University of Pennsylvania Working Papers in Linguistics

Volume 7
Issue 2 *Current Work in Linguistics (Working Papers)*

Article 4

1-1-2002

Empirical Studies of Centering Shifts and Cue Phrases as Embedded Segment Boundary Markers

Kate Forbes
University of Pennsylvania, forbesk@babel.ling.upenn.edu

Eleni Miltsakaki
University of Pennsylvania, elenimi@babel.ling.upenn.edu

Empirical Studies of Centering Shifts and Cue Phrases as Embedded Segment Boundary Markers

Empirical Studies of Centering Shifts and Cue Phrases as Embedded Segment Boundary Markers

Kate Forbes and Eleni Miltsakaki

1 Introduction

Previous experiments in discourse have shown that subjects intuitively perceive discourses as being constructed from smaller discourse segments, and generally agree that segment boundaries correspond to an interpretation of topic shift or discourse goal fulfillment.[1] However, when asked to indicate the exact placement of segment boundaries, their responses are often uncertain and differ by one or more utterances. This is precisely the heart of the difficulty of this issue, which has created significant obstacles in the research effort to automate the indentification of segment boundaries.

This paper suggests that some of the confusion about where to place discourse segment boundaries can be lessened by the use of discourse segment boundary *markers*. In the Grosz and Sidner stack model of discourse, discourse segment goals (intentions) underlie discourse segments; the fulfillment of discourse segments goals achieves an overall discourse goal. Processing a discourse segment creates a focus state containing the objects, properties and relations relevant to that segment. The focusing structure is modeled as a stack, thus allowing segments to be ordered either hierarchically or linearly with respect to other segments.[2] Within each discourse segment, Centering Theory (Joshi and Kuhn (1979), Joshi and Weinstein (1981), Grosz and Sidner

[1]According to Passonneau and Litman (1993), the most reliable criterion enabling human subjects to perform segmentation is *speakers' intentions*. However, the task becomes much harder when subjects are faced with longer, hierarchically-structured texts.

[2]In the stack model the hierarchical or linear order order is achieved in the following way. Processing a discourse segment creates a focus state containing the objects, properties and relations relevant to that segment. The focusing structure is modeled as a stack. Elements can be placed on top (pushed) or taken off top (popped), but at any point in time only the topmost element on the stack is accessible. If pushing and popping correspond respectively to the initiation and completion of processing, then the stack models the order with which elements are processed. If an element is pushed and then popped from the stack before a second element is pushed, the order of processing the two elements is linear. On the other hand, if a second element is pushed on top of an element already on the stack, then the processing of the second element must be completed (popped) before the processing of the element lower in the stack. In

(1986)), has been proposed as a model of discourse coherence, which tracks the movement of entities through a focus state by one of four possible focus shifts. Using the stack model, this paper investigates whether Centering shifts and other discourse segment boundary markers can be used to identify the hierarchical structure of a discourse.

This paper builds on previous work (Forbes (1999), Passonneau and Litman (1997), Grosz and Sidner (1986)), observing that a) Rough Shifts often correspond to the intuitive boundaries of embedded discourse segments in goal-oriented discourse, b) the Rough Shift transition marking the end boundary of an embedded segment often corresponds to a Continue transition, if the embedded section is ignored (popped-out), and c) Informationally Redundant Utterances (IRUs) mark the boundaries of embedded segments. Additional boundary markers, i.e., cue-words and phrases, are also investigated. Though such markers might serve a variety of methods for producing a replicable method of discourse segmentation, in this paper we classify them as PushCues, PopCues, and LinearCues, for use in a simple stack-based algorithm for segmenting discourse.

This paper is organized as follows: In Section 2 we discuss related work in automating discourse segmentation, and in Section 3 we present our own four-step procedure and describe its performance on five dialogues. In Section 4 we discuss these results, and Section 5 concludes with a discussion of possible future work.

2 Related Work

Research in automated discourse segmentation has been guided by two main approaches: the lexical cohesion approach (Morris and Hirst (1991), Youmans (1991), Hearst (1994), Kozima (1993), Reynar (1994)) and the discourse cues approach. In this section we focus our attention on the latter.

Passonneau and Litman (1997) proposed two sets of algorithms for linear segmentation based on linguistic features of discourse. With the first set they evaluated the correlation of discourse segmentation with three types of linguistic cues: referential pronoun phrases, cue words and pauses. With the second set they used error analysis and machine learning. An important result of this

this case, the order of processing of the two elements is hierarchical: the processing of the second element is embedded in the processing of the first element. Thus, by pushing and popping focus states, discourse segments (i.e. the processing of their relevant objects, properties and relations) will be ordered either hierarchicially or linearly with respect to other segments.

work was that linear segmentation algorithms based on any one type of linguistic cues performed much poorer than algorithms which utilized linguistic cues from multiple sources. Our works benefits from this insight in that we, too, employ multiple types of features for the indentification of hierarchical segmentation.

