EMPIRICAL ANALYSIS OF THREE DIMENSIONS OF SPOKEN DISCOURSE: SEGMENTATION, COHERENCE AND LINGUISTIC DEVICES*

Rebecca J. Passonneau
Department of Computer Science
Columbia University
New York, NY 10027
becky@cs.columbia.edu

Diane J. Litman
AT&T Bell Laboratories
600 Mountain Avenue
Murray Hill, NJ 07974
diane@research.att.com

Abstract

Certain spans of utterances in a discourse, referred to here as *segments*, are widely assumed to form coherent semantic or pragmatic units, and to correlate with a variety of linguistic devices. Currently, however, there is only weak agreement or preliminary research on many fundamental issues relating to the nature of segments, and to the interaction between segmentation, coherence, and linguistic devices. In this paper we present results from our ongoing study of discourse segmentation in a corpus of spontaneous, narrative monologues, in order to present our take on these issues. First, we present an empirical methodology for collecting discourse segmentation judgements from naive subjects, where speaker intention is used as the segmentation criterion. We present a methodology for evaluating the statistical reliability of our human judgements, and show that our empirical results are indeed statistically significant, despite a large amount of variability in the data. Second, we develop three discourse segmentation algorithms (based on the linguistic devices referential noun phrases, cue words, and pauses, respectively), in order to see whether linguistic associations proposed in the literature can be used by natural language processing systems to perform segmentation. We evaluate the segmentations produced by our algorithms, using information retrieval metrics to quantify their success in predicting boundaries produced by a majority of human subjects. While our results suggest that it would be possible to algorithmically achieve human level of performance, this would appear to require numerous knowledge sources and the ability to vary within and across speakers. Finally, we qualititively demonstrate interactions with coherence, by suggesting that subjects agree on the intentions associated with segments, and that correlations with both segmentation and lingustic devices increases for particularly coherent, and coherent but discontinuous portions of the corpus.

*This paper combines and expands on two previous papers: [27] [36]. The authors wish to thank W. Chafe, K. Church, J. DuBois, B. Gale, V. Hatzivassiloglou, M. Hearst, J. Hirschberg, J. Klavans, D. Lewis, E. Levy, K. McKeown, E. Siegel, and anonymous reviewers for helpful comments, references and resources. 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.

1 DISCOURSE SEGMENTATION

A discourse consists not simply of a linear sequence of utterances,[1] but of meaningful semantic or pragmatic relations among utterances. Each utterance 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 such units and linguistic features of utterances is almost universally acknowledged in the literature on discourse. For example, previous work argues for an interdependence between particular cue words and phrases such as *anyway*, and their location relative to an utterance or text, (e.g., [19] [13] [40] [8]); the distribution and duration of pauses relative to multi-utterance units (e.g., [11] [18] [5] [3]); and the interdependence between the form of discourse anaphoric noun phrases and the relation of the current utterance to a hierarchical model of utterance actions, or to a model of focus of attention (e.g., [12] [13] [40] [44] [43]). However, there are a variety of distinct proposals regarding how to model the interdependence among the three dimensions of 1. sequences of semantically and pragmatically related utterances, 2. the units or relations they reflect, and 3. lexico-grammatical or prosodic features. We refer to these dimensions respectively as segmentation, coherence, and linguistic devices. For purposes of this paper we have deliberately avoided adopting a particular theoretical framework in order to pose open-ended questions that we address through an empirical investigation.

Fig. 1 presents an excerpt from a transcription of a spoken narrative taken from the Pear stories corpus (Chafe [6]) that we use in our study. We use the excerpt to illustrate the interdependence of segmentation, linguistic devices, and coherence. Chafe's transcriptions (partly modified for Fig. 1) represent all lexical articulations including repeated and incomplete words and phrases; non-lexical articulations such as *uh, um, tsk*; vowel lengthening as indicated by '-'; location and duration of pauses; and prosodic phrase units. A period or question mark at the end of a line indicates the end of a prosodic unit with utterance final intonation (i.e., assertion or question) and a comma at the end of a line indicates the end of a more intermediate prosodic unit (roughly corresponding to the distinction between intermediate and complete prosodic phrases in [2] [37]). We have numbered the prosodic phrases sequentially, with the second field corresponding to the intermediate units within complete prosodic phrases.

We have grouped together certain sequences of utterances and given them segment labels, showing only the first two utterances of segment 3. Three of the many linguistic devices which correlate with this segmental structure are pauses, cue words, and referential noun phrases. First, long pauses consisting of silence or of non-lexical articulations (e.g., throat clearing) occur before the first proposition of all 3 segments. Second, assuming that *a-nd* in 9.1 is a cue word signalling a relation of 9.1 to the discourse, (rather than part of the proposition conveyed by *he goes up the ladder*), then 2 segments (1 and 3) begin with the cue words *okay* and *a-nd*. Finally, the discourse anaphoric pronoun *he* in 9.1 is used to refer to an entity first introduced in segment 1 as *a farmer* and later referred to as *he* in every remaining utterance of the segment. Despite the subsequent introduction of another male entity in segment 2, *a man with a goat*, the reference of the boxed pronoun in segment 3 is not ambiguous between the two male entities. This in part reflects differences in the semantic content of segments 1 and 2, and the close semantic relation of 9.1 to segment 1. It illustrates a phenomenon referred to as a discourse pop ([12] [40] [13] [9]) in which a suspended discourse goal, e.g., describing the farmer, is later resumed. The indentation in Fig. 1 thus reflects that segment 3 resumes the communicative goal suspended after segment 1, whereas segment 2 addresses a distinct communicative goal. Our investigations are directed at determining

[1] We use the term utterance to mean a use of a sentence or other linguistic unit, whether in text or spoken language.

SEGMENT 1
1.1 Okay.
2.1 [4.4 [.8 *sniff* [2.5] *throat clearing* [1.55] tsk [.35] There's a farmer,
2.2 [.15] he looks like ay uh Chicano American,
2.3 [.5] he is picking pears.
3.1 [4.85 [2.5] A-nd u-m [3.35]] he's just picking them,
3.2 he comes off of the ladder,
3.3 [3.5+ [.35] a-nd] he- u-h [.3] puts his pears into the basket.

SEGMENT 2
4.1 [1.0 [.5] U-h] a number of people are going by,
4.2 [.35+ and [.35]] one is [1.15 um] /you know/ I don't know,
4.3 I can't remember the first . . . the first person that goes by.
5.1 [.3] Oh.
6.1 A u-m a man with a goat [.2] comes by.
7.1 [.25] It see it seems to be a busy place.
8.1 [.1] You know,
8.2 fairly busy,
8.3 it's out in the country,
8.4 [.4] maybe in u-m [.8] u-h the valley or something.

SEGMENT 3
9.1 [2.95 [.9] A-nd um [.25] *throat clearing* [.35]] he goes up the ladder,
9.2 and picks some more pears.
. . .

Figure 1: Discourse Segment Structure: A Typical 'Pop'

what regularities there are in the interaction of segmentation, coherence and the three types of linguistic devices mentioned here: pauses, cue words and referential noun phrases.

While there is strong consensus on the need to model one or more aspects of the interaction among segmentation, coherence and linguistic devices, there is either weak consensus or only preliminary research on many fundamental issues. In section §2, we present an overview of various models of the three dimensions of discourse we address, with a particular emphasis on attempts at empirical verification. Then we present interim results of our ongoing empirical study of segmentation of narrative discourses from Chafe's pear story corpus [6]. We address the following questions:

- are discourse segments objective units that correspond directly to more abstract semantic or pragmatic units?
- how can theories of discourse segmentation be empirically verified?
- what is the nature of the boundaries between segments; e.g., do they have precise locations?
- what linguistic devices correlate with segment boundaries and to what degree?

This paper presents our take on the questions itemized above. Because the questions are interrelated, each section of the paper contributes gradually to our evolving view. In section 3, we present our empirical methodology and the results of an analysis of segmentation data we collected from a population of naive subjects. The segmentation criterion was based on a view of communication as a type of action (cf. Austin [1]), hence related to intentionality in the sense of Grosz and Sidner [13]. Our results, consisting of a quantitative evaluation of the degree of reliability among subjects, demonstrate an extremely significant pattern of agreement on segment boundaries. However, we also point to a large amount of variability in the data, including imprecision in the location of certain boundaries, and differences in what we refer to as boundary strength. In section 4, we quantitatively evaluate the performance of three algorithms based on noun phrases used to refer, cue phrases, and pauses. Our ultimate goal is to determine whether the correlations are strong enough to be useful in NLP systems, a question we return to in the concluding section. Our results suggest that it would be possible to algorithmically achieve human levels of performance given multiple knowledge sources. However, due to variability within and across speakers, we believe the most effective algorithm would need to dynamically adjust to different cues, possibly dependent on user modelling. In section 5 we relate our empirical results to coherence. First, we present a qualitative assessment of how well subjects agree on their abstractions of what segments are about. Then, we look in detail at an episode that occurs in all 10 narratives in our corpus, but which varies in size from 1 to 12 segments. This gives us one context in which to examine the relationship among segments, segment boundaries, and supra-segmental coherence. One of the algorithms is re-evaluated on this episode. Finally, we examine discourse pops, an example of another type of supra-segmental coherence. In the conclusion we discuss implications for natural language understanding and generation systems.

