Knowledge-Based Approaches to the Segmentation of Oral History Interviews

Pengyi Zhang

Dagobert Soergel

MALACH Technical Report

University of Maryland. College of Information Studies

May 2006

Abstract

This paper applies discourse knowledge to the segmentation of speech transcripts. The paper reviews literature on discourse structure, as well as approaches used in text segmentation and speech segmentation, identifies what features are used and how the features are combined in these approaches. After reviewing the literature, a three-part study is conducted to answer the following three research questions:

- Are discourse-markers indicators of segment boundaries in oral history interviews?

- Are questions good indicators of segment boundaries? Could questions be used as segment boundary or segment continuation indicators?

- Do the discourse structures proposed by Labov and Waletzky (1967, 1997) and Stein and Glenn (1979) hold for oral history interviews? How could this knowledge be used in automatic segmentation?

Methodology, results and analysis of each part of the study are described. Major findings include trends in segmentation and answers to these questions. Limitation of the study is discussed. The paper also suggests future research topic relates to segmentation and discourse analysis.

1. Introduction

This paper proposes approaches to apply discourse knowledge to speech segmentation. Segmentation of speech and text has been studied for more than a decade. Since long streams of unstructured text or speech are very difficult to process by human or by computer programs, segmentation becomes essential, especially for speech retrieval. Segmentation can be used in many other tasks, such as:

- **Information retrieval:** (Bawa, Manku, & Raghavan, 2003; Hearst, & Plaut, 1993; Hearst, 1997; Kim, Candan, & Dönderler, 2005; Reynar, 1999; Yaari, 1997) Segments of documents/audio tapes instead of entire documents/tapes relevant to a query can be retrieved and presented to users, which will save users' time not to read/listen to the entire documents/tapes. Segmentation can also support indexing of documents.

- **Text/Speech navigation:** (Choi, 2000; Kim, 2005) Topic segmentation can also be used to support browsing and navigation. Especially in the case of speech, segmentation would allow users to start browsing at any start point of a segment.

- **Summarization:** (Marcu, 1997; Reynar, 1999) Segments, the output of segmentation, can be used by summarization algorithms to weight the relatively importance of the units in a text.

- **Anaphora resolution:** (Kozima, 1993) Segment boundaries provide valuable restrictions for identification of referents of pronouns and referential none phrases.

- **Language modeling:** (Beeferman, Berger, & Lafferty, 1999) Segmentation deals with the structure of the language, and thus is useful to modeling language.

Segmentation deals with the problem of automatically dividing a stream of text or speech into topically homogeneous blocks (Hearst, 1997), so it is often referred as topic segmentation as well. Chafe (1976) suggested that as a speaker moves from focus to focus, there are certain points at which there may be a more or less radical change in space, time, character configuration, event structure, or even world. At points where all these change are in a maximal way, a topic boundary is present. The segmentation task is to find these boundaries and mark them up to form topically cohesive units.

Among other approaches, discourse structure has been found to contribute to the formation of segments and thus has been applied to text and speech segmentations (Grosz & Sinder, 1986; Passonneau & Litman 1997). Linguistic features, such as speech prosody, cue phrases and cue-words, and nominal reference, are partly conditioned by and thus reflect discourse structure. These features have been used to find segment boundaries in text or speech (Dharanipragada, et. al, 1999; Franz, et. al, 1999; Galley, et. al, 2003; Passonneau & Litman 1997; Reynar, 1999; Tür, et. al, 2001).

This paper attempts to support the general claim that discourse structure is an important feature for information access in spoken language. However, discourse structure largely depends on the domain and genre of the text. This paper deals with text transcriptions of interviews of Holocaust survivors. This paper proposes approaches to segmentation of oral history interviews by applying knowledge about discourse structure (personal experience narratives, Labov & Waletzky, 1967; Story Grammar, Stein & Glenn, 1979) and questions as an indicator of topicality (van Kuppevelt, 1995).

This paper is interested in the following research questions: (in the order of increasing complexity)

- Are discourse-markers indicators of segment boundaries in oral history interviews?

- Are questions good indicators of segment boundaries? Could questions be used as segment boundary or segment continuation indicators?

- Do the discourse structures proposed by Labov and Waletzky (1967, 1997) and Stein and Glenn (1979) hold for oral history interviews? How could this knowledge be used in automatic segmentation?

This paper is organized as follows:

Section 2 gives a review of the literatures on segmentation and discourse analysis.

Section 3 describes a three-part study addressing the above three research questions. Section 3 has four subsections:

- Section 3.0 Data: The MALACH collection
- Section 3.1 Discourse markers
- Section 3.2 Interview questions as boundary indicators
- Section 3.3 Discourse structure of Personal Experience Narratives

Each subsection includes methodology, results, and analysis.

Section 4 concludes the major findings, discusses limitations of the study, and suggests a segmentation approach combining multiple sources of evidence based on the findings.

2. Background and Literature Review

2.1 Discourse Structures

Linguists have been long studying the structure of different discourses. Rhetorical Structure Theory (Mann, & Thompson, 1987; Mann, Matthiessen, & Thompson 1992) is often used in linguistic analysis to define relations among utterances. It is one of the most widely used discourse theories in natural language processing (NPL), for example, in the generation and summarization of text (Carlson, Marcu, & Okurowski, 2003; Hovy, 1993; Marcu, 1997).

Rhetorical Structure Theory (RST) describes texts in a rich and highly connected context, and makes predictions about their characters and effects. It describes functions of text units and the relations between them. An example set of RST relations (Mann, & Thompson, 1987) is show in figure 1:

- Circumstance
- Solutionhood
- Elaboration
- Background
- Enablement and Motivation
    - Enablement
    - Motivation
- Evidence and Justify
    - Evidence
    - Justify
- Relations of Cause
    - Volitional Cause
    - Non-volitional Cause
    - Volitional Result
    - Non-volitional Result
    - Purpose
- Antithesis and Concession
    - Antithesis
    - Concession
- Condition and Otherwise
    - Condition
    - Otherwise
- Interpretation and Evaluation
    - Interpretation
    - Evaluation
- Restatement and Summary
- Restatement
- Summary
- Other Relations
    - Sequence
    - Contrast

Figure 1. One set of RST relations

The introduction of RST is simplified and focused on structure and relations between text units (called text span in RST). Many other details have been left out. RST is widely used in discourse analysis. However, since RST's primary aim is discourse analysis – to provide paths or mappings both form situation to language, explaining how and why some uses of language are chosen, and from language and situation to effect, explaining why particular use of language succeeded or failed (Mann, Matthiessen, & Thompson, 1992), the relations defined are more linked to language use and are general enough to apply to different domains and genres. In terms of discourse structure of a particular genre, RST may not provide enough detail at the practical level to analyze it. For example, in narratives, the relation of sequenced action appears more often than in other types of text, but in RST sequence is labeled only as one of the "other relations".

In terms of discourse structure of narrative text, Story Grammar (Stein, & Glenn, 1979) and structure of Personal Experience Narratives (Labov, & Waletzky, 1967; 1997) have been very influential on later work on discourse analysis.

Based on their study of comprehension of stories by elementary school children, Stein and Glenn (1979) developed a theory of Story Grammar, saying that a story (event) consists of the following elements:

- **Setting:** introduction of main characters, as well as the time and place for the story action.
- **Initiating Event:** An action or happening that sets up a problem or dilemma for the story.
- **Internal Response:** The leading characters' reactions to the initiating event.
- **Attempt:** An action or plan of the leading characters to solve the problem.
- **Consequence:** The result of the actions.
- **Reaction:** A response by the characters to the consequence.

Similarly, the work of Labov and Waletzky (1967; 1997) on personal experience narratives (PEN), finds that a fully developed narrative may include clauses or sequences of clauses with the following functions, roughly in this order:

- **Abstract** is the initial clause(s) in a narrative that reports the entire sequence of events of the narrative, summarizing the story to come. It may or may not occur depending on the language style of the teller.
- **Orientation** clauses introduce characters, temporal and physical settings, situation, and the identities of the participants and their initial behavior. Orientations usually occur near the beginning of a narrative, but may be interjected at other points when needed. The characteristic of orientation tense in English is past tense and past progressive tense.
- **Complication** (or **complicating action**) is a sequential clause that reports a next event in response to a potential question, "And what happened [then]?"
- **Evaluation** is part of the narrative, which reveals the attitude of narrator towards the narrative by emphasizing the relative importance of some narrative units as compared to others. Evaluation often follows a complicating action or a sequence of complicating actions. Often a sequence of complications and evaluations lead up to their climax, the point of maximum suspense, which is the most reportable event.
- **Resolution** is the ending or outcome of a narrative. Usually it is the set of complicating actions that follow the most reportable event. It releases the tension and tells what finally happened.

- **Coda** may appear at the end of the story. The teller may announce via a coda that the story is over (For example, "And that was that."), and bring the narrative back to the time of telling so that the question "what happened then" is no longer appropriate.

Figure 2 shows a possible mapping between the two theories of structure of an event-based discourse.