Passonneau (1998) investigated the relationship of Centering transition types with segment boundaries.[3] Two versions of Centering were computed: version A as in Brennan, Walker-Friedman, and Pollard (1987) and version B as in Kameyama, Passonneau, and Poesio (1993). Correlations were very poor and Passonneau (1998) concluded that Centering Transitions do not directly reflect segmental structure. Note, however, that in both versions A and B the Centering transitions were significantly modified. In version B, the authors defined new transitions which differ significantly from the original Centering Transitions and in version A the Rough-Shift transition was collapsed with the Smooth-Shift transition. In our study, we specifically show that the Rough-Shift transition plays a significant role in the identification of embedded segments. It is possible that the significance of the Rough-Shift transition was overlooked in Passonneau's study due to the fact that the segment boundaries identified by human raters in their corpus were mostly linear. In version B, the authors defined new transitions which differ significantly from the original Centering Transitions.

Grosz and Sidner (1986) have shown that Informationally Redundant Utterances (IRUs) indicate embedded segments. Walker (1993) argues that, with respect to a well defined task, IRUs are used by resource-limited agents as a discourse strategy to improve the efficiency of completing a task. The distribution of IRUs in her corpus indicates that IRUs function as markers of returning to a superior segment, an observation compatible with her claim that IRUs reestablish the salience status of an earlier proposition. An example is shown below. The IRU is capitalized.

(1) H: ...but I would suggest this - if all of these are 6 month certificates and I presume they are

(2) E: yes

(3) H: then I would like to see you start spreading some of that money around

(4) (Discussion about retirement investments)

(5) but as far as the certificates are concerned, I'D LIKE THEM SPREAD OUT A LITTLE BIT -THEY'RE ALL 6 MONTHS CERTIFICATES....

[3] An overview of Centering Theory is given in Section 3.2.

An additional marker of return to a superior segment was observed in Forbes (1999). Forbes (1999) noticed that (at least in goal-oriented discourse) certain Rough-Shift transitions would in fact be Continue transitions if they were computed with respect to the last utterance appearing before an embedded segment. This preliminary work opened up the possibility that identifying this pattern of behavior could in fact be utilized for the identification of an embedded segment.[4]

Our current research adopts an overall similar approach to discourse segmentation in that we, too, identify linguistic cues for discourse segmentation. The novelty in our approach is that we focus our investigation on the exploration of markers of embedded (and not linear) segments and we develop a methodology for building an exhaustive list of such markers.[5] Placing our work in the framework of the stack model, we seek to build a system which will recognize *pushes* as the start of an embedded segment and *pop-outs* as closing off such embeddings. We start off with the hypothesis that Rough-Shifts, IRUs and other discourse cues are indicators of embedded structures and develop a methodology for gradually identifying such cues.

3 The Study

In this study we apply a four-step procedure for discourse segmentation to five dialogues: 1) utterance level segmentation, 2) coreference tagging, 3) computing Centering transitions. In the fourth step, a machine learning technique is used to classify discourse segment boundary markers as PushCues, PopCues, and LinearCues; these cues then function as conditions in a simple stack-based algorithm for segmenting discourse into hierarchical and parallel levels. First, we intuitively segment five discourses according to their goal structure, as discussed in 3.1. We then perform the first three steps stated above, discussed in 3.2.[6] As discussed in 3.3, we then divide the five discourses into three sets, corresponding to three stages: training, retraining, and evaluation. We initialize our algorithm with the RoughShift PushCue, (H), the RoughShift + Continue PopCue, (H+C) and the IRU. In both the training and the evaluation stages we test the ability of this algorithm to correctly retrieve the segments we labeled,

[4]Building on this result, in the current study, we call this pattern RoughShift + Continue and classify it as a PopCue.

[5]Earlier work (Hirschberg & Litman, 1993), has already shown that certain cues (i.e., *now*) are used to signal the beginning of a subgoal.

[6]Algorithms do exist for these three steps, but to reduce error we performed them manually.

and investigate the validity of any additional segments which the algorithm retrieves. In the training stages we add to the algorithm any boundary markers which it does not already contain.