2 RELATED WORK

We do not have space here to provide a complete review of existing claims regarding the interdependence of segmentation, coherence, and linguistic devices. Instead, we mention certain works in order to illustrate these points. First, many have argued that understanding or generating coherent discourse depends on an appropriate model of the relation among segmentation, coherence and linguistic devices. Second, few explicit claims are made regarding segmentation, but it implicitly plays a role in much work on discourse comprehension and generation. Third, semantic and pragmatic coherence among utterances subsumes both rhetorical and intentional relations.

Claims regarding segmental structure are most explicit in the work of Grosz and Sidner [13], who propose a tri-partite discourse structure of three homomorphic levels: linguistic structure, which is roughly equivalent to what we refer to as segmental structure (but including linguistic devices), intentional structure, reflecting the goals of the discourse participants, and attentional state. They [13] make the strong claim that these three levels are isomorphic. Relations among segments, and among focus spaces, are analogous to hierarchical relations among discourse segment purposes (DSP), or the intentions underlying individual segments. The significant relations among segments/DSPs/focus spaces pertain to whether whether one DSP *dominates* or *precedes* another, not to the specific types of relations among DSPs.

In much other work on discourse, segmental structure is an artifact of coherence relations among utterances, and few if any specific claims are made regarding segmental structure per se. We use coherence here in a general way to refer to semantic, lexical, referential, intentional, rhetorical, or other relations among utterances. For Hobbs [20], understanding a discourse depends on understanding a variety of semantic and intentional coherence relations among utterances. Over time, he has made use of about half a dozen such relations, such as elaboration, exaplanation, occasion, and so on. He accounts for many other aspects of discourse coherence, such as assigning referents to noun phrases, as a by-product of understanding these inter-utterance relations. Polyani [38] similarly lays out a specific set of relations among utterances, but comes closer to [13] in proposing a mapping from semantic and pragmatic relations among utterances to a tree structure of discourse constituent units (DCUs). As in Reichman [40] and Fox [9], Polanyi [38] draws on work in interactional analysis [42] to distinguish between *open* or *closed* DCUs, and to predict the status of DCUs relative to the discourse parse tree. *Open* dcu's are located only on the right frontier of the tree, corresponding to the current focus stack of [13] (cf. discussion in Webber [47]). Also, as in [13], accessibility of discourse entities is predicted directly by the structure of the discourse tree.

Another tradition of defining relations among utterances arises out of Rhetorical Structure Theory [29], an approach which informs much work in generation (cf. discussion in [30]). Moore and Paris [30] argue that a generation system must have access to an explicit representation of intentions as well as rhetorical relations in order to handle phenomena such as clarifications of previous explanations. In related work, Moore and Pollack [31] note problems that arise out of an assumption that a single RST relation must represent the relation among consecutive discourse elements. They attempt to disentangle the intentional from the informational relations in RST. While others (cf. [13] [11] [29]) have pointed out that the relation of an utterance to the global discourse model may be ambiguous, Moore and Pollack [31] argue that the relation between intentional and informational structure is in some cases not isomorphic. They present an example of an utterance that is the nucleus of a set of utterances at the intentional level (e.g., a dominating DSP in terminology of [13]), but is not the nucleus when the same set of utterances is viewed on the informational level.

Although Moore and Pollack [31] do not directly address the question of segmentation, their claim that an utterance can simultaneously be a nucleus (dominating) and a satellite (subordinate) raises a number of questions about the adequacy of trees or stacks as the appropriate means of representating global structure in discourse, and of the ability to predict constraints between this structure and linguistic devices. Although all of the studies cited above have carried out detailed analyses of individual discourses or representative corpora to arrive at more or less substantive hypotheses, we believe that there is a need for many more rigorous empirical analyses of discourse corpora of various modalities (e.g., spoken, written), genres (e.g., task oriented, narrative), registers (e.g., conversational, formal), and so on. In the remainder of this section, we discuss recent empirical work on the relation among segmentation, coherence and linguistic devices.

Recently, more rigorous empirical work has been directed at establishing objectively verifiable segment boundaries. Despite the observation that segmentation is rather subjective [28] [22], several researchers have begun to investigate the ability of humans to agree with one another on segmentation, and to propose methodologies for quantifying their findings. Grosz and Hirschberg [11] [18] asked subjects to structure three AP news stories (averaging 450 words in length) according to the model of Grosz and Sidner [13]. Subjects identified hierarchical structures of discourse segments, as well as local structural features, using text alone as well as text and professionally recorded speech. Agreement ranged from 74%-95%, depending upon discourse feature. Hearst [15] asked subjects to place boundaries between paragraphs of three expository texts (length 77 to 160 sentences), to indicate topic changes. She found agreement greater than 80%. In [16], Hearst asked subjects to segment 13 texts of 1800-2500 words, but doesn't report agreement levels. We present results of an empirical study of a large corpus of spontaneous oral narratives, with a large number of potential boundaries per narrative. Subjects were asked to segment transcripts using an informal notion of speaker intention. We found agreement ranging from 82%-92%, with very high levels of statistical significance (from $p = .114 \times 10^{-6}$ to $p \leq .6 \times 10^{-9}$).

One of the goals of such empirical work is to use the results to correlate linguistic cues with discourse structure. By asking subjects to segment discourse using a non-linguistic criterion, the correlation of linguistic devices with independently derived segments can be investigated. Grosz and Hirschberg [11] [18] derived a discourse structure for each text in their study, by incorporating the structural features agreed upon by all of their subjects. They then used statistical measures to characterize these discourse structures in terms of acoustic-prosodic features. Morris and Hirst [32] structured a set of magazine texts using the theory of Grosz and Sidner [13]. They developed a lexical cohesion algorithm that used the information in a thesaurus to segment text, then qualitatively compared their segmentations with the results. Hearst [15] derived a discourse structure for each text in her study, by incorporating the boundaries agreed upon by the majority of her subjects. Hearst developed a lexical algorithm based on information retrieval measurements to segment text, then qualitatively compared the results with the structures derived from her subjects, as well as with those produced by Morris and Hirst. In more recent work, Hearst [16] presents two implemented segmentation algorithms based on term repetition. The boundaries produced by the algorithms are compared to the boundaries marked by at least 3 of 7 subjects, using information retrieval metrics. Iwanska [21] compares her segmentations of factual reports with segmentations produced using syntactic, semantic, and pragmatic information. We derive segmentations from our empirical data based on the statistical significance of the agreement among subjects, or *boundary strength.* We develop three segmentation algorithms, based on results in the discourse literature. We use measures from information retrieval to quantify and evaluate the correlation between the segmentations produced by our algorithms and those derived from our subjects.

3 INTENTION-BASED SEGMENTATION

3.1 DESIGN OF STUDY

If segments are taken to be a definitional construct, i.e., if by definition certain lexical items or other constructions are presumed to indicate the onset or closure of a segment, then the theorist should be at pains to motivate such definitions, e.g., through functional analysis of discourse corpora. But if segments are taken to be an independent construct that directly correlate with the focus of attention (cf. [13]) or the speaker's specific rhetorical purposes (cf. [28]), the definition of segment must be formulated independently of the distribution of linguistic devices. Only then does the hypothesis that the distribution of certain linguistic devices correlate with discourse segments have substantive content. We take this second tack, using an empirically derived database of segments.

First, a criterion for discourse segmentation must be elucidated. Unfortunately, the correspondence between discourse segments and more abstract units of meaning is poorly understood (cf. [31]). A number of alternative proposals have been presented which directly or indirectly relate segments to intentions [13], RST relations [28] or other semantic relations [38] [20]. In our study, we ask subjects to segment discourse using communicative intention as the segmentation criterion, as in [13]. Most importantly, by asking subjects to segment discourse using a non-linguistic criterion, the correlation of linguistic devices with *independently* derived segments can be investigated.

A subject pool must then be selected. Discourse segmentation has typically been performed by researchers, rather than obtained experimentally by empirical methods. We use "naive" subjects, as opposed to trained researchers in the area of computational discourse. In particular, the subjects are introductory psychology students at the University of Connecticut and volunteers solicited both from electronic bulletin boards at Columbia University and by the authors. Our decision to use naive subjects provides us with a large pool of subjects, and also helps ensure that subjects are not performing segmentation using the linguistic theories that we hope to independently validate.

The type of segmentation data to be collected must be determined. To simplify data collection,

...
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 ... 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.
...
In the wide left hand margin, say in abbreviated form what the speaker is doing. ... [Here] is an example of how to proceed. 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.
	The first part of the movie just sets up ...

Figure 2: Excerpt from Instructions

we asked subjects to perform a linear segmentation, as described below. We did not ask subjects to identify the type of hierarchical relations among segments illustrated in Fig. 1. In a pilot study we conducted, subjects found it difficult and time-consuming to identify non-sequential relations. Given that the average length of our narratives is 700 words, this is consistent with previous findings [41] that non-linear segmentation is impractical for naive subjects in discourses longer than 200 words. In addition, issues of intersegment relatedness (e.g., whether there is a standard set of relations that exist between segments) are still poorly understood.

A linear segmentation consists of dividing a transcript of a narrative into sequential segments. Our corpus consists of 20 narrative monologues about the same movie, taken from Chafe [6] (N≈14,000 words). Each narrative was segmented by 7 subjects.[2] Subjects were instructed to identify each point in a narrative where the speaker had completed one communicative task, and begun a new one. These points were thus the boundaries of segments. Subjects were also instructed to briefly identify the speaker's intention associated with each segment. Intention was explained in common sense terms and by example. An excerpt from the instructions is shown in Fig. 2.