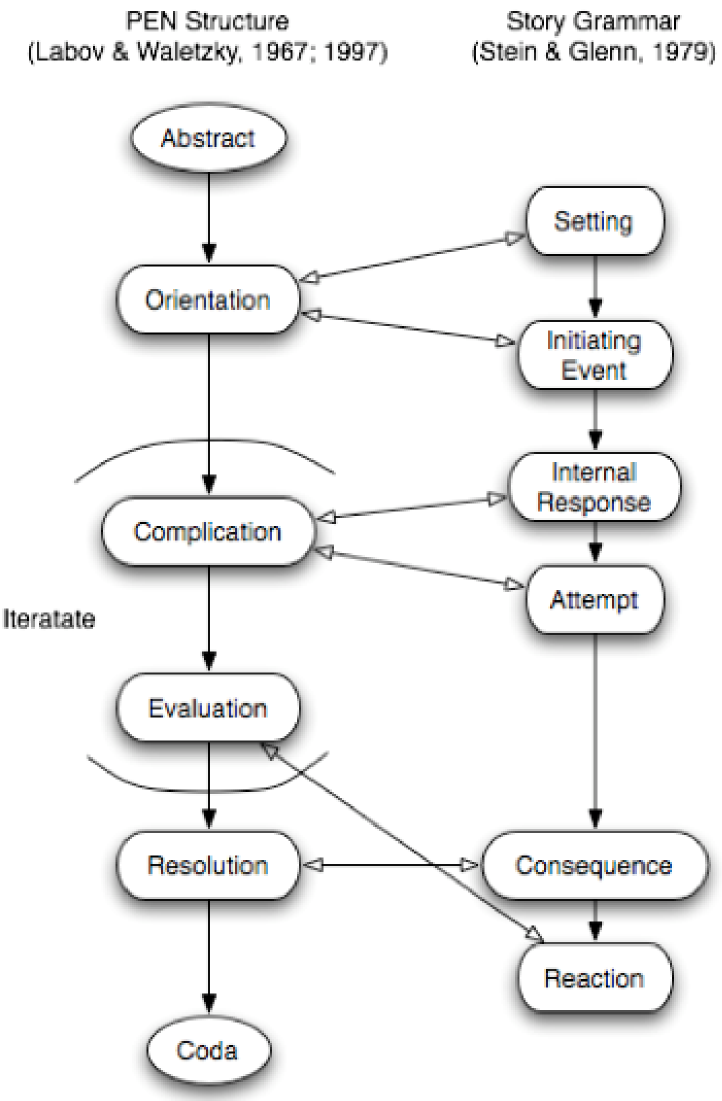

Figure 2: Mapping between Stein & Glenn (1979) and Labov & Waletzky (1967; 1997)

As can be seen from the above mapping, the basic elements in the two theories correspond to each other, except for the optional abstract and coda. The order of these elements differs slightly from Story Grammar to PEN Structure. In Story Grammar, all the events in the story are told, and at last a reaction from the teller to the story is given. In PEN structure, a complicating act is often paired with an evaluation, and resolution of the story comes at the end.

There are different ways for human to create a discourse using similar structures, consciously or unconsciously. Among other indicators of discourse structure, discourse markers are usually related to coherence relations. The study of discourse markers constitutes an extensive area of research in itself. Schiffrin (1987; 2001) believes that discourse markers can have both local and global functions. Discourse markers can not only connect propositional meaning, but also determine the structure of exchange. Discourse markers can impose constraints on the implications the hearer can draw from the discourse (Blakemore, 1992; 2002). They can also act as cohesive devices that cue coherence relations, marking transition points within a sentence, between sentences, or between turns both at the local level and global level (Louwerse, & Mitchell, 2003).

Another indicator is questioning. Questioning, as part of a discourse, is found to play an important role both in forming the structure of a discourse and in contributing to its topicality. Topicality is an important or even a central point of investigation in several theories and views about discourse structure, directly or indirectly (van Kuppevelt, 1995). In discourse analysis, questions and responses are usually considered as adjacency pairs which consist of adjacently ordered first and second pair parts, with the first part setting up constraints on the second (Sacks, 1967, cited from Stenström, 1988). The demanding nature of questions sets restrictions on forthcoming answers. Evidences show that some types of discourse are initiated by questions, while some others are sustained through questioning (Mishler, 1975).

2.2 Segmentation

Given a sequence of (written or spoken) words, the aim of topic segmentation is to find the boundaries where topics change. Segmentation algorithms use multiple sources of evidence for deciding segment boundaries. Often a set of potential boundaries is detected, and segment boundaries are picked from this set. Different approaches can be characterized by the types of evidence they use, the combination mechanism of multiple sources of evidence, and the means by which they detect segment boundaries out of a set of potential boundaries.

This section reviews approaches to segmentation, and includes the following subsections:

2.2.1 Features used in segmentation

2.2.2 General approaches

2.2.3 Machine learning in segmentation

2.2.4 Approaches using primarily word/concept distribution

2.2.5 Approaches combining multiple sources of evidence

2.2.6 Applying discourse structures to segmentation

2.2.1 Features Used in Segmentation

A number of features have been used in segmentation algorithms, including:

- Word/concept distribution features
- Discourse related features
- Multi-media features

Table 1 summarizes what features are used in text segmentation, used in speech segmentation with Automatic Speech Recognition (ASR), and used in video segmentation. An uppercase X indicates that a feature has been used and a lowercase x indicates that the feature may be used but not found in the literature reviewed.

Table 1: Features Used in Segmentation

Feature		**Text**	**Speech with ASR**	**Video**
Word/concept distribution	Word frequency	X	X	
	Lexical similarity score of vocabulary	X	X	
	Introduction of new vocabulary	X	X	
	Semantic relationships	X	x	
	Repetition of named entities	X	x	
	Cue-word / cue phrase	X	X	
Discourse	Pre-assigned potential boundary	X		
	Discourse structure	x	x	
	Cue-word / cue phrase	X	X	
	Pronoun usage	X	x	
	Referential noun phrase usage	X	x	
	Questions	x	x	
	Cohesion between different	x	x	

		units		
Multi-media features	Non-speech events	Pause	X	x
		Laughing	x	x
		Crying	x	x
	Speech prosody	Stress-pattern	X	
		Tune	X	
	Video cues	Facial expression		x
		Body movement		x

Word/concept distribution rely the content of text. Word/concept distribution features includes:

- Lexical repetition (Reynar, 1994)
- Lexical similarity across a potential boundary (Hearst, 1997; Kozima, 1993; Yaari, 1997)
- Introduction of new vocabulary (Bestgen, 2006; Choi, Wiemer-Hastings, & Moore, 2001; Franz, et. al, 2003)
- Semantic relationships (Bolshakov, & Gelbukh, 2001; Reynar, 1999).
- Cue-words/phrases fall between word/concept distribution and discourse-related features, depending on the type of cue-words an approach looks for. Often segmentation approaches do not distinguish between content-words and discourse markers used as cue-words (Franz et. al, 2003).
- Repetition of named entities (Reynar, 1999) is also good indicators of topic of where two sections are likely to be talking about the same topic.

Discourse features rely on the structure of the text and the relations among text units. Often they are used in combination with word/concept distribution features. Discourse features includes:

- Pre-assigned potential boundaries (Bolshakov, & Gelbukh, 2001; Hearst, 1997; Kozima, 1993; Reynar 1994) for example sentence boundaries are used as potential segment boundaries. Often the computed segment boundaries are adjusted to the nearest paragraph or section boundaries.
- Discourse structure (Grosz, & Sidner, 1986): Discourse structure is used, although not to an extensive level.
- Cue-word / cue phrase (Franz et. al, 2003): the presences of certain cue phrases or cue-words that tend to appear near the segment boundaries are useful. Such methods tend to be domain-specific because their dependence on the style of the text.
- Pronoun usage (Passoneau, & Litman, 1997; Reynar, 1999)

- Referential noun phrase usage (Passoneau & Litman, 1997)
- Question is a type of discourse, which this paper will be studying.
- Cohesion between different units.

Multi-media features are features associated with audio or video cues, which are often combined with word/concept distribution and discourse features in segmentation, including:

- Non-speech events such as duration of pauses (Dharanipragada et. al, 1999; Franz, et al, 2003) are also used in speech segmentation as an indicator of potential segment boundary.
- Speech prosodies such as speaker change and silences are already used in speech segmentation tasks. They are especially useful in multi-party conversations (Galley et. al, 2003). Furthermore, emphasis and intonation of speech, short vs. long phonemes may be very useful cues not only in classifying boundaries but also in detecting the focus of a topic (Tür, et. al, 2001). Rate of speech (Franz, et. al, 1999) is also useful in speech segmentation.
- Video cues such as body movement and facial expressions.

2.2.2 General Approaches

Often a segmentation program contains two major steps: detecting potential boundaries and selecting real boundaries from the potential boundary set.

Potential Boundary Candidate Detection

Detection of potential boundaries in text documents is relatively easy. Often text is pre-segmented by the author of the text. For some approaches (for examples, Bolshakov, & Gelbukh, 2001; Hearst, 1997; Kozima, 1993; Reynar 1994), this step is simplified. Sentence boundaries, paragraph boundaries, or section boundaries are used as potential segment boundaries.

For speech segmentation it is not as easy as for text segmentation. Since speech is not punctuated and well-organized as text, each point between two words could be a potential boundary position, thus the total number of “potential boundary positions” could be very large, making the effort of classifying them even harder. Researchers (Tür, et. al, 2001) have adopted speech prosody and cue words to perform this task.

Selecting Segment Boundaries

To select real segment boundaries from potential boundary candidates, a variety of approaches haven been taken. The program can work bottom-up, top-down, or sequentially through the text.

Clustering (Bestgen, 2006; Choi, Wiemer-Hastings, & Moore, 2001; Eichmann, et. al, 1999; Yaari, 1997) works bottom up to cluster together similar text units. It would usually contain the following steps:

- First, the text is partitioned into elementary units. An elementary unit could be a certain number of words, a sentence, or a paragraph.
- The most similar consecutive units are combined to a larger unit. This similarity may be defined by lexical cohesion, discourse cohesion or the combination of the two.
- Then, the program repeats until there is only one unit left.