3.1 Goal Structure of the Dialogues

As suggested by Grosz and Sidner (1986), discourse segmentation, or parceling a discourse into subgoals, arguably serves an information packaging purpose, the communicative effect of which is *comprehension*. In order to investigate what cues are available in a discourse to indicate its subgoal boundaries, we sought discourse that displayed a clear goal-oriented structure. We randomly chose five moderately-lengthed dialogues from the Harry Gross Financial Radio Talk Show.[7]

As shown in Forbes (1999), these dialogues can generally be described by the primary goal 'Obtain Answer to Financial Question'. In order to achieve this goal, each caller, in collaboration with Harry Gross, has to fulfill the following structured series of subgoals (assumed to be mutually known to both caller and Harry):

- GOAL: Obtain Answer to Financial Question
    - SUBGOAL: Caller Greet Harry Gross, Harry Gross Greet Caller
    - SUBGOAL: Caller Ask Financial Question
        - SUBGOAL: Caller Describe Details of Financial Question
    - SUBGOAL: Harry Gross Answer Financial Question
        - SUBGOAL: Harry Gross Clarify Financial Question/Answer
    - SUBGOAL: Caller Say Goodbye, Harry Gross Say Goodbye

Though slight variations were found in individual dialogues, with this structure as a guide we manually labeled the likely parallel and embedded segment boundaries that were present.

3.2 Annotating the Dialogues

Below we give a brief overview of Centering (Walker, Joshi, and Prince (1998), inter alia) and then discuss our methodology for performing the first three steps in our procedure for discourse segmentation.

---

[7]Thanks to Julia Hirschberg and Martha Pollack for originally transcribing this corpus and Ellen Prince for pointing it out to us.

	$Cb(U_i) = Cb(U_i - 1)$	$Cb(U_i) \neq Cb(U_i - 1)$
$Cb(U_i) = Cp$	Continue	Smooth-shift
$Cb(U_i) \neq Cp$	Retain	Rough-shift

Table 1: Table of transitions

A Brief Overview of Centering

In Centering Theory, each discourse segment consists of utterances designated as U_i. Each utterance U_i evokes a *set* of discourse entities, the FORWARD-LOOKING CENTERS, $Cf(U_i)$. The highest-ranked entity in $Cf(U_{i-1})$ *realized* in U_i is the BACKWARD-LOOKING CENTER, Cb. The highest-ranked member in U_i is the PREFERRED CENTER, Cp. The members of the Cf list are ranked as follows:

Subject>Indirect Object>Object>others

Four types of transitions are defined reflecting variations in the degree of topic continuity and are computed according to Table 1.

Discourse coherence is computed according to the following *ordering rule*.

Ordering rule Continue is preferred to Retain, which is preferred to Smooth-shift, which is preferred to Rough-shift.

3.2.1 Step 1: Utterance-Level Segmentation

Following Miltsakaki (1999), we define the utterance as the traditional 'sentence', i.e. the main clause and its accompanying subordinate and adjunct clauses constitute a single utterance unit. Self-corrections do not constitute independent utterance units.

Following Eckert and Strube (1999) we tag single utterances as <I>, acknowledgments (e.g. 'yes', 'sure') as <A>, and acknowledgments followed by an utterance (including answers to questions) as <A-I>. Centering transitions are computed only for <I>, <I-A> and <A-I>.

3.2.2 Step 2 and Step 3: Coreference and Centering

Coreference is done manually in this study; each entity is given a unique REF number. [8]

[8] We do not assume a Centering-based algorithm for anaphora resolution.

Centering the Dialogues. We use standard Centering Theory, except that we compute transitions across segment boundaries. Following Miltsakaki and Kukich (2000), we tag only the Cp within an utterance, and tag remaining entities as OTHER. The Cb, the Cp, and the transition (Tr) are listed next to each utterance.

Following Walker and Prince (1995) we treat the Cf ranking of multiple evoked entities in complex NPs (e.g. ‘his mother’, ‘software industry’) as ordered from left to right. As Walker and Prince (1995) have pointed out, restricting the relationship of $\mathrm{Cb}(\mathrm{U}_{i-1})$ and $\mathrm{Cb}(\mathrm{U}_i)$ to strict coreference is inadequate due to cases of functional dependence between evoked entities. We do not attempt to solve this problem in this paper. However, the discourses we investigate contain both inferrables and discourse deixis, and thus we were forced to address the issue. Based on the intuition that inferrables and discourse deixis do *not* indicate a Rough Shift transition we decided to link them to the the Cf set of the previous utterance. We did that by giving instances of deixis and inferrables a unique REF number added to the Cf set, as is done with all other newly introduced entities. In this way, we accommodated cases in which a pronominal referred back to a deictic. Additionally, however, we added to the Cf set the REF numbers of all the entities in the preceding utterance that were evoked in the part of text referred to by the deictic. These REF numbers were added according to their ranking. An example follows in (1). Indefinite plural noun phrases (e.g. ‘people’, ‘they’, ‘we’) are ranked lowest in the Cf list. An example follows in (2).