Finally, to simplify data analysis, we restricted subjects to placing boundaries between the prosodic phrases identified in [6]. In principle, this makes it more likely that subjects will agree on a given boundary than if subjects were completely unrestricted. However, previous studies have shown that the basic unit subjects use in similar tasks corresponds roughly to a breath group, prosodic phrase, clause, or intonation unit [6] [41] [18] [45]. Thus one effect of using smaller units would have been to artificially lower the probability that two subjects would agree on the exact same boundary site, and conversely, to artificially inflate the significance when subjects agree.

Fig. 3 shows the boundary responses of subjects at each potential boundary site for a portion of the excerpt from Fig. 1.[3] Note that the identify of each of the 7 subjects is indicated by using a distinct letter of the alphabet. (The intentions associated with each subject's segmentations will be illustrated in later figures).Line spaces *between* prosodic phrases represent potential boundary sites. Note that a majority of subjects agreed on only 2 of the 11 possible boundary sites: after 3.3 (n=6) and after 8.4 (n=7). (The symbols NP, CUE and PAUSE will be explained later.) Agreement

[2] Except in rare cases, no subject segmented more than one narrative.

[3] The transcripts presented to subjects did not contain line numbering or pause information (pauses indicated here by bracketed numbers.)

3.3	[.35+ [.35] a-nd] he- u-h [.3] puts his pears into the basket.	
	$\boxed{\text{6 SUBJECTS (a, b, c, d, e, g)}}$	NP, PAUSE
4.1	[1.0 [.5] U-h] a number of people are going by,	
		CUE, PAUSE
4.2	[.35+ [.35] and] one is [.s] um /you know/ I don't know,	
4.3	I can't remember the first ... the first person that goes by.	
	$\boxed{\text{1 SUBJECTS (d)}}$	PAUSE
5.1	[.3] Oh.	
	$\boxed{\text{1 SUBJECTS (e)}}$	
6.1	A u-m.. a man with a goat [.2] comes by.	
	$\boxed{\text{2 SUBJECTS (d, e)}}$	PAUSE
7.1	[.25] It see it seems to be a busy place.	
		PAUSE
8.1	[.1] You know,	
8.2	fairly busy,	
	$\boxed{\text{1 SUBJECTS (c)}}$	
8.3	it's out in the country,	
		PAUSE
8.4	[.4] maybe in u-m [.8] u-h the valley or something.	
	$\boxed{\text{7 SUBJECTS (a, b, c, d, e, f, g)}}$	NP, CUE, PAUSE
9.1	[2.95 [.9] A-nd um [.25] [.35]] he goes up the ladder,	

Figure 3: Excerpt from 9, with Boundaries

among subjects was far from perfect, as shown by the presence here of 4 boundary sites identified by only 1 or 2 subjects. However, as we describe in subsequent sections, the probabilities of the observed patterns of agreement are extremely low, thus extremely statistically significant.

3.2 RELIABILITY ISSUES

Once data is collected from an empirical study, how can and should it be evaluated? Many data evaluation techniques and methodologies already exist outside the area of computational discourse. We draw on this body of knowledge to provide us with both a vocabulary for defining the issues, as well as with a set of analytic techniques.

We evaluate the segmentation data using *reliability*, as opposed to a weaker notion of *agreement*, and a stronger notion of *validity*. Krippendorff's [25] introduction to content analysis nicely illustrates the meanings of and differences among these terms:

> To test *reliability*, some duplication of efforts is essential. A reliable procedure should yield the same results from the same set of phenomena regardless of the circumstances of application. To test *validity*, on the other hand, the results of a procedure must match with what is known to be "true" or assumed to be already valid. It follows that, whereas reliability assures that the analytical results represent something real, validity assures that the analytical results represent what they claim to represent. ... *Reliability data* require that at least two coders independently describe a possibly large set of recording units in terms of a common data language. ...If agreement among coders is perfect for all units, then reliability is assured. If the agreement among coders is not better than chance, ...reliability is absent.

In other words, an agreement figure in isolation is often deceptive, since it does not indicate how the figure compares with the agreement figure that would occur by chance. We report the level of agreement in our data, but ultimately use a reliability measure based on comparing the observed distribution relative to a chance model. The validity of the subjects' segmentations is currently beyond our grasp, as the "true" segmentation of a discourse is not independently known. We later refer to the statistically significant boundaries as empirically validated, but by this we mean only that we use them as a target against which to validate how closely algorithm performance can approximate human performance.

3.3 STATISTICAL RESULTS

Since it is less informative than a reliability measure, we omit detailed discussion of percent agreement among our human subjects (but cf. [36]). However, previous work has used percent agreement as a summary statistic to report consistency across subjects performing various types of discourse segmentation (cf. section 2). For purposes of comparison, we give a brief discussion of our percent agreement results here.

As defined in [10], *percent agreement* is the ratio of observed agreements with the majority opinion to possible agreements with the majority opinion. In our empirical study, agreement among from 4 to 7 subjects on whether or not there is a segment boundary between two adjacent prosodic phrases constitutes a *majority opinion*. Given a transcript of length n prosodic phrases, there are n-1 possible boundaries. Thus the total *possible agreements* with the majority corresponds to the number of subjects times n-1. Total *observed agreements* equals the number of times that subjects' boundary decisions agree with the majority opinion.

Table 1 presents the data from Fig. 3 as a matrix where the 11 columns represent possible boundary slots and the 7 rows represent the subjects' responses. Column totals of 3 or less represent a majority opinion that the slot is a non-boundary; conversely, totals of 4 or more represent a majority opinion that the slot is a boundary. There are $11 \times 7 = 77$ possible agreements with the majority opinion, and 71 observed agreements. Thus, percent agreement for the excerpt as a whole is 71/77, or 92%. Percent agreement over our corpus averaged 89% (σ=.0006; max.=92%; min.=82%). The low standard deviation, or spread around the average, also shows that subjects are consistent with one another.

Non-boundaries, with an average percent agreement of 91% (σ=.0006; max.=95%; min.=84%), showed greater agreement among subjects than boundaries, where average percent agreement was 73% (σ= .003; max.=80%; min.=60%). This partly reflects the fact that non-boundaries greatly outnumber boundaries, an average of 89 versus 11 majority opinions per transcript. Because the average narrative length is 100 prosodic phrases and boundaries are relatively infrequent (average boundary frequency=15%), percent agreement among 7 subjects is largely determined by percent agreement on non-boundaries. Thus, total percent agreement could be very high, even if subjects did not agree on any boundaries. But percent agreement on boundaries alone is also high. Furthermore, as we show next, the reliability is extremely high.

There are various ways of arriving at a reliability figure. For example, a reliability statistic can quantify how close a set of observations is to a predicted distribution, or how far it is from a random one. Krippendorff [25] presents an agreement coefficient that indicates the extent to which an observed distribution such as that of Table 1 resembles a maximum rather than a chance agreement distribution (where chance agreement is estimated from the coders' distributions). His method also allows subjects to make more than binary distinctions in their decisions. In addition, there are various ways to model randomness of a data set, depending on the dimensions that are

Subject	Potential Boundary Slots										
	3.3,4.1	4.1,4.2	4.2,4.3	4.3,5.1	5.1,6.1	6.1,7.1	7.1,8.1	8.1,8.2	8.2,8.3	8.3,8.4	8.4,9.1
a	1										1
b	1										1
c	1								1		1
d	1			1		1					1
e	1				1	1					1
f											1
g	1										1
Total	6	0	0	1	1	2	0	0	1	0	7

Table 1: Matrix Representation of Boundary Data, Excerpt from 9

allowed to vary freely. We use Cochran's $\mathcal{Q}$ [7] to evaluate the reliability of our segmentation data.[4] Cochran's $\mathcal{Q}$ is a summary statistic that compares an observed matrix such as Table 1 to a random model given by allowing the locations of a subject's boundaries to vary, but restricting the total number of boundaries assigned by any subject to be equal to the observed total. In other words, column totals can vary randomly, but row totals are fixed.

We represented our segmentation data as a set of $i \times j$ matrices, such as Table 1 with height i=7 subjects and width j=n-1 prosodic phrases. Thus a row total corresponds to the total number of boundaries assigned by subject i. In Table 1 (j=11), subject d assigned 4 boundaries. The probability of a 1 in any of the j cells of the row is thus 4/11, with $\binom{11}{4}$ ways for the 4 boundaries to be distributed. Taking this into account for each row, Cochran's test evaluates the null hypothesis that the number of 1s in a column, here the total number of subjects assigning a boundary at the jth site, is randomly distributed. Where the row totals are u_i, the column totals are T_j, and the average column total is $\overline{T}$, the statistic is given by:

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

$\mathcal{Q}$ approximates the χ^2 distribution with j-1 degrees of freedom [7]. Our results indicate that the agreement among subjects is extremely highly significant. That is, the number of 0s or 1s in certain columns is much greater than would be expected by chance. For the 20 narratives, the probabilities of the observed distributions range from $p = .114 \times 10^{-6}$ to $p \leq .6 \times 10^{-9}$.

The potential boundary sites can be sorted into two classes, boundaries versus non-boundaries, depending on how the majority of subjects responded. Partitioning $\mathcal{Q}$ by the 7 values of j shows that $\mathcal{Q}_j$ is significant for each distinct $T_j \geq 4$ across all narratives (see [36]). For T_j=4, $.0002 \leq p \leq .30 \times 10^{-8}$; probabilities become more significant for higher levels of T_j, and the converse. At T_j=3, p is sometimes above our significance level of .01, depending on the narrative. Thus we use T_j=4 as the threshhold for distinguishing empirically validated boundaries from non-boundaries.