Top-down methods break the entire document into parts. There are different ways of breaking down the document. Usually some weak link or valley values of some similarity/possibility under thresholds are identified. Reynar (1994) identifies the topic boundaries by plotting the distribution of word repetitions and put the boundaries at the least dense points of the graph. Hearst (1997) places topic boundaries at the locations of valleys in the similarity measure (Salton, & McGill, 1983), and are then adjusted with known boundaries (paragraphs, sections). Other researchers (Bolshakov, & Gelbukh, 2001; Hearst, 1997; Kozima, 1993; Lin, 2004) compute valley values of similarity/probability score or use statistical methods to identify the points that maximize the overall segmentation probability (Beeferman, 1999; Kehagias, et. al, 2004; Utiyama, & Isahara, 2001).

Boundary classification approaches (Galley et. al, 2003; Passonneau & Litman, 1997; Reynar 1999; Tür, et. al 2001) work sequentially from the beginning to the end of the text. Each potential boundary position is considered, and a classifier is used to decide whether it is a segment boundary or not. If the score of some features reaches some threshold, the position where it appears is considered a segment boundary.

Boundary classification approaches often use cue words/phrases. Grosz and Sidner (1986) found that cue phrases play an important role in signaling segment changes, where the word Okay indicates 93% of the segment change in their collection. Automatic computation of cue words has three steps:

- First, it computes word probability to appear in boundary position.
- Second, it selects words with the highest probability.
- Third, it removes non-cues. This step could be done with or without human supervision.

Possible Refinements

Sometimes, after the segment boundaries are selected from the set of potential boundary candidates, some refinement may be needed to reduce error. This could be seen as part of the previous step. For example, Franz and others (2003) used boundary classification to select topic boundaries based on some features (for example, speech pause and cue-words), and then removed non-topic ones based on other features (for example, similarity between vocabulary usage).

2.2.3 Machine Learning in Segmentation

Machine learning methods are used by segmentation approaches in learning both the features (Beeferman, 1999; Franz, et. al, 2003) and the rules of combining features (Kehagias, et al, 2004). For example, the CMU method (Beeferman, 1999) uses a statistical framework for feature selection. The program assigns a probability that there exists a boundary to the end of every sentence. The probability distribution is computed by building a log-linear model which weights a large set of features of the surrounding text. The features are automatically selected from lexical discourse cues and incorporated topical word usage into the model by building two statistical language models.

Machine learning methods are mostly data-driven, and are able to avoid dependence on manual specification of domain specific knowledge as in the other two types of approaches. However, there may be patterns that machine-learning methods cannot discover. One possible reason is data sparseness. Some patterns may not occur frequently enough in the training corpus for machine learning methods to learn the patterns.

To compare these approaches, content-based approaches are used from the very early stage of segmentation studies and lasted over time. They also provide basic features for other approaches. However, content-based approaches work relatively well in some domain, such as broadcast news, because the vocabulary change from story to story is often very significant (for example, from car accident to basket ball games). But for some other genre where the vocabulary is similar from topic to topic, content-based approaches alone are not good enough to distinguish the slight differences. Approaches that combine multiple sources of evidence are heavily used by researchers in this area and seem to have satisfactory performance. They often combine many features using a decision tree or a maximum entropy model. These features include not only lexical usage but also non-speech cues and speech prosody as well. The rules combining these features are manually specified based on the characteristics of the corpus and thus is domain or genre-dependent. Machine learning methods are relatively new and have its advantage over the other two types of approaches in terms of less human intervention and are usually not domain specific. But it faces other problems such as data sparseness.

2.2.4 Approaches Using Primarily Word/Concept Distribution

Word/Concept distribution features are used extensively in segmentation. Some approaches use only word/concept distribution features, while other approaches use primarily word/concept distribution features and combine these features with other features. These approaches rely on the differences of lexical usage on the two sides of a potential boundary. The larger the difference, the more indicative of a boundary would be.

Introduction of new vocabulary is adopted by many approaches, probably because lexical data is the easiest to quantify. Youmans (1991) used is called VMP (vocabulary management profile). VMP simply counts the number of new vocabulary terms introduced in an interval of text; once the number of new vocabulary exceeds some threshold, a topic changes is considered to occur, and a new topic is detected.

Similarity between chunks of words is also used. TextTiling segmentation system (Hearst, 1997) assigns a score to each potential boundary based on a cosine similarity measure (Salton, & McGill, 1983) between chunks of words appearing to the left and right of the potential boundary. Topic boundaries are placed at the locations of valleys in this measure, and are then adjusted to coincide with known paragraph boundaries.

Lexical repetition is largely used to locate topic boundaries in a stream of text. Kozima (1993) used mutual similarity of words in a sequence of text as an indicator of text structure. Reynar (1994) presented a method that finds topically similar regions in the text by graphically modeling the distribution of word repetitions and put the boundaries at the least dense points of the graph.

Additional lexical recourses are used. Bolshakov and Gelbukh (2001) used collocation of words and semantic links from a large database (CrossLexia system) to identify the cohesion boundaries. Methods that use additional knowledge allow for a solution to problems caused by the use of hyperonyms or synonyms. For example, a sentence belonging to a topic may not share common words with other sentences, but still are semantically related to taken as one topic unit.

2.2.5 Approaches Combining Multiple Sources of Evidence

Rather than relying solely on word/concept distribution, some approaches use multiple features: not only the words/concepts in the text, but also the relationships between words, sentences, and paragraphs (Dharanipragada, et. al, 1999; Eichmann, et. al, 1999; Galley et. al, 2003; Passonneau & Litman, 1997).

More often than not, a single feature alone may not be strong enough to support segmentation. Often multiple sources of evidence are used to reach a higher confidence level. To combine multiple sources of evidence, usually decision trees or other types heuristic rule-based methods are used (for example, Franz, et. al, 1999, 2003; Tür, et. al 2001).

Franz, and others (2003) explored automatic segmentation of the collection being used in this study. They combined word/concept distribution features (vocabulary similarity across a potential boundary), discourse features (cue-words/phrases), and multimedia features (duration of events marked as non-speech) using a decision tree. They used a two-step approach:

- **Step 1** hypothesizes boundaries at non-speech events: use a decision tree (binary) based probabilistic model to compute the probability of a boundary at every non-speech point in the ASR transcript. The interval peaks are compared with a threshold value to hypothesize document boundaries.
- **Step 2** is the refinement stage: to remove boundaries around which the stories are topically similar.

2.2.6 Applying Discourse Structures to Segmentation

Segmentation, as a problem dealing with topic, and discourse structure, of which topicality is an important or even a central point, are inherently related to each other. Researchers have been developing theories of discourse structure and applying them to the area of automatic segmentation of text.

Grosz and Sidner (1986) apply discourse structure to segmentation. They define discourse structure as having three parts: the structure of the sequence of utterances (called the linguistic structure), a structure of purpose (called the intentional structure), and the state of focus of attention (called the attentional state). They have made suggestions for automatic processing of discourse structures mentioned in their model, and their work has been very influential in segmentation literature.

Discourse markers (also referred as cue-words or cue phrases in the segmentation literature) have been used as evidence of segment boundaries (Dharanipragada et. al, 1999; Eichmann, et. al, 1999; Galley et. al, 2003; Lin 2004; Passonneau & Litman, 1997; Tür, et. al, 2001). However, they are often used together with other types of cue-words.

Discourse parsing intends to capture the structure of discourse. For example, Rhetorical Structure Theory is used to represent the tree structure of discourses (Carlson, Marcu, & Okurowski, 2003; Marcu, 2000). Discourse structure theories have been used in other text mining areas, such as summarization (Marcu, 1997; 2000) and discourse generation (Hovy, 1993). However, discourse structures, i.e., the global and local structure of the text, have not been investigated much in connection with segmentation. Segmentation programs use discourse indicators such as cue-words and speech prosody to identify segment boundaries, but often neglect the structure of the text.

In this paper, I am interested in some discourse features that might be able to contribute to segmentation of the oral history interviews. These features include: discourse markers, questions, and discourse structure of the personal experience narratives. The next section describes the study.

3. Methodology, Results, and Analysis

Section 3 describes a three-part study addressing the research questions. It has four subsections. Subsection 3.0 describes the collection used in this study. Subsection 3.1, 3.2, and 3.3 describes each part of the study and are organized by increasing complexity.

3.0 Data: The MALACH Collection

This paper used manual transcripts of some testimonies in this collection, but the ideas may also be applied to Automatic Speech Recognition transcripts.

3.0.1 ASR, Human indexed

MALACH (Multilingual Access to Large Spoken Archives) is a project aiming to facilitate storage and retrieval of multilingual spoken archives (Oard, et. al, 2004; Soergel & Oard, 2005). It contains testimonies of interviews with Holocaust survivors. Human indexers segmented the testimonies into several topical units, and indexed the interviews with index terms. Some testimonies are manual transcripts, in order to assist and evaluate automatic speech recognition. In this paper, manual transcripts were used for discourse analysis. There were 5 testimonies (with some tapes missing) in all.

3.0.2 Narrative and Conversational Text

The MALACH collection is composed of interviews with Holocaust survivors. This corpus consists of unconstrained and natural speech of people from different parts of the world, of different gender and age.

Interviews are mostly narratives. Their purpose is to let the survivors tell stories about their lives before, during and after the Holocaust. The interview contents are very homogenous in terms of vocabulary use. Thus the content-based approaches based on similarity across potential boundaries will not be as effective. Most of them are personal life experiences, thus the structure of PEN (Labov, & Waletzky, 1967; 1997) is very likely to apply. Time and location play important roles in forming the narrative. They are also person-oriented, thus character changes between topics might be useful in segmenting the narrative.