1. <I><CP REF=1’>I</CP>’m assuming that <OTHER REF=‘11(10,9)’>that</OTHER> is the case</I>

2. <I>h. in reality what <OTHER REF=‘2’> we </OTHER> have to recognize is that <CP REF=‘1’> you </CP> don’t have <OTHER REF=‘3’>losses</OTHER></I>

3.3 Retrieving Segmentation Cues

We divide the annotated dialogues into three sets, corresponding to three stages: two for training, two for retraining, and one for evaluation. As shown in Table 2, we initialize our algorithm with the RoughShift PushCue (H), the RoughShift + Continue PopCue (H+C), and the IRU embedded segment boundary marker (IRU). We do not assume any Linear Cues. However, we annotated linear segments, so, in addition to the list of Push and Pop Cues, we report numbers of linear segments and cues. Interestingly, we did not identify any

PushCues	PopCues	LinearCue
H	H+C	none
U_i	$IRU_i(=U_i)$	

Table 2: Initial Algorithm

linear cues. In Table 2, we designate the utterance to which the IRU refers as U_i and we indicate this relationship with $IRU_i(=U_i)$.

3.3.1 Stage 1: Training on 'Andy' and 'Eleanor'

We present below part of the 'Andy' dialogue, to exemplify the testing and training of the algorithm. The 'level' on the stack at which each discourse segment resides is indicated by indentation. For ease in reading, only relevant OTHER references are tagged. Cue phrases and IRUs are in italics.

- U1 <I>a. hi <OTHER REF='2'>harry</OTHER> this is <CP REF='1'>andy</CP></I> <A> h. welcome andy</A> **Cb=1 Cp=1 Tr=none**
  - U2 <I>a. uh *first* like to say <CP REF='1'>i</CP>'m glad to hear <OTHER REF='2'> you</OTHER>'re back on in the afternoons</I> **Cb=1 Cp=1 Tr=C**
- U6 <I>a. uh <OTHER REF='3'><CP REF='1'>*my*</CP>

  question</OTHER> *is a tax one.*</I> **Cb=none Cp=1 Tr=H**
  - U7 <I> uh ><CP REF='1'>I</CP> bought <OTHER REF='4'>

    a property</OTHER> a resort property in 1978 </I> **Cb=1 Cp=1 Tr=C**
    - U8 <I><CP REF='6'>the purpose</CP> in buying <OTHER REF='4'>this property</OTHER>was basically appreciation and tax shelter</I> **Cb=4 Cp=6 Tr=H**
  - U9 <I>*and uh each year on my irs return* <CP REF='1'> *I've*</CP> *claimed a loss*</I> **Cb=none Cp=1 Tr=H**
  - U10 <I> uh the income on the property being a resort and a seasonal kind of rental the income uh it's very unlikely <CP REF='7'>the income</CP> will ever exceed <OTHER REF='8'>the expenses</OTHER></I> **Cb=none Cp=7 Tr=H**

- U12 <I> *so* <OTHER REF='9'>*each year*</OTHER> *on my return* <CP REF='1'>*I've*</CP> *had a loss*</I> **Cb=none Cp=1 Tr=H**

• U14 <OTHER REF='3'><CP REF='1'>*my*</CP> *question*</OTHER>*is, how long can I claim a loss?*</I> **Cb=none Cp=1 Tr=H**

U1-U5 corresponds to the first subgoal 'Greet'. Though this segment is generally ordered linearly with respect to 'Ask Question', in 'Andy''s case, there is an additional subgoal embedded within the 'Greet' goal, which could be referred to as 'Encourage Harry's Work'. We argue that this subgoal is embedded due to the likelihood that the speaker does not intend to return to this 'topic', nor does it pertain directly to his overall goal. This segment is cued by the PushCue 'first' in U2; we thus add it to our algorithm. The end of this segment is cued by the PopCue H+C, because the transition from U1 to U6 would be computed as a CONTINUE (C) if the embedded segment (U2-U5) were omitted.

We find two IRU pairs in this excerpt: (U6,U14), and (U9, U12); the algorithm treats these utterances as marking the boundaries of embedded segments, and embeds the intervening utterances within them. U6-U14 correspond to the subgoal 'Ask Question'. U7 initiates the subgoal 'Describe Question'.[9] We label U8 as an embedded segment; in it the caller refers only parenthetically to his purpose; in U9 he returns to the line of thought (background information) pursued in U7. This embedding is correctly labeled by the algorithm due to the H transition in U8 and the H+C transition in U9. In U10-U11, embedded by the IRU (U9, U12) and redundantly signaled by the H transition of U10, the caller adds an explanation to complete the subgoal 'Describe Question'.

Table 3 presents the results of the algorithm on the 'Andy' dialogue. As shown, the algorithm did not retrieve two segment boundaries endings; one of these is cued by the phrase, 'in any event', and the other by a 'reverse wh-cleft'; we thus added these cues to the algorithm.[10] The algorithm additionally

[9]The subgoal initiated in U7 may be 'redundantly' cued by 'an explicit tense reference'. In these dialogues, these references are usually to the past, and are used when supplying background information. The exact nature of this PushCue deserves further study. We did not find a strong correlation between tense change and segment boundaries; tagging each utterance for tense (present and past) we found: 1) Embedded segment begins/ends when change/no change in Tense, 2) Change in Tense when no Embedded segment begins/ends. However, Iida (1998) shows that change of tense signals the start of or return to a superior segment.