3.4 FUZZY BOUNDARIES

Despite the statistically significant agreement among subjects in identifying common segment boundaries, there is a lot of variability in the number and precise locations of segments proposed by each subject. Fig. 4 illustrates a case in point. It shows 7 prosodic phrases from narrative 12. Seven subjects agreed on boundary A, located at phrase (15.1,16.1). Two subjects located boundary B at (16.4,17.1), and three different subjects located it at (17.1,18.1). Thus a majority agree that boundary B occurs somewhere in the same vicinity, although no location meets the threshhold (T_j

[4] We thank Julia Hirschberg for suggesting this test.

15.1 [3.1 [7.5] A-nd [.45] then [1.15 tsk..]] he [.45] the goat [.25] and the goatman.. disappeared.

A 7 subjects (a, b, c, d, e, f, g)

16.1 [.55] A young boy on a bicycle,

16.2 [.45] that was much too big for him,

16.3 [1.35] rode [.45] thee.. from the [.2] direction in which the goat [.25] person had come,

16.4 [.8] towards the man picking the [.75] the pearpicker [3.0 [1.45.. tsk..]] a-nd [1.25]] stops.

B 2 subjects (a, c)

17.1 [1.3] Beside the baskets.

B 3 subjects (d, e, g)

18.1 [.5] While the man was upsta he had gone up the tree.

Figure 4: Indeterminacy of Boundary Location

≥ 4) established in the preceding section. Furthermore, as discussed in section 5, the 4 or 5 phrases between A and B describe a coherent sub-event of a well-defined episode of the Pear narratives. This constitutes conceptual motivation for classifying them as a coherent segment despite the imprecision in the location of boundary B. We are still developing a reliable method for integrating fuzzy boundary locations into our analysis (cf. §3.5). For present purposes, however, we quantify correlation of linguistic cues using data like boundary A, where a majority of subjects agreed on the precise location. In the next subsection, we describe the information retrieval metrics we use to evaluate the performance of an individual subject or algorithm against the consensus. Then in the following subsection, we present data quantifying the range of variation among subjects. This provides a baseline for interpreting the results for the algorithms.

3.5 USE OF INFORMATION RETRIEVAL METRICS

Information retrieval (IR) metrics are designed to quantify, e.g., what proportion of documents retrieved by a given query are *correct*, what proportion of *desired* documents are retrieved, and so on. We use IR metrics to quantify what proportion of statistically validated boundaries are recognized, and so on. We make two caveats about the use of IR metrics. First, IR metrics are overly conservative, given the binary classification of ideal and actual performance into recognition of boundaries versus non-boundaries. As noted above, such a strict classification fails to address near misses like boundary B in Fig. 4. In future work we will remedy this by using sums of boundary weights within a moving window of 3 or more boundary slots, or by using string matching techniques. Second, interpreting the values of the IR metrics requires some notion of how hard the task is. In the next subsection, we establish a baseline for comparison by quantifying the performance of human subjects using the IR metrics.

Fig. 5 schematicatically presents the distributions used to compute four IR metrics. The column marginals, or sums, (i.e., a+c and b+d) represent the classification of possible boundary locations (pairs of sequential prosodic phrases) into empirically validated boundaries and non-boundaries. The row marginals (i.e., a+b and c+d) represent the hypotheses of an individual subject or algorithm for a given narrative regarding boundaries and non-boundaries. Each cell represents an intersection of an ideal classification with a hypothesized one, thus 'a' represents the set of correctly hypothesized boundaries, 'b' represents the set of incorrectly hypothesized boundaries, and so on. The formulas for the 4 IR metrics we use to summarize such a table are also given in Fig. 5. Ideally, the number of incorrect hypotheses for boundaries (c) and non-boundaries (b) both equal 0, giving a value of 1 for both recall and precision, and a value of 0 for both fallout and error rate. We will see in the next section that average human performance is far from ideal.

Test Responses	Empirically Validated Responses	
	Boundary	Non-Boundary
Boundary	a	b
Non-Boundary	c	d

Recall = a/(a+c)
the ratio of correctly hypothesized boundaries to empirically derived boundaries

Fallout = b/(b+d)
the ratio of incorrectly hypothesized boundaries to empirically derived non-boundaries

Precision = a/(a+b)
the ratio of correctly hypothesized boundaries to hypothesized boundaries

Error rate = (b+c)/(a+b+c+d)
the ratio of incorrect hypotheses to the table total

Figure 5: Information Retrieval Metrics

3.6 BASELINE HUMAN PERFORMANCE

Ultimately, we quantify the correlation of linguistic devices with segment boundaries using IR metrics. However, interpreting the results depends on first understanding the inherent variability in the human subjects' data.

Given a particular narrative, we take boundaries agreed on by a majority of subjects to be empirically validated. Since the 7 subjects providing data on any one narrative all assign boundaries at a different rate, the performance of any one subject is likely to deviate from the consensus. Table 2 gives an overall picture of the human subjects' performance. Row 1 repeats ideal scores for the 4 metrics (cf. Fig. 5), for convenient reference. The average performance across all subjects on all narratives, given in row 2, shows in particular that average precision is relatively poor, with roughly only half the hypothesized boundaries being 'correct'. This results from the wide variation in the rate at which subjects assign boundaries. The average frequency of boundary assignment is 15%, but the standard deviation (σ=.07) reflects the large spread around the average (max=.41; min=.03). When a subjects' rate is higher than average, the value of 'b' (cf. Fig. 5) increases and precision decreases.

Table 2 also shows the best and worst values for each metric achieved by any subject (rows 4 and 6), and the best and worst overall performance for a given subject (rows 5 and 7). To rate overall performance for a particular subject, we compute a sum of deviations (column 6): the sum of differences between the observed and ideal values for each metric. In the worst possible performance, recall and precision equal 0 instead of 1, and fallout and error equal 1 instead of 0, thus the summed deviation would equal 4. Note that the values for the best metrics are close to ideal for recall (1), fallout (.01) and error (.04). The subject with the best overall score also was the single subject with the best precision (.86), and notably achieved both high precision and recall. This subject assigned 7 boundaries, 6 of which were validated boundaries, in a relatively short narrative that had a total of 7 validated boundaries. Note that although this subject had only 1 type 'b' error and one type 'c' error, recall and precision are not higher because there were relatively few boundaries.

A single subject happened to have both the worst fallout (.35) and the worst error rate (.33), but had a relatively high recall (.81) and moderately bad precision (.26; i.e., roughly midway between the worst of .09 and the average of .55). This reflects this subject's very high rate of assigning boundaries (41.0% compared with the average of 22.7%), leading to higher than average errors of

	Recall	Precision	Fallout	Error	Deviation
Ideal	1.00	1.00	.00	.00	.00
Average	.74	.55	.09	.11	.90
σ	.20	.17	.06	.05	.34
Best Metric	1.00	.86	.01	.04	.19
Best Subject (narr 3)	.86	.86	.02	.04	.34
Worst Metric	.20	.09	.35	.33	2.39
Worst Subject (narr 6)	. 20	.09	.13	.17	2.01

Table 2: Individual Human Performance

both types 'b' and 'c'. The subject with the worst recall (.2) and precision (.09) had only moderately bad fallout (.13) and error rate (.17) because, having a boundary rate equal to the average, this subject's errors of type 'b' and 'c' were still low relative to 'd' despite being high relative to 'a'. Thus interpreting Table 2 requires an understanding of the interdependencies among the IR metrics and the typically low rate of assigning boundaries.

4 CORRELATION OF BOUNDARIES AND LINGUISTIC CUES

In the preceding section, we demonstrated that the reliability of humans' ability to recognize segment boundaries is highly significant, despite the open-endedness of the task presented to the subjects, despite what is mostly likely an overly rigid notion of *segment boundary location*, and despite the variability of human performance. Our next goal is to determine whether segments reflect processes or structures that would be useful for generating or understanding extended discourse, regardless of their role in on-line human processing. We tackle this goal in two ways. First, in this section, we quantify the degree to which linguistic features correlate with segment boundaries, thus providing automatic means for recognizing or generating segment boundary location.

4.1 THREE ALGORITHMS

In principle, the process of determining whether the empirically validated segment boundaries correlate with linguistic devices would require a complex search through a large space of possibilities, depending on what set of linguistic devices one examines, and what features are used to recognize and classify them. Our interest is primarily in providing an initial evaluation of existing hypotheses, rather than in developing a method to search blindly through the space of possibilities. Thus we have begun by looking at three devices whose distribution or surface form has frequently been hypothesized to be conditioned by segmental structure: referential NPs, cue words and pauses.

REFERENTIAL NOUN PHRASES. Distributional analyses of discourse anaphoric NPs in various kinds of monologic, conversational or textual discourse has generated specific hypotheses about how surface form correlates directly or indirectly with segmental structure (cf. e.g., [12] [44] [43], [13] [40] [26] [34] [35]). We use a segmentation algorithm based on features of the surface form of referential NPs developed out of our previous work on formulating specific discourse constraints. For further details on the motivation for the algorithm design, cf. [36] [33].

