardly any trees, golden grass waving in the breeze, all like that.
`introduces the Indians trying to steal the horse`	And some Indians try to steal his horse. Some Sioux. Actually, the ones who try first are kids. But his horse is smart and gets away and comes back to him.
`making friends with the Sioux`	The Sioux start to visit him, and they make friends,
`how they start to communicate, and learn each other's word for buffalo`	and the first word they can understand is buffalo. He prances around like a buffalo, and one Indian keeps saying "This man has lost his mind." Cause he looks ridiculous, he really does. But the other Indian says "No." "He's trying to make us understand–buffalo." So the Indian says to Dunbar that's the guy's name, Lieutenant Dunbar. He says, "Ichtonka". That's the Sioux word for buffalo. And Dunbar says, "Yeah! Buffalo!" And so they go like that: "Buffalo." "Ichtonka." "Buffalo." "Ichtonka." And they all smile at each other.

Acknowledgments
The authors wish to thank J. Catlett, W. Chafe, K. Church, W. Cohen, J. DuBois, B. Gale, V. Hatzivassiloglou, M. Hearst, J. Hirschberg, J. Klavans, D. Lewis, E. Levy, K. McKeown, E. Siegel, and anonymous reviewers of previous papers for helpful comments, references and resources. In addition, the authors wholeheartedly thank the anonymous reviewers for their very thorough commentary, and extremely helpful suggestions. Both authors' work was partially supported by DARPA and ONR under contract N00014-89-J-1782; Passonneau was also partly supported by NSF grant IRI-91-13064. Bellcore did not support Passonneau's work.

References

Abney, Steven P. 1990. Rapid incremental parsing with repair. In *Proceedings of the 6th New OED Conference: Electronic Text Research*, pages 1–9.

Anderson, Anne H. et al. 1991. The HCRC Map Task corpus. *Language and Speech*, pages 351–366.

Aone, Chinatsu and Scott W. Bennett. 1995. Evaluating automated and manual acquisition of anaphora resolution strategies. In *Proc. of 33rd ACL*, pages 122–129.

Brieman, Leo, J. Friedman, R. Olshen, and C. Stone. 1984. *Classification and Regression Trees.* Wadsworth and Brooks.

Butterworth, Brian. 1975. Hesitation and semantic planning in speech. *Journal of Psycholinguistic Research*, 4:75–87.

Butterworth, Brian. 1980. Evidence from pauses in speech. In Brian Butterworth, editor, *Language Production.* Academic Press, London, pages 155–176.

Carberry, Sandra. 1990. *Plan Recognition in Natural Language Dialogue.* MIT Press, Cambridge, MA.

Carletta, Jean. 1996. Assessing agreement on classification tasks: the kappa statistic. *Computational Linguistics.*

Chafe, Wallace L. 1980a. The deployment of consciousness in the production of a narrative. In W. L. Chafe, editor, *The Pear Stories: Cognitive, Cultural and Linguistic Aspects of Narrative Production.* Ablex Publishing Corporation, Norwood, NJ.

Chafe, Wallace L. 1980b. *The Pear Stories: Cognitive, Cultural and Linguistic Aspects of Narrative Production.* Ablex Publishing Corporation, Norwood, NJ.

Cochran, William G. 1950. The comparison of percentages in matched samples. *Biometrika*, 37:256–266.

Cohen, Robin. 1984. A computational theory of the function of clue words in argument understanding. In *Proc. of COLING84*, pages 251–258, Stanford.

Dale, Robert. 1992. *Generating Referring Expressions: Constructing Descriptions in a Domain of Objects and Processes.* MIT Press, Cambridge, MA.

Duncan, Starkey D. and Donald W. Fiske. 1977. *Face-to-face Interaction: Research, Methods and Theory.* Lawrence Erlbaum Associates, Hillside, New Jersey.