[10]Wh-clefts have found support in the literature as an indication of a completed goal,

PushCue	# of instances	PopCue	# of instances	LinearCue	# of instances
U_i	3	$IRU_i(=U_i)$	3	none	3
H	6	H+C	4		
'first' (added)	1	'in any event' (added)	1		
		reverse wh-cleft (added)	1		

Table 3: Training on 'Andy'

retrieved a segment that we had not labeled, but which corresponded to a parenthetical. The three linear segment boundaries were not retrievable.

Eleanor: We found this dialogue difficult to analyze; our segmentation was thus very rough-grained. As shown in Table 4, The algorithm did not retrieve two segment boundaries that we had labeled; we added the PushCue *now in addition* and the PopCue *I was wondering* + *H* to retrieve them. In the latter case, the Cue phrase would have been incorrectly characterized as a Pushed segment, due to the simple H marker. The H PushCue and H+C PopCue retrieved a segment boundary that we had not labeled but which did not seem implausible. And again, none of the linear segments were cued. The final algorithm at the end of Stage 1 training is shown in Table 5.

3.3.2 Stage 2: Retraining on 'Jim' and 'Judy'

Jim: As shown in Table 6, one embedded segment boundary beginning was not retrieved by the algorithm; we added the new PushCue, 'oh now hang on' to retrieve it[11]. All linear segment boundaries remain uncued.

Judy: As shown in Table 7, the algorithm did not retrieve one embedded segment boundary ending; we added the PopCue *right* + *H*, because otherwise that segment boundary would have been incorrectly characterized as a Pushed segment. All linear segment boundaries remain uncued. Our final algorithm is shown in Table 8.

---

see Ball (1994) and Delin and Oberlander (1995).

[11]In this dialogue, we found possible additional PopCues 'well now' and 'ok', but did not include them because we were not certain of the underlying goal structure. Another embedded segment boundary ending was 'redundantly' cued by 'in that case'.

PushCue	# of instances	PopCue	# of instances	LinearCue	# of instances
U_i	2	$IRU_i(=U_i)$	2	none	3
H	1	H+C	1		
'first'	0	'in any event'	0		
		reverse wh-cleft	0		
'now in addition' (added)	1	H + 'I was wondering' (added)	1		

Table 4: Training on 'Eleanor'

PushCues	PopCues	LinearCue
U_i	$IRU_i(=U_i)$	none
H	H+C	
'now in addition'	reverse wh-cleft	
first	'in any event'	
	H + 'I was wondering'	

Table 5: Final Algorithm: Stage1

PushCue	# of instances	PopCue	# of instances	LinearCue	# of instances
H	2	H+C	3	none	3
'oh now hang on' (missed)	1				

Table 6: Retraining on 'Jim'

PushCue	# of instances	PopCue	# of instances	LinearCue	# of instances
H	2	H+C	1	none	3
IRU	1	IRU	1		
		H + 'right' (missed)	1		

Table 7: Retraining on 'Judy'

PushCues	PopCues	LinearCue
U_i	$IRU_i(=U_i)$	none
H	H+C	
'now in addition'	reverse wh-cleft	
first	'in any event'	
'oh now hang on'	H + 'I was wondering'	
	H + 'right'	

Table 8: Final Algorithm: Stage2

3.3.3 Stage 3: Evaluation on 'Susan'

The 'Susan' discourse (see Appendix) contains three people; it was used to evaluate the algorithm. As shown in Table 9, only one PushCue 'you see' was missed (U18), and a corresponding PopCue, 'so' (U20). Additionally, there appears to be an embedded goal within the 'Greet' subgoal (U3-U4), though there was no corresponding Cue phrase. As before, no linear segments were cued.

PushCue	# of instances	PopCue	# of instances	LinearCue	# of instances
H	5	H+C	5	none	4
'you see' (missed)	1	'so' (missed)	1		

Table 9: Testing on 'Susan'

4 Discussion

A number of issues arise concerning the retrieval of PushCues and PopCues. The first is that parentheticals and subgoals are not clearly differentiated. We did not always label parentheticals, but if a parenthetical causes a Rough-Shift, our algorithm will label it as embedded. Moreover, cue phrases themselves can be either ambiguous or redundant. Such cases require the investigation of more dialogues, to determine their exact status.[12] Finally, we found no cue phrases for parallel segment boundaries. While overt statements like *hello* and *thank you* could be treated as cues for the 'Greet' and 'Goodbye' segments, we would need some method of 'defaulting' in the other cases.