The input to the NP algorithm consists of lists of 4-tuples representing all the referential NPs in each narrative, where each tuple is <LOC, NP, INDEX, INFER>. LOC stands for location, and is represented in terms of the sequential position of the prosodic phrases and syntactic clauses in which the NP occurs.[5] NP stands for the surface form of the noun phrase, permitting identification of definite NPs and their classification into phrasal versus pronominal NPs. INDEX stands for the discourse entity evoked by each NP (cf. Karttunen [24], Webber [46], Heim [17] and Kamp [23]). If an NP evokes evokes an entity that is already in the discourse model, it is assigned the same index. Finally, INFER is a set of inferential relations between the denotation[6] of the NP and other objects in the discourse context. It has long been noticed that discourse anaphoric NPs are also licensed given a strong inferential link to an existing entity (cf. [12] [39] [4] [14] [20]). The output is a set of boundaries B, represented as ordered pairs of adjacent clauses: (C_n,C_{n+1}). This output is then mapped to pairs of adjacent prosodic phrases.[7]

Essentially, the NP algorithm has three tests for determining whether the next pair of adjacent clauses (C_i, C_{i+1}) represents a boundary. Any NP can provide a direct link to the preceding clause through coreference or an inferential relation; a definite pronoun can provide a link to the current segment as a whole. Specifically, a clause pair (C_i, C_{i+1}) is not a boundary under one of the following conditions:

- coreference: the same INDEX is used for an N-tuple in C_{i+1} and an N-tuple in C_i
- an inferential link: an INFER relation for an NP-tuple in C_{i+1} links an INDEX in C_{i+1} to an INDEX in C_i
- definite pronominal reference to an entity in the current segment: a third person definite pronoun in C_{i+1} receives an INDEX that is also used for some NP-tuple from any prior clause, up to the last boundary that was assigned.

If all 3 tests fail, C_{i+1} is determined to begin a new segment, and (C_i,C_{i+1}) is added to the current set of boundaries.

The excerpt in Fig. 3 shows two boundaries assigned by the NP algorithm. Boundary (3.3,4.1) is assigned because the sole (non-pronominal) NP in 4.1, *a number of people*, evokes an entity that is entirely new to the discourse, and this entity cannot be inferred from any entity mentioned in 3.3. Thus all 3 tests fail. The next boundary that is assigned is (8.4,9.1). The entity evoked by the full NP *the ladder* is not evoked in 8.4, and it is not inferentially linked to any of the entities in 8.4, thus failing the first two tests. The third person pronoun *he* evokes an entity, the farmer, that was last mentioned in 3.3, and 1 NP boundary has been assigned since then, thus failing the third test.

CUE WORDS. Cue words (e.g., "now") are words that are sometimes used to explicitly signal the structure of a discourse. Our segmentation algorithm based on cue words uses a simplification of one of the features shown by Hirschberg and Litman [19] to identify discourse usages of cue words. Hirschberg and Litman [19] examine a large set of cue words proposed in the literature and show that certain prosodic and structural features, including a position of first in prosodic phrase, are highly correlated with the discourse uses of these words. The input to our cue word algorithm is a sequential list of the prosodic phrases constituting a given narrative. The output is a set of

[5]The syntactic clauses consist roughly of tensed clauses that are neither verb arguments nor restrictive relative clauses.

[6]In theory, relevant inferential relations may involve either the discourse entity evoked by an NP or the intended referent (cf. [35]), but for present purposes we conflate the entity and referent indices.

[7]Because potential boundary locations assigned by the NP algorithm (between adjacent clauses) can differ from the potential boundary locations for the human study (between adjacent prosodic phrases), the data is first normalized, as described in [36]. Normalization eliminates a total of 5 boundaries in 3 of the narratives (out of 213 in all 10).

	Recall	Precision	Fallout	Error	Deviation
Avg. Human	.74	.55	.09	.11	.79
Avg. NP	.50	.30	.15	.19	1.53
σ	.17	.10	.06	.06	.29
Best (narr 3)	.71	.50	.10	.13	1.02
Worst (narr 2)	.44	.20	.27	.31	1.94
Avg. Cue	.72	.15	.53	.50	2.16
σ	.17	.06	.07	.07	.27
Best (narr 3)	.86	.22	.44	.40	1.76
Worst (narr 6)	.40	.04	.60	.60	2.76
Avg. Pause	.92	.18	.54	.49	1.93
σ	.09	.05	.06	.06	.17
Best (narr 2)	.94	.23	.48	.42	1.73
Worst (narr 18)	.73	.16	.57	.53	2.21

Table 3: Evaluation for $T_j \geq 4$

boundaries B, represented as ordered pairs of adjacent phrases (P_n,P_{n+1}), such that the first item in P_{n+1} is a member of the set of cue words summarized in Hirschberg and Litman [19]. That is, if a cue word occurs at the beginning of a prosodic phrase, the usage is assumed to be discourse and thus the phrase is taken to be the beginning of a new segment. Fig. 3 shows 2 boundaries (CUE) assigned by the algorithm, both due to *and*. Because *and* occurs frequently in phrase initial position, the cue algorithm assigns boundaries at a high rate. As noted below, this results in low precision.

PAUSES. Grosz and Hirschberg [11] [18] found that in a corpus of recordings of AP news texts, phrases beginning discourse segments are correlated with duration of preceding pauses, while phrases ending discourse segments are correlated with subsequent pauses. We use a simplification of these results in our algorithm for identifying boundaries in our corpus using pauses. The input to our pause segmentation algorithm is a sequential list of all prosodic phrases constituting a given narrative, with pauses (and their durations) noted. The output is a set of boundaries B, represented as ordered pairs of adjacent phrases (P_n,P_{n+1}), such that there is a pause between P_n and P_{n+1}. Unlike Grosz and Hirschberg, we do not currently take phrase duration into account. In addition, since our segmentation task is not hierarchical, we do not note whether phrases begin, end, suspend, or resume segments. Fig. 3 shows 7 boundaries (PAUSE) assigned by the algorithm.

4.2 PERFORMANCE OF ALGORITHMS

In this section, we discuss the performance of the three algorithms first to evaluate their success at locating segment boundaries, but then to draw generalizations regarding the nature of segment boundaries. Unsurprisingly, the performance of the three algorithms as quantified by the IR metrics varies with the amount of knowledge they exploit. More enlightening than comparing the algorithms with one another is to compare them with human performance. We first present and discuss the performance of each algorithm averaged across the narratives as compared with the average of human subjects. The results suggest that levels approximating average human performance could eventually be achieved. However, the key lesson we draw from analyses of variation in the data is that no single strategy for identifying segments would be sufficient. Individual discourses vary significantly in how reliably humans or algorithms can segment them (variation across speakers). Also, there may be inherent limitations to algorithms that rely on a consistent strategy, because boundaries within the same discourse may be signalled by different means (variation within speakers).

Table 3 shows the performance of the three algorithms (NP, CUE, PAUSE) using the 4 IR metrics, along with the summed deviation. Again, the summed deviation ($0 \leq$ deviation ≤ 4) is the sum of the deviations of each metric from ideal performance. For convenient comparison, the first row in Table 3 repeats the human subjects' average performance. The remainder of Table 3 consists of 3 subtables, with 4 rows per algorithm. The first row of each subtable gives average values across the 10 narratives, and the second gives the standard deviation.[8] We use the summed deviations to identify the individual narratives on which each algorithm performs best and worst, shown in the 3rd and 4th rows of the subtables. NP exploits more knowledge than CUE and PAUSE,reacting to one of 3 conditions instead of a single one. This is reflected in the performance differences, with NP performing less well than humans but better than CUE and PAUSE. The average summed deviation for NP (1.53) is almost twice as bad as for humans (.79), but is roughly 75% that for PAUSE (1.93) and CUE (2.16). Recall and precision show the biggest difference between human performance and NP; they are also more sensitive to small changes in performance because values of 'a', 'b' and 'c' tend to be much smaller than values of 'd'. NP recall is .50 compared with .74 for humans; NP precision is .30 compared with .55 for humans. NP error and fallout are closer to human performance (.15 vs. .09; .19 vs. .11). PAUSE (.92) has much higher recall than humans, but PAUSE and CUE have very low precision (.15 and .18) and very high fallout and error rate (ranging from .49 to .54).

A second dimension to consider in observing that NP performance is closer to human performance than PAUSE and CUE is that the NP algorithm and human consensus result in relatively fewer boundaries. On average, subjects assigned boundaries 15% of the time (σ=.07). Given the average narrative length of 102 boundary slots, this makes each segment an average of 15 phrases long. NP assigns boundaries using narrow criteria reflecting an assumption that in the default case, each subsequent clause of a narrative continues the current segment rather than starting a new one. It assigns boundaries at a rate of 19%. In contrast, CUE and PAUSE rely on features that occur frequently, and consequently assign boundaries relatively often.

The difference in performance between the NP algorithm versus CUE and PAUSE suggests that by exploiting even more knowledge performance might be further improved. In fact, we are optimistic that all three algorithms can be improved by discriminating among duration of pauses (cf. [11]), types of cue words (e.g., signalling pops [13]), and by adding further features to differentiate uses of referential NPs (e.g., grammatical role information, cf. [34]).

As another illustration of how increased knowledge can enhance performance, Table 4 presents results for various composite algorithms that locate a boundary wherever two or more of NP, PAUSE and CUE locate a boundary. 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. As noted, recall of combined algorithms is limited by the lowest recall of the individual algorithms: the average recall of .49 for NP/(CUE or PAUSE) (row 5) is

[8] The figures for NP differ from those in [36], where we indavertently gave the metrics for boundaries defined by a threshold of agreement among at least 5 instead of 4 subjects. Another difference is that formerly the algorithm was applied by hand; here, the coding of NP-tuples is still only partly automated, but application of the algorithm is fully automatic.