Flammia, Giovanni and Victor Zue. 1995. Empirical evaluation of human performance and agreement in parsing discourse constituents in spoken dialogue. In *Eurospeech 1995.*

Gale, William, Ken W. Church, and David Yarowsky. 1992. Estimating upper and lower bounds on the performance of word-sense disambiguation programs. In *Proc. of 30th Association for Computational Linguistics (ACL)*, pages 249–256, Newark, Delaware.

Grosz, Barbara and Julia Hirschberg. 1992. Some intonational characteristics of discourse structure. In *Proc. of the International Conference on Spoken Language Processing (ICSLP).*

Grosz, Barbara J. 1977. *The Representation and Use of Focus in Dialogue Understanding.* Ph.D. thesis, University of California, Berkeley.

Grosz, Barbara J., Aravind K. Joshi, and Scott Weinstein. 1995. Centering: A framework for modeling the local coherence of discourse. *Computational Linguistics*, 21.2:203–226.

Grosz, Barbara J. and Candace L. Sidner. 1986. Attention, intentions and the structure of discourse. *Computational Linguistics*, 12:175–204.

Hearst, Marti A. 1993. TextTiling: A quantitative approach to discourse segmentation. Technical Report 93/24, Sequoia 2000 Technical Report, University of California, Berkeley.

Hearst, Marti A. 1994. Multi-paragraph segmentation of expository text. In *Proc. of 32nd ACL.*

Hindle, Donald. 1983. Deterministic parsing of syntactic non-fluencies. In *Proc. of 21st ACL*, pages 123–128.

Hirschberg, Julia. 1990. Accent and discourse context: Assigning pitch accent in synthetic speech. In *Proceedings of the Eighth National Conference on Artificial Intelligence (AAAI)*, pages 952–957.

Hirschberg, Julia. 1991. Using text analysis to predict intonational boundaries. In *Proceedings of the Second European Conference on Speech Communication and Technology.*

Hirschberg, Julia and Barbara Grosz. 1992. Intonational features of local and global discourse structure. In *Proc. of the Darpa Workshop on Spoken Language.*

Hirschberg, Julia and Diane Litman. 1993. Empirical studies on the disambiguation of cue phrases. *Computational Linguistics*, 19(3):501–530.

Hirschberg, Julia and Christine H. Nakatani. 1996. A prosodic analysis of discourse segments in direction-giving monologues. In *Proc. of 34th ACL*, pages 286–293.

Hirschberg, Julia and Janet Pierrehumbert. 1986. The intonational structuring of discourse. In *Proc. of 24th ACL.*

Hobbs, Jerry R. 1979. Coherence and coreference. *Cognitive Science*, 3(1):67–90.

Hwang, Chung H. and Lehnart K. Schubert. 1992. Tense trees as the 'fine structure' of discourse. In *Proc. of 30th ACL*, pages 232–240.

Isard, Amy and Jean Carletta. 1995. Replicability of transaction and action coding in the Map Task Corpus. In *AAAI 1995 Spring Symposium Series: Empirical Methods in Dicourse Interpretation and Generation*, pages 60–66.

Kozima, Hideki. 1993. Text segmentation based on similarity between words. In *Proc. of 31st ACL (Student Session)*, pages 286–288.

Krippendorff, Klaus. 1980. *Content Analysis: An Introduction to Its Methodology.* Sage Publications, Beverly Hills, CA.

Levy, Elena. 1984. *Communicating Thematic Structure in Narrative Discourse: The Use of Referring Terms and Gestures.* Ph.D. thesis, University of Chicago.

Lewis, David D. and J. Catlett. 1994. Heterogeneous uncertainty sampling for supervised learning. In W. W. Cohen and H. Hirsh, editors, *Proceedings of the Eleventh International Conference on Machine Learning (ML-94)*, pages 148–156. Morgan Kaufmann.

Linde, Charlotte. 1979. Focus of attention and the choice of pronouns in discourse. In T. Givon, editor, *Syntax and Semantics: Discourse and Syntax.* Academic Press, New York, pages 337–354.