A number of issues arise concerning our proposal for segmenting discourse. First, we have seen that Cue phrases used to indicate embedded segments in one dialogue are not often used in other dialogues, even by Harry himself. It will require the investigation of many more dialogues to see if this is a problem for the algorithm. It would not be tractable if the list grew exponentially with the number of dialogues investigated, or if the use of a cue phrase/word varied. Another potentially difficult problem lies in extracting the relevant sense of Cues (e.g. *first* as a topic marker versus *first* in its other uses).[13] Finally, we found that statistically, H, H+C and IRUs were the most frequent indicators of segment boundaries. A tractable implementation of the algorithm must thus avoid the need to keep track of the exact phrasing of all previous utterances. For the shifts, the algorithm need only search to find a referent in a preceding push, but for IRUs the situation might prove more difficult.

5 Future Work

This study can be seen as a preliminary stage of research into automating discourse segmentation. A number of issues arise for future research: undetermined Centering concepts (deictics, inferrables, complex NPs, inter alia); the role of tense and aspect; the disambiguation of the various uses of cue words,

[12]This issue is exemplified by the PushCue *now in addition* in the 'Eleanor' dialogue; the segment intuitively feels embedded, but the Cue seems indicative of a parallel segment.

[13]Litman (1993), Hirschberg and Litman (1993) discuss the distinction between structural and sentential cues. They use 'orthography' to disambiguate them. Their experiments with machine learning showed that certain cues, in various sentence positions, are always used for discourse purposes, including 'now' and 'OK'.

and the determination of how the use of cues may vary use within and across speakers and discourses. More generally, the issue remains of how to determine the optimal methodology for identifying the strategies speakers use to signal segments.

We have investigated the possibility of producing a replicable method of discourse segmentation. We have concluded that discourse segments are signaled by a combination of diverse factors or features. In our opinion, discourse segmentation is thus an area in which a combination of linguistic and statistical approaches will yield optimal results; a wide variety of discourse must be investigated to determine the breadth and use of the set of features involved. We have shown that insights from theoretical approaches such as Centering Theory can provide a starting set of features, which then can be used in an annotation effort for a variety of text genres. Such features, fed into a machine learning project and considered along with lexical cues and information structures, will yield optimal combinations of features that correlate with segment boundaries. We leave this project for future work.

Appendix

SUSAN

- U1.<I>s. hello <CP REF='1'>harry</CP></I> <A>h. yes</A> **Cb=none Cp=1 Tr=none**

- U2.<I>s. this is <CP REF='2'>susan</CP></I> <A>h. <OTHER REF='1'></OTHER> welcome susan</A> **Cb=2 Cp=2 Tr=S**

- U3.<I>s.<CP REF='2'></CP>so glad to have<OTHER REF='1'>you</OTHER> back </I> **Cb=2 Cp=2 Tr=C**

- U4.<I><CP REF='2'>i</CP>'ve been telling all my <OTHER REF='3'>friends and neighbors</OTHER> to put <OTHER REF='1'>you</OTHER>on</I> <A>h. thank you very much</A> **Cb=2 Cp=2 Tr=C**

  LINEAR SEGMENT BOUNDARY IRU?

- U5.<I>s.<CP REF='1'>i</CP> have <OTHER REF='4'>a question</OTHER>,</I> **Cb=2 Cp=2 Tr=C**

  - PUSHCUE: H

- U6.<I>if <OTHER REF='5'>you</OTHER> have <OTHER REF='6'>a certificate </OTHER> and they and you get <OTHER REF='7'>your interest</OTHER>like say for 81, um do you have to put <CP REF='7,6'>that</CP> on your <OTHER REF='8'>income tax</OTHER></I> or could you wait until you exchange the certificate</I><A>h. fred?</A> **Cb=none Cp=7,6 Tr=H**
- LINEAR SEGMENT BOUNDARY
- U7.<I>f. unfortunately <OTHER REF='5'>you</OTHER> have to report <CP REF='7,6'>that</OTHER> in <OTHER REF='9'>the year in which it was earned </OTHER></I> **Cb=7,6 Cp=7,6 Tr=C**
  * PUSHCUE: H
  * U8.<I>and<OTHER REF='5'> you</OTHER>will get <CP REF='10'>a form 1099</CP> from uncle sam - from <OTHER REF='11'>the bank</OTHER> in most cases</I> **Cb=5 Cp=10 Tr=H**
- POP: H+C
- U9.<I>and <CP REF='7'>it</CP>will be reportable before <OTHER REF='5'> you</OTHER> cash in <OTHER REF='6'>the certificate</OTHER></I> <A>s. ok</A> **Cb=none Cp=7 Tr=H**

POP H + C (IRU?)