	Recall	Precision	Fallout	Error	Deviation
NP/Pause	.47	.42	.08	.13	1.42
NP/Cue	.36	.34	.09	.15	1.59
Pause/Cue	.69	.29	.29	.29	1.66
NP/Cue/Pause	.34	.47	.05	.12	1.43
NP/(Cue or Pause)	.49	.34	.12	.17	1.52

Table 4: Combined Algorithms

only somewhat lower than the average recall of .50 for NP alone and the average recall of .69 for the PAUSE/CUE algorithm (row 3) is somewhat lower than that of .72 for CUE alone. However, the composite algorithms use narrower criteria for boundaries, which should reduce the number of false positives. The precision of the combined algorithms is indeed higher than any of the algorithms alone. NP/PAUSE has the best combined algorithm performance as measured by the summed deviation. We believe it would be possible to algorithmically achieve the performance levels of the average human subject by some combination of additional knowledge and judicial use of multiple knowledge sources. However, the differences across and within narratives indicate inherent limitations to this approach.

Standard deviations (σ) of the various metrics for human performance are included in Table 2 and for algorithm performance in Table 3. Overall, the standard deviations for recall and precision are higher than for fallout and error. This reflects the high proportion of non-boundaries in the data, i.e., values of 'd' are always relatively large, minimizing the effect of errors relative to 'd'. The largest spread occurs among humans: for recall, σ=.20 and for precision, σ=.17. There is wide difference among subjects, even on the same narrative. For example, for narrative 3, 2 subjects achieved maximum recall of 1 whereas recall of a third subject was equal to the observed minimum of .2. Differences among subjects on the same narrative presumably result in part from different levels of attention and ability on the part of the subjects. However, since subjects bring different goals and different kinds of knowledge to the segmentation task, the variation must also partly reflect different interpretations of the same discourse (cf. section 2). As we note in the conclusion, this is an issue that should be addressed in designing natural language understanding systems.

As noted earlier, the performance differences in the algorithms primarily reflects differences in the amount of knowledge they use. But performance differences across narratives may also reflect differences in the narrative strategies used by the speakers. Levy [26], for example, identifies several distinct speaker strategies for reidentifying major and minor characters based on an analysis of narrative monologues about the television dramatization of *Brideshead Revisited.* On the one hand, some speakers seem to produce narratives that because of greater consistency in speaker strategy, or other factors to be determined, are segmented more reliably by more than one algorithm. Narrative 3, for example, yields the best scores for NP and CUE as well as the best scoring human. On the other hand, some speakers may rely more exclusively on a strategy that affects one type of linguistic feature, making other surface features irrelevant. This might explain why narrative 2 shows the worst performance for NP but the best performance for PAUSE.

To summarize our interim conclusions regarding the correlation of linguistic cues with segment boundaries, our results show that it may be possible to replicate average human performance, but that doing so would require multiple sources of knowledge. For example, the NP algorithm looks at multiple features of NPs, and its best performance (recall=.71; precision=.50; fallout=.10; error rate=.13) is close to the average of humans performance (recall=.74; precision=.55; fallout=.09;

error rate=.11). Combining NP and PAUSE raises average precision by a third and causes small improvements in fallout and error rate without significantly degrading recall. However, a close look at different aspects of variation in the subjects' reliability at identifying boundaries, and of the spread in performance on individual narratives suggests that in addition to using multiple knowledge sources, automated methods for recognizing segments through surface features would ultimately need to adapt dynamically to different strategies within and across speakers.

5 FROM BOUNDARIES TO SEGMENTS

The previous section asked whether specific locations for segment boundaries can reliably be identified by humans or automated methods. In this section, we address the question of how the results on segment boundaries pertain to coherence. First, we illustrate the degree to which subjects agreed on the intentions they assigned to segments by examining one segment in detail. Then we examine two multi-segment phenomenon. In one case, we take a semantically coherent multi-segment unit consisting of an episode mentioned in all 10 narratives. We discuss the relation of segment boundaries to segments in this episode, focussing on the issue of fuzzy boundary location. We re-examine one of the algorithms on the episode, using fuzzy boundary location. Finally, we examine the more general phenomenon of discourse pops, illustrated in our first example in Fig. 1, where the relation among distinct segments is pragmatically defined. We discuss the nature of the segment boundaries in all the pops involving a definite pronoun in the resumption segment. We also discuss the semantic coherence of the segment that is popped over.

5.1 DISCOURSE SEGMENT INTENTION

Inspection of subjects' descriptions of speaker intention suggests that subjects not only agree on the boundaries of segments, but also agree on the narrator's intentions for the segment. As yet we have not found a suitable formal methodology for analyzing the reliability of the data regarding intentions. For example, the matrix for representing boundary decisions (recall Table 1) is not suitable for representing segment intentions. In particular, since each subject associates intentions with a different set of segments, it is not possible to have the same set of columns across subjects when specifying intentions. In addition, when deciding whether a potential site is a boundary or not, subjects are restricted to a dichotomous choice. In contrast, subjects are not restricted to choosing intentions from a pre-defined set of values. Here we informally examine the qualitative agreement among subjects on identification of speaker intention.

Consider the segment that can be derived from Fig. 6. The segmentation shown in Fig. 6 contains 2 boundaries proposed by 1 subject each, 1 boundary proposed by 5 subjects, and 1 proposed by all 7 subjects. The first 3 of these comprise a fuzzy boundary in the sense illustrated above with Fig. 4; 7 distinct subjects locate a boundary at or near (14.1,14.2). This yields a single discourse segmentation for the excerpt spanning from an initial fuzzy boundary at (12.5,13.1,14.1,14.2) to (15.4,16.1). However, the subject who located the initial boundary earlier than the 6 other subjects, 'g' at (12.5,13.1), associates a completely distinct intention with the segment.

Fig. 7 presents the intentions attributed to the narrator by the 7 subjects. The 5 subjects who agreed on exactly the same segment boundaries (a-e) all indicate that the speaker's intentions pertain to describing the audio characteristics of the movie. Subject f identifies the segment to include one more phrase (cf. Fig. 6), and characterized the segment purpose as *techniques used in the movie.* Thus f identifies essentially the same intentional structure. However, subject g, 12.5 [2.1 [.2] a-nd [1.2]] his [.3] bicycle hits a rock.

1 SUBJECT (g)

13.1 Because he_i's looking at the girl.

1 SUBJECT (f)

14.1 [.75] {ZERO-PRONOUN_i} Falls over,

5 SUBJECTS (a, b, c, d, e)

14.2 [1.5 [1.35] uh] there's no conversation in this movie.

15.1 [.6] There's sounds,

15.2 you know,

15.3 like the birds and stuff,

15.4 but there.. the humans beings in it don't say anything.

7 SUBJECTS (a, b, c, d, e, f, g)

16.1 [1.0] He_i falls over,

Figure 6: Excerpt from 6

Subject	Annotation of Narrator's Intention
a	Digression to describe sound track
b	No verbal communication [i.e., speaker describes lack thereof]
c	Describes that it is a silent movie with only nature sounds
d	Speaker describes sound techniques used in movie
e	Explain that there is no speaking in movie
f	Techniques use in the movie
g	Because he's looking at the girl when the boy falls no one else cares about him

Figure 7: Segment spanning 14.2 through 15.4

who began the segment with 13.1, offers an intention that has no relevance to the sound track or cinematic properties of the movie. Subject g thus not only identifies a different segment boundary, but also a different overall purpose.

Presumably, the first 6 subjects depicted in Fig. 7 abstract from the fact that the three full clauses in the segment all refer to auditory characteristics, as signaled by the lexical items *conversation* in the first clause, *sounds* in the second, and *say* in the third. Here we have generalized further from the subjects annotations to note that each asserts something about the movie's auditory character. In general, for segments delimited by high agreement boundaries, a single formulation of speaker purpose can be generalized from the data provided by the 7 subjects. We believe that such data provides evidence that when asked to, subjects perform the same kinds of abstraction across related utterances described in [38] and elsewhere.

5.2 THE THEFT EPISODE

One of the key events in the narrative is an episode in which a young boy, one of the main characters in the narrative, steals a basket of pears. We use this episode to illustrate the interaction between

1. boy arrives on bike
2. boy steals the one full basket of pears
3. boy rides away with basket balanced on bike handlebars
4. boy is distracted by an approaching girl
5. boy loses hat
6. boy's bike hits rock
7. boy falls over
8. pears spill all over the ground

Figure 8: Events in the Theft Episode

the sometimes imprecise location of segment boundaries, and segment content. We show that with a fuzzier definition of boundary, the NP algorithm performance improves significantly, particularly on precision. However, variation across narratives remains high, reinforcing our conclusion that maximum performance could not be achieved without relying on multiple types of features.

By abstracting across the descriptions of the theft episode in our 10 narratives in a manner similar to the method described by Polanyi [38], the 8 actions listed in Fig. 8 can be identified. In all but one narrative, the episode consists of a contiguous sequence of prosodic phrases. Earlier, we used Fig. 4 to illustrate a fuzzy boundary. Recall that five subjects agreed that a boundary occurred within 2 adjacent boundary slots: (16.4,17.1) or (17.1,18.1). If this were a boundary, it would separate a segment describing the first event listed in Fig. 8 from a segment describing event 2 of the episode. Here we define a fuzzy boundary to be a contiguous sequence of inter-phrasal locations PP_i to PP_n such that the total number of distinct subjects identifying PP_i to PP_n is at least 4. By this definition, boundary B in Fig. 4 is a fuzzy boundary beginning at (16.4,17.1) (2 subjects) and ending at the next adjacent location (17.1,18.1) (3 subjects). Using fuzzy boundaries adds 3 boundaries out of a total of 46 for all the segments where the theft episode is described.