Interviews are also conversational by nature. The conversational nature makes the collection different from broadcast newswire in Topic Detection and Tracking (TDT) which are well-written articles spoken by radio or television reporters. However, it is also different from everyday conversations, which are turn-by-turn base. The interviewer usually controls the pace and the topic of the interview. Questions and change of speakers thus also play an important role in forming topic segments.

3.1 Discourse Markers

This section describes an experiment with discourse markers. The results show that discourse markers in this collection are not good indicators of segment boundaries. Readers may want to skip this section and go to section 3.2.

3.1.1 Methodology

Franz and others (2003) have examined cue-word position and frequency at segment boundaries of this corpus. But they looked at the correlation at the general level, regardless of the types of words. The cue-words in their experiment contain both content-words or discourse markers.

Our research question here is how much would discourse markers alone contribute to the segmentation of oral history interviews.

This study used a set of the discourse markers from Schiffrin (1987) and Taboada (2006), and first examined their frequency in the corpus. Markers that never occurred in this corpus were removed from the list.

The following discourse markers were examined:

Okay
Anyway
Um
Well
Still
Oh
As
So
Uh
When
Or
Also
Then
And
But
Because
If
Since
Otherwise
Though
However

The aim is to find out which discourse marker(s) is more likely to appear at boundaries. To measure the likelihood, relative frequency is used to measure.

RF = frequency around boundaries / overall frequency in the text.

10, 20, and 40 words "around the boundaries" are experimented.

3.1.2 Results and Analysis

A program counted the number of occurrence of the discourse markers listed in the above section. Table 2 shows the frequencies of the cue-words in the full-text and around boundaries.

Table 2: Frequencies of Cue-words

Cue Word	Absolute Frequency in Full-text	Absolute Frequency at Boundaries	Relative frequency around boundaries*
Okay	73	35	**0.48**
Anyway	22	3	0.14
Um	97	10	0.10

Cue Word	Absolute Frequency in Full-text	Absolute Frequency at Boundaries	Relative frequency around boundaries*
Well	107	9	0.08
Still	70	5	0.07
Oh	43	3	0.07
As	176	9	0.05
So	616	29	0.05
Uh	2678	124	0.05
When	263	12	0.05
Or	242	10	0.04
Also	79	3	0.04
Then	160	5	0.03
And	3204	94	0.03
But	360	10	0.03
because	275	4	0.01
If	120	1	0.01
Since	9	0	-
otherwise	7	0	-
though	7	0	-
however	3	0	-

*N=20, e.g. 10 words before boundary and 10 words after boundary. N=10 and N=40 gave similar results.

Among the 22 cue words, “okay” appears more often at the boundary in comparison with its frequency in the overall transcript. It seems that “okay” is a better predictor than “anyway” or other discourse markers listed. However, it is difficult to know from the above analysis how good or confident a discourse marker is.

3.2 Questions in Oral History Interviews as Boundary Indicators

3.2.1 Methodology

In order to understand how questions can be used as segment boundary or segment continuation indicators, some specific research questions are considered:

- What types of questions are asked during the interview?
- What content areas are asked about in the interview questions?
- Is there any relation between the type or content of questions and their positions in a segment?

A qualitative research methodology was used.

Data

I used 4 testimonies that are manual transcripts and are manually segmented by topic: testimonies no. 9, no. 17, no. 55, and no. 1124. Table 3 shows numbers of questions and segments in each testimony.

Table 3: Numbers of questions and segments

	Testimony no.	No. of questions	No. of segments
Training data	17	51	36
	55 (missing tape 1)	23	14
Test data	9 (missing tape 1)	40	36
	1124	46	26
Total		160	112

There were 160 questions in all testimonies asked by interviewers. The sample I analyzed contains 74 questions in two randomly selected testimonies (no. 17 and no. 55) asked by interviewer(s). I coded the 74 questions to extract patterns, and used these features in a question classifier to distinguish question boundary questions from segment continuation questions. I tested the program on the other two testimonies (no. 9 and no. 1124) with 86 questions.

Definition

In the transcripts of the interviews, three types of utterance are marked:

- Questions (marked with “Q”)
- Answers (marked as “A”)
- Operational speech for tape recording purposes, for example “this is roll two”, which are sometimes marked as “O” by human indexers.

This paper adapted the definition of question as verbalization that has the illocutionary force of a question without necessarily being phrased in an interrogative form (Krone, 1993), and considered a declarative statement that requests corroboration from the respondent, for example, “tell us about the trip to America”, as a question.

In some cases, interview questions are interrupted by the interviewee for further clarification and are marked as separate Qs. For example:

- **Q** yeah tell me how did it feel when you finally knew that you can come out and say you’re a Jew
- **A** to these people there
- **Q** in general

In this case, this paper considered the separated Qs as one question, which is “yeah tell me how did it feel when you finally knew that you can come out and say you’re a Jew in general.”

In some other cases, interview questions are interrupted by other events such as changing or testing tapes. For example:

- **Q** did your sister move in with you after your mother left
- **O** [UM] [pause]
- **O** [noise] I [unintelligible] okay I’m rolling
- **Q** after your mother left

Again, the separate Qs are considered as one question, which is “did your sister move in with you after your mother left?”

Some of the interviewer’s speech is not a question. Sometimes it is a repetition of what have been said by the interviewee, and sometimes it is a little reminder or hint from the interviewer to let the talk continue. For example:

- **Q** How many inmates were in [unintelligible]?
- **A** A few thousand and everybody was four or five
- **Q** infected infect-
- **A** with lice
- **Q** yeah
- **A** yes everybody…

In this case, only the first Q is considered a question.

Coding

Most of the interviews contain speech of an interviewer and an interviewee. The interviewer asks questions to let the interviewee talk about their personal experiences during the Holocaust period. The interviewee usually starts his/her talk by responding to the question and goes beyond that. Sometimes, the interviewer interrupted the interviewee and asks questions about specific events. The interviewee’s talk dominates the interview, and there is no significant signal or cue used for turn taking.

Sometimes, a testimony contains speech of multiple speakers. Usually there are family members of the primary interviewee. Speaker change is also a good indicator of segment boundaries (Galley et. al, 2003). However, this paper only deals with the speaker changes where there is a question.

I classified the questions in the interviews as “segment boundary questions” and “segment continuation questions” based on the segments boundaries. This is consistent with Mishler’s observation (1975) that there are types of questions that initiate a discourse and types of questions through which a discourse sustained. In addition, the position of questions in a segment was coded (beginning of a segment, in the middle of a segment, or toward the end of a segment). A number of segments contain only one question.

I coded both segment boundary questions and segment continuation questions against Graesser's (1992) taxonomy of questions and inquiries to see whether any patterns appear in terms of the relationship between a question's Graesser category and its presence/absence at segment boundaries. I analyzed the subject and content of the questions. In addition, I also looked at the length of questions and answers.

To ensure reliability of the coding, all questions were coded for each type of coding (position of questions, question types, purpose of questions, answer length) separately so that the coder worked with one set of categories at one time.

Other features, such as the presence of particular words or phrases, are also examined, for example, the word "then" in question "what happened then?" often results in a continuous complicating actions of an event. In question "what year was that", word "that" indicates that the time to appear in the answer is often associated a previous event and should be the same segment with the answer prior to this question.

3.2.2 Results and Analysis

Since the collection consists of interview-based conversations, the testimonies of interviews contain turns of conversation lead by questions that are posed by the interviewer. It is not surprising that some segments start with one of these questions. By analyzing these questions I found that half of the segments are bounded by questions (25 out of 50 in testimony no. 17 and 55).

However, not all questions start a new segment, so here two different types of questions are defined: segment-boundary questions and segment-continuation questions. As named, segment-boundary questions serve as triggers and boundaries to start new segments (topics). Segment-continuation questions serve as bridges to continue a segment across the interviewer and interviewee, and they usually ask the interviewee to "tell me more" about a topic.

Table 4 and 5 show numbers of questions and segments in training data and test data:

Table 4: Numbers of questions and boundaries in the training data

	Boundaries marked by question	Boundaries not marked by question	All boundaries
Segment-boundary questions	25	25	50
Segment-continuation questions	49	N/A	N/A
All questions	74	N/A	N/A

Table 5 Numbers of questions and boundaries in the test data

	Boundaries marked by question	Boundaries not marked by question	All boundaries
Segment-boundary questions	20	42	62
Segment-continuation questions	46	N/A	N/A
All questions	86	N/A	N/A

Questions were carefully examined. The type and content of the questions were examined in connection with their positions in a segment.

Types of Questions

Graesser's typology of questions was used. Table 6 explains Graesser's categories and illustrates the abstract specification with an example (if possible) from this study.

The categories, as noted by Graesser, are not mutually exclusive. In our study, a particular question can be assigned to more than one category. For example, the following questions can be assigned to multiple categories:

Example 1

Tell me how did it feel when you finally knew that you can come out and say you're a Jew in general while you were still in Poland

This question belongs to both directive and judgmental categories. In this sample of 74 questions, 7 questions were assigned with two categories.

Table 6: Graesser's Taxonomy of Questions

Question	Abstract Specification	Example
1. Verification	Is a fact true? Did an event occur?	Did you have any non-Jewish friends?
2. Comparison	How is X similar to Y? How is X different from Y?	
3. Disjunction	Is X or Y the case? Is X, Y, or Z the case?	Did you come from a religious or a secular home?
4. Concept completion	Who? What? When? Where? What is the referent of a noun argument slot?	What year was that?
5. Definition	What does X mean? What is the superordinate category and some properties of X?	
6. Example	What is an example of X? What is a particular instance of the category?	
7. Interpretation	How is a particular event interpreted or summarized? How is a pattern of information interpreted or summarized?	
8. Feature specification	What qualitative attributes does entity X have? What is the value of a qualitative variable?	When were you realizing the danger of this anti-Semitism?
9. Quantification	What is the value of a quantitative variable? How much? How many?	Alice how much did you know or were you aware of the Judenrat?