- U10.<I>- and<CP REF='2'></CP>just one more <OTHER REF='4'>short one</OTHER> <OTHER REF='1'>harry</OTHER></I> <A>h. sure</A> **Cb=2 Cp=2 Tr=H**
- U11.<I>s.<CP REF='2'>my</CP> <OTHER REF='12'>husband</OTHER>retired, and uh <CP REF='13'>this</CP> is the first year,</I> **Cb=2 Cp=2 Tr=C**
  - PUSHCUE: H
  - U12.<I>now what's <CP REF='14'>this</CP> about quarterly income tax papers, that <OTHER REF='5'>you</OTHER>'re supposed to report <OTHER REF='15'>every 3 or 4 months</OTHER>?</I> **Cb=none Cp=14 Tr=H**

- U13.<I>h.<OTHER REF='5'>you</OTHER>'re talking about <CP REF='14'>an estimated tax return</CP></I> **Cb=14 Cp=14 Tr=C**
- U14.<I>s. yes, do <CP REF='2'>i</CP> have to do <OTHERREF='14'>that</OTHER> this year?</I> **Cb=14 Cp=2 Tr=R**
- LINEAR SEGMENT BOUNDARY
- U15.<I>f. well,susan i guess <CP REF='16,2,14'>the answer to that</CP>depends in part as to how much <OTHERREF='2'><OTHER REF='17'>your</OTHER> husband</OTHER> will get in retirement and whether or not <OTHER REF='18'>they</OTHER> will be taking withholding out of <OTHER REF='19'>retirement payment</OTHER> </I> <A>s. oh</A> **Cb=14 Cp=16,2,14 Tr=R**
- U16.<I>f. you see - all <CP REF='14'>an estimated payment</CP> is is an attempt for <OTHER REF='18'>the government</OTHER> to treat people who do not have <OTHER REF='20'> salary income</OTHER> on the same basis as people who do get salary income</I> <A>s. uh huh</A> **Cb=14 Cp=14 Tr=C**
- U17.<I>f. where <CP REF='18'>they</CP> in effect hold some of <OTHER REF='20'>it</OTHER> as they go along -</I> **Cb=18 Cp=18 Tr=S**
- U18.<I>so <CP REF='18'>they</CP>'re saying we're not gonna let <OTHER REF='5'>you</OTHER>keep your <OTHER REF='20'>own money</OTHER> all year,</I> **Cb=18 Cp=18 Tr=C**
- U19.<I><CP REF='18'>we</CP> want you to send <OTHER REF='20'>it</OTHER> in to us as we go along</I> **Cb=18 Cp=18 Tr=C**

POP H+C

- U20.<I>s. well how.. in other words <CP REF='2'>i</CP>'ll probably go up to <OTHER REF='21'>the bureau of internal revenue</OTHER> because i still don't understand <OTHER REF='17/14'>it</OTHER></I>, **Cb=none Cp=2 Tr=H**

- U21.<I>and <OTHER REF='5'>you</OTHER> have to put down <OTHER REF='14'></OTHER></OTHER><CP REF='22'>what you got for the four months</CP></I> **Cb=14 Cp=14 Tr=H**

- U22.<I>h. no no <CP REF='14'>it</CP>'s not done on <OTHER REF='22'>that basis</OTHER>–</I>]bf Cb=14 Cp=14 Tr=C

  - PUSHCUE: H

  - U23.<I>no <OTHER REF='5'>you</OTHER> have to really have to estimate what your <OTHER REF='23'>tax</OTHER> will be for the current year..</I> <A>s. for the whole year</A> **Cb=none Cp=23 Tr=H**

  - U24.<I>h. and then, yes, and then </CP> split <OTHER REF='23'>it</OTHER></I> <A>s. uh huh, um</A> **Cb=23 Cp=23 Tr=S**

  POPCUE: H+C

- U25.<I>h.<CP REF='14'>it</CP>'s not an easy thing to do the first year <OTHER REF='2'>you</OTHER>'re in</I> **Cb=none Cp=14 Tr=R**

- U26.<I>s. (ha ha) <CP REF='2'>we</CP>'ll pay somebody to do <OTHER REF='14'>it</OTHER></I> **Cb=14 Cp=2 Tr=R**

- U27.<I>h. no <CP REF='21'>internal revenue</CP> will help <OTHER REF='2'>you</OTHER><OTHER REF='14'></OTHER> if you can get into if you can get to them,</I> **Cb=14 Cp=21 Tr=R**

  - PUSHCUE: H

  - U28.<I><CP REF='25'>the difficulty</CP> is that <OTHER REF='21'>they</OTHER>have slashed their program to ribbons</I> <A>s. uh huh</A> **Cb=21 Cp=25 Tr=H**

  POPCUE H + C

- U29.<I>f. <CP REF='14'>it</CP>'s not a difficult form to fill out</I> **Cb=14 Cp=14 Tr=H**

  LINEAR SEGMENT BOUNDARY

- U30.<A-I>s. ok well thank <CP REF='1'>you</CP> very much</A-I> **Cb=1 Cp=1 Tr=S**
- U31.<A-I>h. thank <CP REF='2'>you</CP> very much for your call susan</A> **Cb=2 Cp=2 Tr=S**

References

Ball, C. 1994. The Origins of the Informative-Presupposition It-Cleft. *Journal of Pragmatics.*

Brennan, S., M. Walker-Friedman and C. Pollard. 1987. A Centering Approach to Pronouns. In *Proceedings of the 25th Annual Meeting of the Association for Computational Linguistics*, pp. 155–162. Stanford, Calif.