Litman, Diane J. 1994. Classifying cue phrases in text and speech using machine learning. In *Proc. of the 12th National Conference on Artificial Intelligence (AAAI)*, pages 806–813.

Litman, Diane J. and James Allen. 1990. Discourse processing and commonsense plans. In P. R. Cohen, J. Morgan, and M. E. Pollack, editors, *Intentions in Communication.* MIT Press, Cambridge, MA.

Litman, Diane J. and Rebecca J. Passonneau. 1993. Empirical evidence for intention-based discourse segmentation. In *Proc. of the ACL Workshop on Intentionality and Structure in Discourse Relations.*

Litman, Diane J. and Rebecca J. Passonneau. 1995a. Combining multiple knowledge sources for discourse segmentation. In *Proc. of 33rd ACL*, pages 108–115.

Litman, Diane J. and Rebecca J. Passonneau. 1995b. Developing algorithms for discourse segmentation. In *AAAI 1995 Spring Symposium Series: Empirical Methods in Dicourse Interpretation and Generation*, pages 85–91.

Mann, William C., Christian M. Matthiessen, and Sandra A. Thompson. 1992. Rhetorical structure theory and text analysis. In William C. Mann and Sandra A. Thompson, editors, *Discourse Description: Diverse Linguistic Analyses of a Fund-Raising Text.* J. Benjamins.

Mann, William C. and Sandra Thompson. 1988. Rhetorical structure theory: towards a functional theory of text organization. *TEXT*, pages 243–281.

Marslen-Wilson, William, Elena Levy, and Lorraine K. Tyler. 1982. Producing interpretable discourse: The establishment and maintenance of reference. In R. J. Jarvella and W. Klein, editors, *Speech, Place and Action.* John Wiley and Sons Ltd., New York, pages 339–378.

Mokros, Hartmut B. 1984. *Patterns of Persistence and Change in the Sequencing of Nonverbal Actions.* Ph.D. thesis, University of Chicago.

Moore, Johanna D. and Cecile Paris. 1993. Planning text for advisory dialogues:

Capturing intentional and rhetorical information. *Computational Linguistics*, 19:652–694.

Moore, Johanna D. and Martha E. Pollack. 1992. A problem for RST: The need for multi-level discourse analysis. *Computational Linguistics*, 18:537–544.

Morris, Jane and Graeme Hirst. 1991. Lexical cohesion computed by thesaural relations as an indicator of the structure of text. *Computational Linguistics*, 17:21–48.

Moser, Megan and Johanna Moore. 1995a. Investigating cue selection and placement in tutorial discourse. In *Proceedings of the 33rd ACL*.

Moser, Megan and Johanna D. Moore. 1995b. Using discourse analysis and automatic text generation to study discourse cue usage. In *AAAI 1995 Spring Symposium Series: Empirical Methods in Dicourse Interpretation and Generation*, pages 92–98.

Moser, Megan, Johanna D. Moore, and Erin Glendening. 1995. Instructions for coding Sherlock explanations: Identifying segments, relations and minimal units.

Nakatani, Christine H., Julia Hirschberg, and Barbara J. Grosz. 1995. Discourse structure in spoken language: Studies on speech corpora. In *AAAI 1995 Spring Symposium Series: Empirical Methods in Dicourse Interpretation and Generation*, pages 106–112.

Passonneau, Rebecca J. 1993. Getting and keeping the center of attention. In R. Weischedel and M. Bates, editors, *Challenges in Natural Language Processing*. Cambridge University Press.

Passonneau, Rebecca J. 1994. Protocol for coding discourse referential noun phrases and their antecedents. Technical report, Columbia University.

Passonneau, Rebecca J. Forthcoming. Using centering to relax Gricean informational constraints on discourse anaphoric noun phrases. *Language and Speech*.