10. Causal antecedent	What caused some event to occur? What state or event causally led to an event or state?	What made you decide to leave Poland?
11. Causal consequence	What are the consequences of an event or state? What causally unfolds from an event or state?	
12. Goal orientation	What are the motives behind an agent’s action? What goals inspired an agent to perform an action?	
13. Enablement	What object or resource enables an agent to perform an action?	How were you fed?
14. Instrumental	How does an agent accomplish a goal? What instrument or body part is used when an agent performs an action?	
15. Procedural		
16. Expectational	Why did some expected event not occur?	
17. Judgmental	The questioner wants the answer to judge an idea or to give advice on what to do.	How did you feel about the fact that other people were starving?
18. Assertion	The speaker expresses that he or she is missing some information.	
19. Request	The speaker politely asks the listener to perform an action.	Alice would you introduce your family to us?
20. Directive	The speaker wants the listener to perform an action and is spoken more forcefully than a request.	Tell us about the trip to US.

The questions fall into ten categories in Graesser's taxonomy. Table 7 shows the categories and their percentage in comparing with reference interviews (White, 1998). Only bolded ones occur in this sample.

Table 7: Percentage of Question Types in Graesser's Categories

Type of question	**Oral history interviews (N=70)**	**Reference interviews (N=600)**
1. Verification	33	48
2. Comparison	0	7
3. Disjunction	1	5
4. Concept completion	32	5
5. Definition	0	2
6. Example	0	2
7. Interpretation	1	0
8. Feature specification	0	1
9. Quantification	4	0
10. Causal antecedent	1	0
11. Causal consequence	0	0
12. Goal orientation	0	2
13. Enablement	2	<1
14. Instrumental	0	<1
15. Procedural	0	<1
16. Expectational	0	1
17. Judgmental	12	12
18. Assertion	0	4
19. Request	7	10
20. Directive	5	<1

Verification questions account for the largest percentage (about one third) among all question types. Concept completion questions appear more often in this testimony then in reference interviews.

Table 8 shows the question types grouped by relative frequency.

Table 8: Question Types by Relative Frequency

Percentage Range	**Question Type**	**Percentage**
>30%	Verification	33%
	Concept Completion	32%
10-30%	Judgmental	12%
5-10%	Request	7%
	Directive	5%
<5%	Quantification	4%
	Enablement	2%
	Causal antecedent	1%
	Conjunctive	1%
	Interpretation	1%

Verification takes about 33% of all the questions. According to Graesser, verification questions normally elicit short answers. However, in this corpus it is not the case. Often the interviewee started by answering the question, and went on and on to try to explain more about the answer. For example:

Example 2

Q: were you interacting with your parents at this point?

A: yes [UH] my father died early he got the typhoid fever he was [UH] ve-very sick we couldn't get [UH] doctor or help in the house …

The interviewee then talked about her father's illness and death, her mother's hard work in the camps, lives of her sister and other family members, and finally how they all survived and was able to be with the families together.

Such verification questions elicit long answers because they are not used to seek for a specific answer as they appear. The interviewer is indeed lack of the knowledge, but she/he is not particularly interested in the yes/no answer of this question, but rather wants to elicit the interviewee's talk. For example, the interviewer does not intend to get a specific answer to the above question. Rather, the interviewer uses this question to give a starting point to let the interviewee talk about her parents.

Concept completion questions take about 32% of all questions. The predominance of concept completion questions is caused by the informal information exchange of the interviews. Usually the interviewer asks a question based on what has been said by the interviewee previously. For example, if an interviewee talks about an event, the interviewee would often ask, "what year was that?" because time was an very important dimension of history. Location and following events are also asked very frequently in concept completion questions.

Judgmental questions account for 12% of all questions. Interviewer asks the interviewee about his/her feelings and attitudes of a previous event.

Request and directive questions each account for about 5-7%. As example 1 shows, request and directive questions often belongs to multiple categories, because it is in the form of a request or directive question, but can ask for opinion (judgmental questions) or location (concept completion).

The following five categories occur less frequently (less than 5%) comparing with other categories:

- Enablement questions, ask how something was accomplished.
- Conjunctive questions, ask whether one situation was true over another.
- Interpretation, ask for explanation of something.
- Quantification, ask for the quantity of something.
- Causal antecedent, ask for the reason why some particular event happened.

Positions of Questions in a Segment

Of all the 74 questions coded, 25 questions are segment boundary questions, and 49 are segment continuation questions.

Figure 3 shows the number of different types of questions ordered by decreasing total frequency.

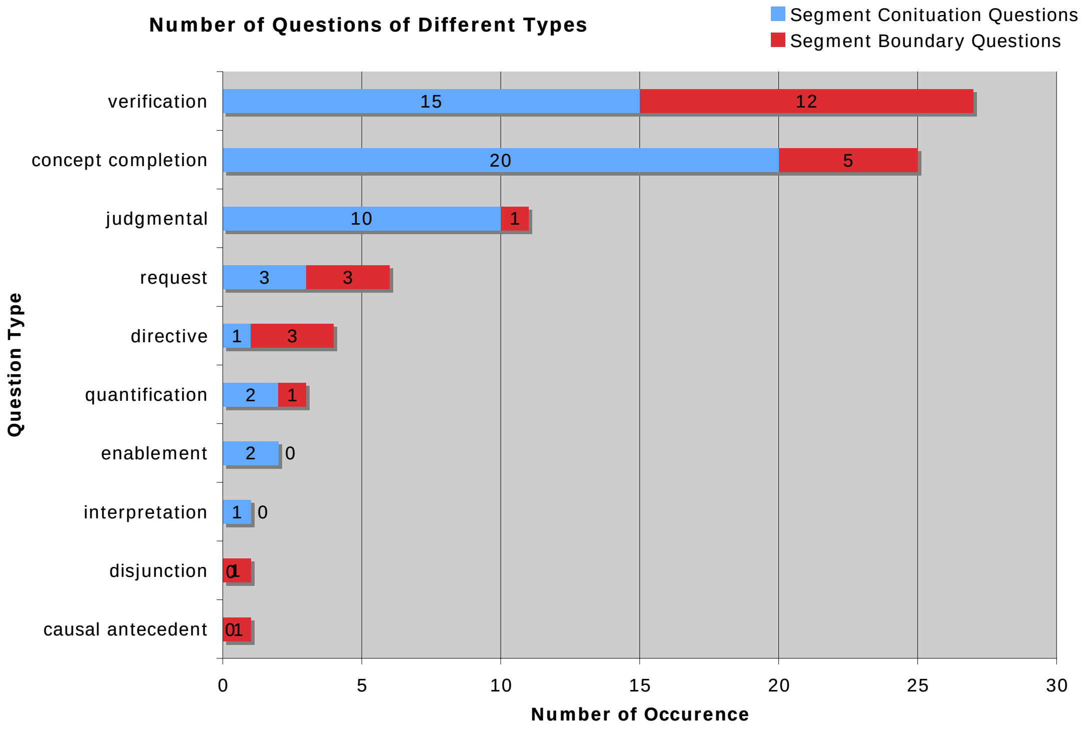

Figure 3: Number of Questions of Different Types

Concept completion questions are mostly segment continuation questions. However, four concept completion questions in the sample are segment boundary questions. They can be easily distinguished from other concept completion questions, because they always provide the context of the concept to be completed, for example:

- What happened **after your mother left?**
- When were **you realizing the danger of this anti-Semitism?**

The bold phrases provide the context of the concept. A segment continuation question would look like:

- What happened **then?**
- When was **that?**

Judgmental questions are mostly segment continuation questions, because they ask for the judgment of some previous mentioned event. Thus the event and its judgment have to belong to one segment.

Other types of questions are not frequent enough to be generalized. However, a reasonable assumption is that enablement and interpretation questions should all be segment continuation questions. Enablement questions ask about how something was accomplished, this “something” has to be mentioned previously, so they belong to one segment. Interpretation questions ask for further explanations of something, again, this “something” has to be mentioned previously.

Other types of questions are very difficult to distinguish from segment continuation questions to segment boundary questions. The content of questions has to be examined.

Types of Questions by Content

1. Segment-continuation Questions

I examined content of segment boundary questions and segment continuation questions separately, and found that the majority of segment continuation questions fall into the following categories and can be well distinguished from segment boundary questions.

1) Graesser’s concept completion questions (38%):

1a) Asking about addition information about an event (33%), e.g. time, location, or more details. For examples:

- What year was that?
- What was the town?

- You're gonna tell us a little bit about that

1b) Asking what happened next: (8%)

- What happened then?

2) Graesser's judgmental, interpretation and enablement questions (20%):

2a) Judgmental: Asking the interviewee to talk about feelings about an event, which is mentioned in the previous speech. For example:

- How did that whole thing feel?
- How did you feel when you found out what happened really to those people that were taken from the ghetto?

2b) Interpretation and enablement questions: Asking why something happened or how something was accomplished, and the thing is mentioned in the previous speech. For example:

- How were you fed?

3) Asking about something that the interviewee just talked about, usually asking the interviewee to explain or clarify a previous condition (30%). The forms of the questions vary largely, but they all contain some words/concepts mentioned by previous talk. For example,

- Did you have any personal contacts with the Judenrat?
  (The interviewee was just talking about the Judenrat in a few of sentences prior to that.)

2. Segment-boundary Questions

However, segment-boundary questions, which are 35% of all questions, vary a lot in format and content. For example, these are some sample segment-boundary questions:

- Did you come from a religious or a secular home?
- Did you have any non-Jewish friends?