Delin, J. and J. Oberlander. 1995. Syntactic Constraints on Discourse Structure: The Case of It-clefts. *Linguistics, 33.*

Eckert, M. and M. Strube. 1999. Dialogue Act, Synchronising Units and Anaphora Resolution. In *Proceedings of Amstelogue '99, Workshop on the Semantics and Pragmatics of Dialogue.*

Forbes, K. 1999. Investigating Discourse Segmentation: Centering and Embedding. University of Pennsylvania manuscript.

Grosz, B., and C. Sidner. 1986. Attentions, Intentions and the Structure of Discourse. *Computational Linguistics, 12*, 175–204.

Hearst, M. 1994. Multiparagraph Segmentation of Expository Text. In *Proc. of the 32nd ACL.*

Hirschberg, J. and D. Litman. 1993. Empirical Studies on the Disambiguation of Cue Phrases. *Computational Linguistics, 19*(3), 501–530.

Iida, M. 1998. Discourse Coherence and Shifting Centers in Japanese Texts. In Walker, M., A. Joshi and E. Prince (Eds.), *Centering Theory in Discourse.* Oxford University Press.

Joshi, A. and S. Kuhn. 1979. Centered Logic: The Role of Entity Centered Sentence Representation in Natural Language Inferencing. In *6th International Joint Conference on Artificial Intelligence*, pp. 435–439.

Joshi, A. and S. Weinstein. 1981. Control of Inference: Role of Some Aspects of Discourse Structure: Centering. In *7th International Joint Conference on Artificial Intelligence*, pp. 385–387.

Kameyama, M., R. Passonneau and M. Poesio. 1993. Temporal Centering. In *Proc of the 31st ACL*, pp. 70–77.

Kozima, H. 1993. Text Segmentation Based on Similarity Between Words. In *Proc. of the 31st ACL (Student Session)*, pp. 286–288.

Litman, D. 1993. Cue Phrase Classification Using Machine Learning. *Journal of Artificial Intelligence Research*, *5*, 53–94.

Miltsakaki, E. 1999. Locating Topics in Text Processing. In *Proceedings of Computational Linguistics in the Netherlands (CLIN'99).*

Miltsakaki, E. and K. Kukich. 2000. Automated Evaluation of Coherence in Student Essays. In *Proceedings of the Workshop on Language Rescources and Tools in Educational Applications, LREC 2000.*

Morris, J. and G. Hirst. 1991. Lexical Cohesion Computed by Thesaural Relations as an Indicator of the Structure of the Text. *Computational Linguistics*, *17*, 21–28.

Passonneau, R. 1998. Interaction of Discourse Structure with Explicitness of Discourse Anaphoric Noun Phrases. In Walker, M., Joshi, A., & Prince, E. (Eds.), *Centering Theory in Discourse*, pp. 327–358. Clarendon Press: Oxford.

Passonneau, R. and D. Litman. 1993. Intention-Based Segmentation: Human Reliability and Correlation with Linguistic Cues. In *,Proceedings of the 31st Annual Meeting of the Association for Computational Linguistics (ACL-93)*, Vol. 19, 3, pp. 148–155.

Passonneau, R. and D. Litman. 1997. Discourse Segmentation by Human and Automated Means. *Computational Linguistics*, *23*(1), 103–139.

Reynar, J. 1994. An Automatic Method of Finding Topic Boundaaries. In *Proc. of 32nd ACL (Studen Session)*, pp. 331–333.

Walker, M. 1993. Discourse Strategies for Attention-Limited Agents. Tech. rep., IRCS Report 93-21.

Walker, M., A. Joshi and E. Prince. 1998. *Centering Theory in Discourse.* Clarendon Press: Oxford.

Walker, M. and E. Prince. 1995. A Bilateral Approach to Givenness: A Hearer-Status Algorithm and a Centering Algorithm. In Fretheim, T., & Gundel, J. (Eds.), *Reference and Referent Accessibility*. Amsterdam: John Benjamins.

Youmans, G. 1991. A New Tool for Discourse Analyis: The Vocabulary-Management Profile. *Language*, *67*, 763–789.

Kate Forbes
Department of Linguistics
University of Pennsylvania
Philadelphia, PA 19104-6305
forbesk@babel.ling.upenn.edu

Eleni Miltsakaki
Department of Linguistics
University of Pennsylvania
Philadelphia, PA 19104-6305
elenimi@babel.ling.upenn.edu