Passonneau, Rebecca J. and Diane J. Litman. 1993. Intention-based segmentation: Human reliability and correlation with linguistic cues. In *Proc. of 31st ACL,*, pages 148–155.

Passonneau, Rebecca J. and Diane J. Litman. To appear. Empirical analysis of three dimensions of spoken discourse: Segmentation, coherence and linguistic devices. In E. Hovy and D. Scott, editors, *Interdisciplinary Perspectives on Discourse*. Springer Verlag, Berlin.

Pierrehumbert, Janet and Julia Hirschberg. 1987. The meaning of intonational contours in the interpretation of discourse. Technical Report TM 11225-870325-07, AT&T Bell Laboratories.

Pitrelli, J., Mary Beckman, and Julia Hirschberg. 1994. Evaluation of prosodic transcription labeling reliability in the ToBI framework. In *Proc. of ICSLP*.

Polanyi, Livya. 1988. A formal model of discourse structure. *Journal of Pragmatics*, pages 601–638.

Quinlan, John Ross. 1993. *C4.5 : Programs for Machine Learning*. Morgan Kaufmann.

Reichman, Rachel. 1985. *Getting Computers to Talk Like You and Me*. MIT Press, Cambridge, Massachusetts.

Reynar, Jeffrey C. 1994. An automatic method of finding topic boundaries. In *Proc. of 32nd ACL (Student Session)*, pages 331–333.

Rotondo, John A. 1984. Clustering analysis of subject partitions of text. *Discourse Processes*, 7:69–88.

Schiffman, Rebecca J. (Passonneau). 1985. *Discourse Constraints on 'It' and 'That': A Study of Language Use in Career-Counseling Interviews*. Ph.D. thesis, University of Chicago.

Song, Fei and Robin Cohen. 1991. Tense interpretation in the context of narrative. In *Proc. of 9th AAAI*, pages 131–136.

Stifleman, Lisa J. 1995. A discourse analysis approach to structured speech. In *AAAI 1995 Spring Symposium Series: Empirical Methods in Dicourse Interpretation and Generation*, pages 162–167.

Swerts, Marc. 1995. Combining statistical and phonetic analyses of spontaneous discourse segmentation. In *Proceedings of the 12th International Congress of Phonetic Sciences (ICPhS 95)*, volume 4, pages 208–211.

Swerts, Marc and Mari Ostendorf. 1995. Discourse prosody in human-machine interactions. In *ESCA Workshop on Spoken Dialogue Systems*, pages 205–208.

Thompson, Sandra A. and T. Ono. To appear. Interaction and syntax in the structure of conversational discourse. In E. Hovy and D. Scott, editors, *Interdisciplinary Perspectives on Discourse*. Springer Verlag, Berlin.

Walker, Marilyn and Steve Whittaker. 1990. Mixed initiative in dialogue: An investigation into discourse segmentation. In *Proc. of 28th ACL*, pages 70–78.

Walker, Marilyn A. 1995. Limited attention and discourse structure. Ms.

Webber, Bonnie L. 1988. Tense as discourse anaphor. *Computational Linguistics*, 14:113–122.

Webber, Bonnie L. 1991. Structure and ostension in the interpretation of discourse deixis. *Language and Cognitive Processes*, pages 107–135.

Weiss, Sholom M. and Casimir Kulikowski. 1991. *Computer systems that learn: classification and prediction methods from statistics, neural nets, machine learning, and expert systems.* Morgan Kaufmann.

Whittaker, Steve and Phil Stenton. 1988. Cues and control in expert-client dialogues. In *Proc. of 26th ACL*, pages 123–130.

Wightman, Colin W. and Mari Ostendorf. 1994. Automatic labeling of prosodic patterns. *IEEE Transactions on Speech and Audio Processing*, 2(4):469–481, October.

Youmans, Gilbert. 1991. A new tool for discourse analysis: The vocabulary management profile. *Language*, pages 763–790.