These questions seem like a yes-no verification question, but they actually serve the role of a prompt. When the interviewer asks a question like these, they do not anticipate a yes-no answer. The interviewee starts from this point, and talks about the story.

Other segment-boundary questions are, for example:

- Alice, how much did you know or were you aware of the Judenrat and when were you realizing the danger of this anti-Semitism?

- What made you decide to leave Poland?
- Tell us (about) the trip to America.

There is no common pattern to these segment-boundary questions. However, since there are only two types of questions and segment-continuation questions are much easier for automatic programs to identify, computer programs can be used to identify segment-continuation questions and treat others as segment-boundary questions.

3.2.3 Automatic Question Classifier

Based on the previous analysis, a question classifier is designed to automatically classify the interview questions into two categories: segment continuation questions and segment boundary questions. The question classification program used the features identified in the above analysis. The algorithm of the program is described as below:

For each question to be classified (in the order of the program logic):

- If it is a judgmental, enablement or interpretation questions, it is a segment-continuation question.
- Else if it is concept completion question and it uses anaphors (“then”, “that”, and “this”), it is a segment-continuation question.
- Else if it uses pronouns (except for “you”, “your”, and “yours”)
- Else if the length of the question <= 4, it is a segment-continuation question.
- Other questions are classified as potential segment boundary questions.

The program is based only on the questions without any context of previous and following answers. The program was tested on two other testimonies (no. 9 and no. 1124, 86 questions in all). The result matrix is show as Table 9 below:

Table 9: Result Matrix

	Continuation	Boundary	Sum
Predict Continuation	40	5	45
Predict Boundary	26	15	41
Sum	66	20	86

For continuation questions, precision = 40/45= 89%; recall = 40/66 = 61%

For boundary questions, precision = 15/41 = 37%; recall = 15/20 = 75%.

15 predicted segment-boundary questions are real boundary questions, which compose about 24% of all segment boundaries (62 segments in the testing data). The program tends to recognize more segment-boundary questions than the real boundary questions. Segment-boundary questions can be used for preliminary segmentation of the text into relatively big chunks, and segment-continuation question can be used as negative evidence of the existence of a boundary. Additional features such as similarity across the potential boundary may be needed to reduce this error rate.

3.3 Discourse Structure of Personal Experience Narratives

3.3.1 Methodology

In order to address the question about the validity of the discourse structures proposed by Labov and Waletzky (1967, 1997) and Stein and Glenn (1979) for oral history interviews, and how could this knowledge be used in automatic segmentation, the following specific research questions was considered:

- How much do PEN and Story Grammar theory agree with each other in terms of labeling narrative clauses?

- How does the structure correspond to segment boundaries?

Data

I randomly picked 10 segments from the 112 segments in 4 testimonies. Half of them contain question(s) and others don't. The sample was not completely random, because I purposefully avoided segments toward the end of the testimonies, which often contain picture showing section or introduction of family members.

Definition

Because the manual transcript does not contain any punctuation, sometimes it is hard to tell the sentence boundaries. So we adapted the definition of clause from Labov and Waletzky (1967, 1997) to be used as the basic text unit to code.

Clause: a clause is a group of words consisting of a subject and a predicate, although, the subject can be implicitly given.

Element: a functional component of the PEN Structure (e.g. abstract) (Labov & Waletzky 1967, 1997) or Story Grammar (Stein & Glenn, 1979) is called an element. It can be composed by one or more clauses.

Coding

I coded each segment against the PEN Structure (Labov & Waletzky 1967, 1997), using the elements in PEN:

- Abstract

- Orientation
- Complication
- Evaluation
- Resolution
- Coda

I also coded each segment against Story Grammar (Stein & Glenn, 1979) elements:

- Setting
- Initiating Event
- Internal Response
- Attempt
- Consequence
- Reaction

I then compared the two coding system based on the coding coverage of the text, order of the elements.

According to Labov and Waletzky (1967), a clause in personal experience narratives can serve two functions, referential or evaluative. Referential clauses have to do with the factorial aspects: time, place, characters, and events. Evaluative clauses (and evaluative aspects of referential clauses) have to do with the attitudinal aspects: evaluative material states or highlights the points of the story.

I also coded the functions of clauses as referential or evaluative.

Again, each type of code was applied at a time so that the coder deals with only one set of categories to ensure the reliability and validity of the coding.

I then analyzed the effectiveness of PEN and Story Grammar, and suggested approaches to identify discourse structure and use it in segmentation.

3.2.2 Results and Analysis

8 segments of 4 testimonies were coded against the PEN Structure and Story Grammar as described in section 3.2.2. Table 10 shows the statistics of the 8 segments. Some segments contain questions from the interviewer, but only the talk of the interviewee was coded.

Table 10: Statistics of Sample Segments

Segment No.	Testimony No.	Segment No. in the Testimony	No. of Words	No. of Questions
1	9	56961	297	0
2		57055	449	1
3	17	62753	604	1
4		63056	312	2
5	55	194185	429	3
6		194385	907	0
7	1124	110650	566	1
8		110904	582	7

Table 11 and Table 12 show how the eight segments fit into the PEN Structure and Story Grammar Theories (see next page). Columns represent segments, and rows represent elements in the theories. A "x" means a clause (or clauses) in that segment was found as corresponding to the element at the head of each row.

Table 11: PEN Structure

PEN Element	1	2	3	4	5	6	7	8	Total Segments having this element
Abstract									0
Orientation	x	x	x	x	x	x	x	x	8
Complication1	x	x	x	x	x	x	x	x	8
Evaluation1		x	x	x	x				4
Complication2		x	x	x	x	x	x	x	7
Evaluation2		x	x	x	x	x			5
Complication3		x	x		x	x		x	5
Evaluation3		x	x		x				3
Complication4			x		x				2
Evaluation4					x				1
Resolution	x	x			x	x	x		5
Coda	x	x							2
Number of Un-coded elements	0	0	0	1	0	0	0	0	1
Order of elements	Y	Y	Y	Y	N	Y	Y	Y	7

Table 12: Story Grammar

Story Grammar	1	2	3	4	5	6	7	8	Total segments having this elements
Settings	x	x	x	x	x	x			6
Initiating event	x	x	x	x	x	x	x	x	8
Internal response	x		x	x	x	x	x	x	7
Attempt	x	x	x	x	x	x	x	x	8
Consequence	x	x	x		x	x	x	x	7
Reaction	x			x					2
Number of Un-coded elements	0	2	3	1	3	1	0	0	10
Order of elements	N	Y	Y	Y	N	Y	Y	Y	6

Coding Coverage

The distinction between internal response and attempt in Story Grammar is very vague in these segments. Usually it is very difficult to tell whether a segment is a response or attempt. They are more like sequenced actions as described in PEN, and often an action is followed by an evaluation.

PEN seems to capture the *complication – evaluation* pairs (which appear fairly frequent in the text) better than Story Grammar. The Story Grammar Theory has difficulty in coding some clauses because it failed to capture some evaluative or judgmental clauses following some actions. For the Holocaust survivors, they had too much stories and feelings to share, not only what happened but also their feelings and evaluation of events. The highest frequency of the *complication – evaluation* pairs appear is four in segment 5.

However, PEN also missed one clause, an evaluation clause that occurs at the end of a segment (in segment 4). This evaluation clause occurs after the resolution of the event(s), and is not evaluation of a particular action, but rather an evaluation of the entire event or topic. Story Grammar has a reaction at the end, which states the speaker's attitudes toward the entire event.

Abstract in PEN never occurs in these segments. In some segments, a question "when was that?" followed by an answer like "that was March nineteen forty seven" is often used to specify the time of an event told, and is often at the end of the segment. It can be seen as a coda, although it is not directly saying the story is over.

Comparing the two schemas, PEN Structure seems to fit better to the corpus. It missed 1 clause while Story Grammar missed 10.

The orientation/settings clauses are easy to identify. They often mention time, location, people and some kind of background information. The resolution is also easy to code, and it often contains cue-words such as “finally”.

Order of Elements

In segment sample 1, settings and initiating events sometimes are reversed or merged together. For example, the following start part of a segment:

> *Example 3*
>
> *“One day my mother decided that it's getting very bad and she made contact with a Polish policeman whom we knew an acquaintance and he was suppose to be on duty a certain day on the outside of the Ghetto now where there were buildings and not all the places were surrounded by barbed wire in some places buildings were the borders”*

“One day my mother decided that it's getting very bad and she made contact with a Polish policeman” is actually an initiating event, and the clauses following it is some kind of background information which could be considered as settings. For these clauses, PEN Structure is better because it combined the two as orientation, which could include both background information and initiating events.

Some background information may be inserted in the middle of a segment. For example, in segment sample no.5, the interviewee gave the background information in the middle of the discourse. Both theories have settings/orientations at the beginning.

Functions of the Elements

Referential clauses took about 75% of all the narratives. Evaluative clauses took about 25% of all the narrative text.

Suggestions for Automatic Analysis of Discourse Structure

As the results show, PEN is a better representation of the structure of narrative texts. It could be a strong evidence of where segments start and ends. However, it is very challenging for a program to automatically identify all types of different elements in this structure.

Although it is difficult to label all elements correctly, the starting part and ending part of the structure are relatively easy to identify. This paper suggests a simplified approach to identify and use PEN structure to segmentation:

Identify starting element of a segment: orientation of a segment usually mentions time, location, and characters. If these items are mentioned within a relatively short text range (within a few clauses), it is very likely that it is an orientation of an experience and would probably start a new segment. Each of the items could be assigned a weight comparing it with a previous item of the same time, for example, a higher weight would be given to a time instance that is ten years later than its previous time instance than a time instance that is one day after its previous time instance. A program could be given some training data and learns from that what are the scoring rules for combining time, location and characters, and those clauses that higher than a threshold would be labeled as orientations.