Using fuzzy boundaries, the average number of segments spanning the sequence is 3.6. The range is from 1 to 11, so there is considerable variation in how many utterances or segments narrators used in relating the episode. One uniformity is that relatively more of the boundaries that are external to the theft episode, i.e., those that begin or end the whole episode, are identified by 6 or more subjects: 18 of the 22 external boundaries.[9] Conversely, 15 of the 24 episode internal boundaries are identified by at most 4 or 5 subjects. The average number of subjects identifying an external boundary is 6.5, compared with 5.3 for internal boundaries.[10] Thus subjects are more likely to identify a boundary marking the break between this episode and other material than one marking intra-episode breaks.

The NP algorithm is also more likely to find an external boundary than an internal one: it finds 17 of the 22 external boundaries but only 10 of the 24 internal ones. This provides a possible functional correlation for our suggestion in [36] that boundaries identified by more subjects are located more reliably by all 3 algorithms; i.e., location of more salient boundaries may correlate with more global discourse intentions.

[9] Narrative 1 has more external boundaries because of an interruption within the episode.

[10] Similar differences between external and internal boundaries exist using the strict notion of boundary.

Table 5 contrasts three performance figures for the NP algorithm: the overall performance figures from Table 3 (row 1), figures for the theft episode alone, but using the same strict notion of boundary location (row 2), and figures for the theft episode using the fuzzy definition of boundary location (row 3). For fuzzy boundaries, we changed the output of the algorithm as follows. Contiguous boundary slots identified by the algorithm were classified as a single boundary; thus whether the algorithm identified one or both locations for B in Fig. 4, it counts as one right answer and no wrong answers. This both penalizes and aids the algorithm: it degrades recall because it becomes impossible to find the occasional occurrences of two distinct boundaries in adjacent slots, but it improves precision.

Table 5 shows an improvement in recall (from .5 to .6) and precision (from .3 to .4), using the strict definition of boundary. We believe this is because in all cases but one narrative, the input consists of a continuous sequence of phrases that is by definition more coherent: it comprises a single episode. Within a whole narrative, there are occasional sequences of phrases whose function is more difficult to interpret, and which are therefore harder to segment. Table 5 also shows an event greater improvement when the input is both a single episode, and boundaries are fuzzy. All scores improve, with precision unsurprisingly showing the greatest improvement.[11]

Perfect scores were achieved on the one narrative (3) in which the whole episode is comprised by a single segment. This segment consisted of 14 phrases. As illustrated in Table 5, the worst performance was on narrative 17, which had 5 segments in the episode (with no difference between strict and fuzzy conditions). Lest one assume that the performance degrades as the number of segments increases, performance was very good for another narrative excerpt consisting of 5 segments: for narrative 12, recall was .80, precision was .67, fallout was .05 and error rate was .07 with a summed deviation of .65. In fact, the performance of the algorithm is far better on 8 of the narratives than on 2 and 17. For example, the average summed deviation of these 8 narratives is only .56, compared with .96 for 2 and 1.74 for 17.

Inititial experiments suggest that improvements to NP would be easier for the 8 narratives on which it already performs well, than on 2 or 17. One condition that seems to improve performance is to require NP to assign a boundary if a completely new entity is introduced that is not inferentially related to entities in the previous utterance. By adding this feature, recall in the 8 narratives would be nearly perfect.[12] However, the effect on narratives 2 and 17 would be quite different. This feature is associated with 4 of the 5 boundaries missed by the algorithm in narrative 17, but with only 1 of the 6 boundaries missed by the algorithm in narrative 2. Recall also that narrative 2 was the narrative yielding the worst overall score for NP (cf. §4.2), with a summed deviation of 1.94. Thus, the narrators in 2 and 17 not only use different patterns of expression from the remaining 8, as illustrated by differences in performance of NP on the theft episode, but also differ from one another.

In sum, NP performance improves when the input is semantically more coherent, and when boundaries are allowed to have fuzzy locations. However, large differences in performance from one narrative to another still exist.

[11] Hearst [16] also finds a big improvement when she admits fuzzy boundaries.

[12] Since the use of new features is still under development, we do not have actual scores on metrics for an NP algorithm using this feature.

	Recall	Precision	Fallout	Error	Deviation
Whole Narrative	.50	.30	.15	.19	1.53
Strict	.60	.40	.17	.20	1.37
Fuzzy	.69	.74	.04	.09	.70
Best Fuzzy (3)	1.00	1.00	.00	.00	.00
Worst Fuzzy (17)	.17	.33	.06	.18	1.74

Table 5: NP Performance on Theft Episode

5.3 DISCOURSE SEGMENT POPS

Although our subjects were not asked to specify structural or semantic relations among segments, here we briefly look at a specific class of the structural discourse relationship *pop*, which we can identify linguistically from our data. In the general case, discourse pops occur when an earlier suspended segment is resumed with a later segment. In this section we restrict our attention to pops where the first complete utterance of the segment contains a third person definite pronoun whose referent is not mentioned in the preceding segment, but rather in a prior segment. The pop in Fig. 1 is an example. Such pops often occur after a segment on which there is strong consensus of a semantic discontinuity. Since the resolution of the reference of the pronoun depends on recognizing a relation between the new segment and a previously suspended segment, we raise the question of how such pops are recognized, and whether the semantic relations are directly or indirectly encoded.

In Fig. 6, one of the main characters of the movie (a boy on a bicycle) is referred to in phrases 13.1 and 14.1. The speaker suspends her description of the boy's activities throughout the next segment ([14.2,15.4]), but resumes reference to the boy in the first utterance of the following segment (16.1) using a third person definite pronoun subject (*he*). This illustrates a specific class of discourse pops, in which a segment resumes an earlier *suspended segment*. In particular, this class of *resumption segments* begins with an utterance in which a third person definite pronoun refers to an entity that is not in the focus space [13] associated with the *intervening segment*. Processing a clause that signals this type of discourse pop involves several tasks. Resolving the pronoun in the initial clause of the resumption segment requires shifting the attentional state [13], since the active focus space (corresponding to the intervening segment) does not contain a representation of the referent. This shift depends on recognizing the termination of the intervening segment, and a continuation relation between the resumption and suspended segments, so that the entities in focus in the suspended segment are again in focus for the resumption segment.

There are 8 discourse pops of this type in the 10 narratives we tested the algorithms on. These 8 examples exhibit various structural and semantic relations to the presumed discourse model. For example, they contain intervening segments that provide more detail, provide general background, or are digressions. In 7 cases, the resumption segment begins with a word that can function as a cue word,[13] but in 4 cases the cue word is *and*, a word whose discourse usage is hard to distinguish and which provides very little semantic information. The cue words in the remaining 3 cases are *so, all right* and *then*, none of which clearly signal attentional change [13] [19]. Non-lexical signals (e.g., *uh, tsk*, false starts) precede the initial clause of the resumption segment in 5 cases, and relatively long pauses ($>$ 1.5 sec.) in 6 (cf. [18]). This suggests that in spontaneous oral discourse, instead

[13] We assume that different uses of cue phrases can be discriminated; cf. [19].

of giving explicit indicators of the structural and semantic relations among segments, speakers provide non-lexical and pausal cues to breaks in segmental structure, relying on the hearer to infer the abstract structural and semantic relations. For example, the semantics of the predication *fall over* at the onset of the resumption segment is arguably very dissimilar to the predications occurring in the intervening segment, possibly supporting the inference that the new clause is unrelated to the intervening segment. In contrast, the two clauses which end the suspended segment (at 14.1) and begin the resumption segment (at 16.1) are semantically identical and structurally parallel, supporting the inference that 16.1 resumes the segment containing 14.1.

In Section 5.1 we used the segment shown in Fig. 6 to suggest that subjects agree on the narrator's intentions for reliable segments. In general, pops occur when there is particularly strong semantic discontinuity with the previous segment. Our data reflects this observation, in that the average strength of all boundaries separating intervening and resumption segments in our training corpus is 6.29, while the average strength of all empirically reliable boundaries is only 5.3. Furthermore, the intentions associated with intervening segments as a class are particularly similar across subjects, as illustrated previously in Figs. 6 and 7, and as will be illustrated now.

Consider the excerpt in Fig. 9, where the narrator refers to a boy (who has fallen off his bicycle) in phrase 21.2, suspends this segment to describe a toy belonging to another boy, then resumes reference to the fallen boy in phrase 24.1 of the resumption segment. Note that the pronoun in 24.1 is not ambiguous between a reference to the first boy (mentioned in 21.2) and the second boy (mentioned more recently in 23.1 and 21.5), due to the "pop" of the intervening segment and the resumption of the suspended segment. The 5 subjects who agree on the intervening segment all indicate that the speaker's intentions pertain to describing the toy (while subject g further subdivided the segment, each of the 3 subintentions was about the toy). It is also interesting to note that 4 of the 5 subjects explicitly refer to the intervening nature of the segment (by using terms such as "suspends", "sidebar", "focus of attention is now", and "again").

6 CONCLUSION

We have argued for the need for extensive, rigorous analyses of discourse corpora in order to determine the nature of the interaction of segmentation, coherence and linguistic devices in discourse. We have presented the results of a study of informal, spoken, monologic narratives that looks at certain aspects of this interaction. Our results show that naive subjects agree significantly often on segment boundaries, using an informal notion of speaker intention as the segmentation criterion. Given that subjects were allowed to use any number of segment boundaries in their linear segmentations, average IR scores across subjects at identifying boundaries will necessarily be far from ideal. Furthermore, there is considerable variation across narratives in the IR scores of individual subjects.