Identify ending elements of a segment: resolutions, which do not occur as frequent as orientation in most of the segments, however, have very lexical evidence such as “finally”, “at last”, and “at the end” in the clauses. Computer programs can also identify the lexical features easily. Coda, as the optional but absolute ending of a narrative, does not occur very frequently. However, whenever a coda like “that’s the story” appears, the program could know for sure that a segment is ended.

The evidence of starting and ending elements can be combined with other features such as questions, lexical similarity and speech prosody to get a better result.

4. Conclusions, Suggestions and Discussions

4.1 Conclusions

This paper reviews the segmentation approaches and identifies certain trends emerging from the literature:

Word/concept distribution features are very important to segmentation. Not only that some early works rely only on content-based features such as word repetition or lexical similarity across a potential boundary, but also that other types of features (discourse, speech prosody, noon-speech events, and video cues) have to work together with at least some content features to come up with segment boundaries.

Segmentation is a multi-step process using evidence from multiple features. More and more approaches try to integrate multiple sources of evidence. Multiple sources of evidence are used to determine whether a given potential boundary candidate is a segment boundary. Usually this process is done in multiple steps, and in each step certain type(s) of evidence may be used. Sometimes, some source of evidence may be used to detect potential boundary candidates, and then other source(s) of evidence may be applied to determine the probability whether this candidate is a real segment boundary. A third source of evidence may be used for refinements such as reduction of error rates.

Discourse related features are under-investigation by segmentation researchers. Discourse, considered as closely related to structure of text, is not studied as extensively as other features. The mostly studied features relating to discourse is cue-words and phrases, which indicate discourse structure to a certain extend. However, several studies using cue-words or cue phrases do not distinguish between discourse markers and other types of words also appearing frequently around segment boundaries. How much do discourse markers contribute to segmentation remains a question. Discourse structure, which was used by other text mining areas such as automatic summarization, is understudied in this field.

This paper also conducts a study trying to examine the following discourse related features as sources of evidence for segmentation of oral history interviews:

- Discourse markers;
- Interview questions;
- Discourse structure of Personal Experience Narratives.

In terms of these three features, this study found the following for the oral history interview collection:

Discourse markers are not very useful to identify segment boundaries. Single-word discourse markers do not seem to contribute to segmentation of this corpus.

Interview questions can be used as segment boundary indicators but do not identify all segments. Two categories of question – segment continuation questions and segment boundary questions, seem to distinguish from each other based on types of questions and content of questions. Thus computer programs may be able to identify the distinction and classify the interview questions into two categories, which can be used as positive evidence (segment boundary questions) to support the existence of a boundary or as negative evidence (segment continuation questions) to reduce the probability of a boundary. However, since the number of interview questions is not large enough, and question distribution is not even through a testimony, other sources of evidence are needed for more fine-grained segments.

Discourse structure of Personal Experience Narratives applies well to oral history interviews, and may be a good indicator of segments. The results show that randomly selected sample segments tend to conform to the PEN Structure. Approaches to automatic identification of part of the structure (starting and ending elements) are suggested.

4.2 Suggestions of an Approach Combining Multiple Sources of Evidence

Based on the findings from this study, here I suggest a multi-step approach combining multiple sources of evidence for segmentation of oral history interviews. This approach contains the following steps, and features that are used in each step are specified under the step.

Potential boundary candidate detection is omitted and every space between two words is considered as a possible boundary.

Step 1: Classify interview questions into segment-boundary questions and segment continuation questions. Use segment boundary questions as to segment the text into big chunks.

Features used:

- Discourse feature: question content and question types; use of pronouns and other anaphors in questions

Step 2: Use discourse structure to identify orientation and resolution clauses. Find potential boundaries before an orientation and after a resolution or coda.

Features used:

- Word distribution: A dictionary to identify time, location, people
- Discourse features: Cue-words such as “finally” “at last” to identify resolution clauses

Step 3: If a potential boundary corresponds to a boundary question (close enough, value of distance considered as “close” to be learned from training data), the boundary is confirmed. Otherwise, compute similarity score (could be introduction of new noun phrases) across a potential boundary within a window.

Step 4: Use segment-continuation questions as negative evidence of existence of a boundary, and remove boundaries that are close enough to a segment-continuation question.

4.3 Discussions

4.3.1 Limitation of the Study

Reliability of coding: In this qualitative research, a lot of coding is conducted. Interview questions (section 3.2) are coded against Graesser’s categories of questions, and segments (section 3.3) are coded against the PEN Structure and the Story Grammar Theory. However, only one coder (author of this paper) participated in coding. Although some mechanism was used (for example, dealing with one set of codes at a time), there might still be reliability problem with coding.

Use of manual transcript: this study used manual transcripts instead of Automatic Speech Recognition (ASR) transcripts. Questions are marked as Q in the manual transcripts. In ASR transcripts, only speaker change could be marked. If work with ASR transcripts, an additional step of marking the questions may be needed before coding and analyzing the questions. Furthermore, ASR transcripts contain recognition errors, which may influence the performance of the question classifier, and the recognition of starting and ending elements in PEN structure.

This paper does not consider evaluation of segmentation, which is an important and yet difficult issue with topic segmentation.

4.3.2 Suggestions for Future Work

Discourse Parser of Personal Experience Narratives: In the paper, one of the findings is that discourse structure of oral history interviews seems to consistently respond to the PEN structure. Researchers have developed discourse parsers for Rhetorical Structure Theory (RST) (Marcu, 1998). However, it works better with discourse structure that are more like a tree structure.

For narrative text, a discourse parser that can identify different elements of an event or story would be very useful for many other tasks such as segmentation, summarization, and information extraction.

Inter-rater Agreement of Segmentation: This paper does not talk about evaluation of segmentation. However, evaluation is a very tough problem for many tasks. One of the reasons is that determining where topic boundaries belong is a subjective task (Passoneau, & Litman, 1993; 1997), and even human judges do not agree among themselves where are the topic boundaries. However, human judges all come up with some “reasonable” boundaries. Furthermore, a human judge may not be consistent when segmenting a document.

One research topic could be to look at the human indexed topic boundaries and examine the rationale for assigning a boundary and to compare them across different judges and across different text segments of a same judge.

Segmentation of unstructured or semi-structured text: So far most of the segmentation systems deal with structured text, for example news articles. However, less attention has been paid to unstructured or semi-structured corpus such as email threads or casual conversations.

It would be interesting to know how much researchers have learned with segmentation of structured text/speech would be useful in unstructured or semi-structured corpus and what are the new challenges are.

Using questions in segmentation of multi-party conversations: Segmentation of multi-party conversations often used combined word/concept distribution features with non-speech features such as pause, speaker change, and overlaps (Galley et. al, 2003). Since question and answers to questions take a large portion of the verbal communication in small group discussions (Pavitt &, Curtis, 1990), questions may be very useful in segmentation of multi-party conversations. A study on questions in multi-party conversations similar to the study described in this paper will probably answer the question.

Moreover, in a multi-party conversation, for example in small group discussions, and especially when multiple topics are going on at the same time, it is very difficult to identify what questions a statement is replying to. To identify question-answer pairs would help to both topic detection and segmentation asks.

Non-contiguous segments: Sometimes, a topic segment may be interrupted by other topics in conversations, and come back again later. This non-contiguous segment brings difficulties to topic segmentation. One may argue that if a topic is separated by another topic, the separated two parts are two segments of the same topic. In that case, this problem is a problem of topic detection – to identify two or more separated segments belongs to the same topic. In each view, the problem of identifying two pieces of text or speech stream belongs together topically is a research topic that needs further investigation.

Suggestions for research on questions: the following research issues relates to questions, not segmentation. This paper found that questions, as a part of discourse structures can be used together within other features to segment speech or text. Questions may also be used in other text-mining tasks, such as summarization and topic detection. Here I suggest some research topics relates to questions and text mining.

Automatic identification of questions: Before questions can be used in automatic segmentation or other tasks, they have to be identified. Automatic identification of questions is a useful pre-processing step for many other tasks. It includes determining whether a statement is a question or not. For some formal text, this is not a problem, because questions are marked with question marks and are using certain words in some fixed way. However, this could be very difficult for conversational types of corpus, because questions can be very diverse and irregular in forms. A question could be in the exact order as a statement. Word/concept distribution alone would not be enough to determine a question; speech prosody such as tunes would be a very important feature to include.

Question classification based on their form and content: Type of question is a very important feature, and is broadly examined in several areas, such as question answering systems and reference interviews. People have different criteria to classify questions based on their forms and content. An automatic question classifier that could classify questions into pre-defined categories would be useful for many tasks, especially for research involving question analysis.

Question Clustering: For a corpus that contains a large number of questions, to cluster questions into topically or functionally related clusters would be useful for retrieval or segmentation of such corpus.

Automatic Question Generation: Questions are found useful in human-computer dialogs. For example, casual conversations are used in recommendation systems (Ginty, & Smyth, 2002) would be able to simulate the shopping experiences with a human seller. A research topic could be to generate interview questions for this oral history project. It would be interesting to see whether a computer system could generate interview questions for letting people talk about their personal experiences.

5. Acknowledgment

The author would like to thank Dr. Dagobert Soergel for his advice on this research. Part of the analysis is built on Graesser's work, and the author would like to thank Dr. Marylin White for suggesting the references.