While none of our 3 algorithms—NP, CUE and PAUSE—performs as well as the human subjects, the results suggest that with additional knowledge, levels approaching human performance could be achieved algorithmically. However, we argue that no algorithm relying on a single class of linguistic device could hope to achieve high levels of performance across speakers due to variations in speaker strategy both across and within narratives. Another problem is that a discourse produced on one occasion may consist of distinct parts that are only weakly related to one another. Thus, when a coherent narrative episode is extracted from our corpus, the NP algorithm performs far better. Its performance improves even further when segment location is defined using fuzzy rather than precise boundary locations.

21.1 [.4] Three boys came out,

21.2 [.75] helped him$_i$ pick himself up,

21.3 [.55] pick up his bike,

21.4 pick up the pears,

5 subjects (b, c, d, e, g)

21.5 [.55] one of them had [.6] a toy,

21.6 which was like a clapper.

2 subjects (a, g)

22.1 [2.4 [1.1] A-nd [1.1]] I don't know what you call it except a paddle with a ball suspended on a string.

2 subjects (f, g)

23.1 [.25 breath] So you could hear him playing with that.

6 subjects (a, b, c, d, e, g)

24.1 [2.25 [.8] A-nd [.95]] then he$_i$ rode off,

Subject	Annotation of Narrator's Intention for Intervening Segment
b	suspends description of events to describe objects boys had, and sound
c	sidebar description of a paddle ball
d	the focus of attention is now the toy
e	toy is introduced. Basically sensory aspects are mentioned again.
g	one of them had a toy; reader trying to name the toy; boy playing with toy

Figure 9: Intervening segment spanning 21.5 through 23.1

We believe that fuzzy boundary location is an important issue. Although the use of naive subjects necessarily produces noisy data, we believe that segment boundary location will turn out to be inherently imprecise. First, the role of certain utterances may be ambiguous, leading to indeterminacy regarding which segments they belong to. We illustrated one example (cf. Fig. 7) in which a subject who identified a distinct initial segment boundary had a clearly divergent interpretation of the narrator's intentions. In this particular case, the subject's interpretation may reflect poor understanding, but in other cases, ambiguity may reflect equally good competing interpretations. Second, a single utterance may simultaneously have multiple functions, leading to distinct segmentations (cf. discussion of Moore and Pollack [31] in §2). Although natural language processing systems must produce a linear sequence of utterances as output, or produce an abstract model from a linear sequence of input utterances, the evidence suggests that ultimately it may be insufficient to assume that utterances contribute to or derive from only a single intention (or plan) at a time.

Speaker variation is another important issue with differing consequences for generation and understanding. Individual speakers have more or less skill at producing coherent discourse, and a similar point can be made regarding individual hearers. Understanding, generation or interactive systems will need to adapt these differences. Presumably, however, a generation system should produce discourse where choices regarding individual linguistic devices as well as linear sequence more closely resemble the performance of the 'best' or most easily understood speakers. Conversely, an understanding system should be designed to accommodate a wide range of hearers, not just the most adept.

References

[1] J. L. Austin. *How to Do Things with Words.* Oxford University Press, New York, 1962.

[2] M. Beckman. Notes on prosody, 1991. Ms.

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

[4] G. Carlson. A unified analysis of the english bare plural. *Linguistics and Philosophy*, 1:413–457, 1977.

[5] W. L. Chafe. 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, 1980.

[6] W. L. Chafe. *The Pear Stories: Cognitive, Cultural and Linguistic Aspects of Narrative Production.* Ablex Publishing Corporation, Norwood, NJ, 1980.

[7] W. G. Cochran. The comparison of percentages in matched samples. *Biometrika*, 37:256–266, 1950.

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

[9] B. A. Fox. *Discourse Structure and Anaphora: Written and Conversational English.* Cambridge University Press, Cambridge, 1987.

[10] W. Gale, K. W. Church, and D. Yarowsky. Estimating upper and lower bounds on the performance of word-sense disambiguation programs. In *Proc. of ACL*, pages 249–256, Newark, Delaware, 1992.

[11] B. Grosz and J. Hirschberg. Some intonational characteristics of discourse structure. In *Proc. of the International Conference on Spoken Language Processing*, 1992.

[12] B. J. Grosz. *The Representation and Use of Focus in Dialogue Understanding.* PhD thesis, University of California, Berkeley, 1977.

[13] B. J. Grosz and C. L. Sidner. Attention, intentions and the structure of discourse. *Computational Linguistics*, 12:175–204, 1986.

[14] J. A. Hawkins. *Definiteness and Indefiniteness.* Humanities Press, New Jersey, 1978.

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

[16] M. A. Hearst. Multi-paragraph segmentation of expository texts. Technical Report 94/790, Computer Science Division (EECS), University of California, Berkeley, 1994.

[17] I. Heim. File change semantics and the familiarity theory of definiteness. In Bauerle, Schwarze, and von Stechow, editors, *Meaning, Use and Interpretation of Language.* Walter de Gruyter, Berlin, 1983.

[18] J. Hirschberg and B. Grosz. Intonational features of local and global discourse structure. In *Proc. of Darpa Workshop on Speech and Natural Language*, 1992.

[19] J. Hirschberg and D. Litman. Empirical studies on the disambiguation of cue phrases. *Computational Linguistics*, 19, 1993.

[20] J. R. Hobbs. Coherence and coreference. *Cognitive Science*, 3(1):67–90, 1979.

[21] Ł. Iwańska. Discourse structure in factual reporting (in preparation), 1993.

[22] N. S. Johnson. Coding and analyzing experimental protocols. In T. A. Van Dijk, editor, *Handbook of Discourse Analysis, Vol. 2: Dimensions of Discourse.* Academic Press, London, 1985.

[23] H. Kamp. A theory of truth and semantic representation. In J. Groenendijk, T. Janssen, and M. Stokhof, editors, *Formal Methods in the Study of Language, Part I*, pages 277–322. Mathematisch Centrum, Amsterdam, 1981.

[24] L. Karttunen. Discourse referents. In J. McCawley, editor, *Syntax and Semantics, Vol. 7: Notes from the Linguistic Underground*. Academic Press, New York, 1976.

[25] K. Krippendorff. *Content Analysis: An Introduction to Its Methodology*. Sage Publications, Beverly Hills, CA, 1980.

[26] E. Levy. *Communicating Thematic Structure in Narrative Discourse: The Use of Referring Terms and Gestures*. PhD thesis, University of Chicago, 1984.

[27] D. Litman and R. Passonneau. Empirical evidence for intention-based discourse segmentation. In *Proc. of the ACL Workshop on Intentionality and Structure in Discourse Relations*, 1993.

[28] W. C. Mann, C. M. Matthiessen, and S. A. Thompson. Rhetorical structure theory and text analysis. In W. C. Mann and S. A. Thompson, editors, *Discourse Description*. J. Benjamins Pub. Co., Amsterdam, 1992.

[29] W. C. Mann and S. Thompson. Rhetorical structure theory: towards a functional theory of text organization. *TEXT*, pages 243–281, 1988.

[30] J. D. Moore and C. Paris. Planning text for advisory dialogues: Capturing intentional and rhetorical information. *Computational Linguistics*, 19:652–694, 1993.

[31] J. D. Moore and M. E. Pollack. A problem for RST: The need for multi-level discourse analysis. *Computational Linguistics*, 18:537–544, 1992.

[32] J. Morris and G. Hirst. Lexical cohesion computed by thesaural relations as an indicator of the structure of text. *Computational Linguistics*, 17:21–48, 1991.

[33] R. J. Passonneau. Coding scheme and algorithm for identification of discourse segment boundaries on the basis of the distribution of referential noun phrases. Technical report, Columbia University, 1993.

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

[35] R. J. Passonneau. A plan based architecture for processing definite and indefinite descriptions in discourse, 1994. Manuscript in review.

[36] R. J. Passonneau and D. Litman. Intention-based segmentation: Reliability and correlation with linguistic cues. In *Proceedings of the 31st Annual Meeting of the Association for Computational Linguistics*, 1993.

[37] J. Pierrehumbert. *The Phonology and Phonetics of English Intonation*. PhD thesis, M.I.T., 1980.

[38] L. Polanyi. A formal model of the structure of discourse. *Journal of Pragmatics*, 12:601–638, 1988.

[39] E. F. Prince. Towards a taxonomy of given-new information. In P. Cole, editor, *Radical Pragmatics*, pages 223–55. Academic Press, New York, 1981.

[40] R. Reichman. *Getting Computers to Talk Like You and Me*. MIT Press, Cambridge, Massachusetts, 1985.

[41] J. A. Rotondo. Clustering analysis of subject partitions of text. *Discourse Processes*, 7:69–88, 1984.

[42] E. Schegloff and H. Sacks. Opening up closings. *Semiotica*, 8:289–327, 1973.

[43] R. J. (Passonneau) Schiffman. *Discourse Constraints on IT and THAT: A Study of Language Use in Career-Counseling Interviews*. PhD thesis, University of Chicago, 1985.

[44] C. L. Sidner. Towards a computational theory of definite anaphora comprehension in English discourse. Technical report, MIT AI Lab, 1979.

[45] S. A. Thompson and T. Ono. Interaction and syntax in the structure of conversational discourse. This volume.

[46] B. Webber. A formal approach to discourse anaphora. Technical Report 3761, Bolt Beranek and Newman Inc., 1978.

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