References

Bawa, M., Manku, G. S., Raghavan, P., (2003). *SETS: Search enhanced by topic segmentation, in the Proceedings of the 26th annual international ACM SIGIR conference on Research and development in information retrieval*, pp. 306-313

Beeferman, D., Berger, A., & Lafferty, J. (1999). Statistical models for text segmentation. *Machine Learning, 34(1–3),* pp.177–210

Bestgen, Y., (2006). Improving text segmentation using latent semantic analysis: a reanalysis of Choi, Wiemer-Hastings and Moore, *Computational linguistics, 32(1)*, pp. 5-12

Blakemore, D., (1992). Understanding Utterances: An Introduction to Pragmatics. Blackwell, Oxford.

Blakemore, D., (2002). Relevance and Linguistic Meaning: The Semantics and Pragmatics of Discourse Markers. Cambridge University Press, Cambridge

Bolshakov, I. A., & Gelbukh, A. F., (2001). Text segmentation into paragraphs based on local text cohesion. In the Proceedings of the 4th International Conference on Text, Speech and Dialogue, pp. 158 - 166

Carlson, L., Marcu, D., Okurowski, M. E., (2003). Building a Discourse-Tagged Corpus in the Framework of Rhetorical Structure Theory. In Current Directions in Discourse and Dialogue, pp. 85-112, Jan van Kuppevelt and Ronnie Smith eds., Kluwer Academic Publishers.

Chafe, W. L., (1976). Givenness, contrastiveness, definiteness, subjects, topics, and point of view. In Li, C. N. (ed.), Subject and Topic, pp.25-55, New York: Academic Press.

Choi, F. Y. Y., (2000). Advances in domain independent linear text segmentation. In *Proceedings of the North American Chapter of the Association for Computational Linguistics,* pp26-33, Seattle, WA, USA.

Choi, F. Y. Y., Wiemer-Hastings, P., & Moore, J., (2001). Latent semantic analysis for text segmentation. In *Proceedings of EMNLP 2001.* Seattle, WA, USA.

Chua, T.-S., Chang, S.-F., Chaisorn, L., & Hsu, W., (2004). Story boundary detection in large broadcast news video archives ñ techniques, experiences and trends. In *Proceedings of the 12th annual ACM international conference on Multimedia*, pp.656-659, New York, NY, USA

Dharanipragada, S., Franz, M., McCarley, J. S., Roukos, S., & Ward, T., (1999). Story segmentation and topic detection in the broadcast news domain. In *Proceedings of the DARPA Broadcast News Workshop*

Eichmann, D., Ruiz, M. E., Srinivasan, P., Street, N., Culy, C., & Menczer, F., (1999). Cluster-Based Filtering for broadcast news. In *Proceedings of the DARPA Broadcast News Workshop*

Franz, M., McCarley, J. S., Ward, T., Zhu, W.-J., (1999). Segmentation and detection at IBM: models and two-tiered clustering. *TDT-1999*

Franz, M., Ramabhadran, B., Ward, T., & Picheny, M., (2003). Automated transcription and topic segmentation of large spoken archives. In *Proceedings of Eurospeech,* Geneva, Switzerland, September 2003, pp.953-956.

Galley, M., McKeown, K., Fosler-Lussier, E., Jing, H., (2003). Discourse segmentation of multi-party conversation, in *Proceedings of the 41st Annual Meeting on ACL*

Graesser, A.C., Pearson, N., & Huber, J., (1992). Mechanisms that generate questions. In Questions and Information Systems, edited by A. C. Graesser, N. Pearson and J. Huber. Hillsdale, NJ: Erlbaum

Grosz, B. J., Sidner D. J., (1986). Attention, intentions, and the structure of discourse. *Computational Linguistics, 12(3),* pp.175-204

Hearst, M. A., Plaut, C., (1993). Subtopic Structuring for Full-Length Document Access. In *Proceedings of the Sixteenth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval,* pp.59-68, Pittsburgh, USA

Hearst, M. A., (1997). TextTiling: segmentation text into multi-paragraph sub-topic passages. *Computational Linguistics, 23 (1),* pp.33-64.

Hovy, E. H., (1993). Automated discourse generation using discourse structure relations. *Artificial Intelligence Journal, 63,* pp.341--385

Kehagias, A., Nicolaou, A., Fragkou, P., & Petridis, V., (2004). Text segmentation by Product Partition Models and dynamic programming. Mathematical and Computer Modelling, 39, pp.209-217

Kim, J. W., Candan, K. S., Dönderler, M. E., (2005). Topic Segmentation of Message Hierarchies for indexing and Navigation Support, in the *Proceedings of the 14th international conference on World Wide Web,* pp.322-331

Kozima, H., & Furugori, T., (1993). Similarity between words computed by spreading activation on an English dictionary. In *Proceedings of the 6th Conference of the European Chapter of the Association for Computational Linguistics (EACL-93),* Utrecht, pp.232-239.

Kozima, H., (1993). Text Segmentation Based on Similarity between Words, in *Proceedings of the 31st annual meeting on Association for Computational Linguistics*, pp 286-288, Columbus, Ohio

Krone, K., (1993). A review and assessment of communication research, *Progress in Communication Sciences, 11,* 179-206

Labov, W., (1997). Some further steps in narrative analysis. *The Journal of Narrative and Life History.* Available online at http://www.ling.upenn.edu/~labov/sfs.html

Labov, W., & Waletzky, J., (1967). Narrative analysis: Oral versions of personal experience. in J. Helm (Ed.), Essays on the verbal and visual arts. Seattle, WA: University of Washington Press. pp.12-44

Litman, D. J., & Passonneau, R. J., (1995). Combining multiple knowledge sources for discourse segmentation. In *Proceedings of the 33rd conference on Association for Computational Linguistics,* pp.108-115, June 26-30, 1995, Cambridge, Massachusetts

Louwerse, M. M., & Mitchell, H. H., (2003). Toward a taxonomy of a set of discourse markers in dialogue: A theoretical and computational linguistic account. Discourse Processes, 35 (3), 243-281

Mann, W. C. & Thompson, S. A., (1987). Rhetorical structure theory: A theory of text organisation. *Technical report ISI/RS,* pp. 87-190

Mann, W. C., Matthiessen, C. M. I. M., Thompson, S. A., (1992). Rhetorical Structure Theory and text analysis. In Mann, W. C., and Thompson, S. A. eds.: Discourse Description: diverse linguistic analysis of a fund-raising text. Amsterdam/Philadelphia: John Benjamins Publishing Company

Marcu, D., (1997). From discourse structures to text summaries. In Mani, I.; Maybury, M. eds.: Intelligent Scalable Text Summarization, pp.82-88

Marcu, D., (2000). The theory and practice of discourse parsing and summarization. Cambridge, MA: The MIT Press

Mishler, E. G., (1975). Studies in dialogue and discourse: II. Types of discourse initiated by and sustained through questioning. *Journal of Psycholinguistic Research, 4(2),* pp.99-121

Oard, D. W., Soergel, D., Doermann, D., Huang, X., Murray, G. C., Wang, J., Ramabhadran, B., Franz, M., & Gustman, S., (2004), Building an information retrieval test collection for spontaneous conversational speech. In *Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval,* pp.41-48, Sheffield, United Kingdom

Soergel, D., & Oard, D. W., (2005). The CLEF 2005 Cross-language speech retrieval text collection introduction.

Passonneau, R. J., Litman, D. J., (1997). Discourse segmentation by human and automated means, *Computational Linguistics, 23(1),* pp.103-139

Pavitt, C., & Curtis, E. (1994). Small group discussion: A theoretical approach (2nd ed.). Scottsdale, AZ: Gorsuch Scarisbrick.

Reynar, J. C., (1994). An automatic method of finding topic boundaries. In *Proceedings of the 32nd annual meeting on Association for Computational Linguistics (ACL94)*

Reynar, J. C., (1999). Statistical models for topic segmentation. In *Proceedings of 37th Annual Meeting of the Association for Computational Linguistics (ACL'99),* pp.357-364

Sacks, H. 1967-1972. Unpublished lecture notes. University of California. As in Stenström, A.-B., 1988, Questioning in Conversation.

Salton, G., McGill, M. G., (1983). Introduction to modern information retrieval. New York: McGraw Hill.

Schiffrin, D. (1987). Discourse markers. Studies in Interactional Sociolinguistics 5, Cambridge: Cambridge University Press.

Schiffrin, D. (2001). Discourse markers: Language, meaning and context. In: Schiffrin, D., Tannen, D., & Hamilton, H. E. (Eds.), The Handbook of Discourse Analysis, Blackwell, Malden, MA, pp. 54-75 Stenström, A-B., (1988). Questioning in conversation. In M. Meyer, ed., Questions and questioning, pp.304-325. Berlin: Walter de Gruyter.

Stein, N. L & Glenn, C. G., (1979). An analysis of story comprehension in elementary school children. In R. O. Freedle (Ed.), New directions in discourse processing (Vol.2)

Stolcke, A., Shriberg, E., T¸r, G., & Hakkani-Tür, D. Z., (1999). Combining Words and Speech Prosody for Automatic Topic Segmentation. In *Proceedings of DARPA Broadcast News Transcription and Understanding Workshop.*

Stokes, N., Carthy, J., & Smeaton, A. F., (2004). SeLeCT: a lexical cohesion based story segmentation system. *Journal of AI Communications, 17(1),* pp.3-12.

Taboada, M. T., (2004). Building Coherence and Cohesion, John Benjamins Pub Co

Taboada, M. T., (2006). Discourse markers as signals (or not) of rhetorical relations, to appear in *Journal of Pragmatics, 2006*

Tür, G., Hakkani-Tür, D., Stolcke, A., & Shriberg, E., (2001). Integrating prosodic and lexical cues for automatic topic segmentation. *Computational Linguistics, 27(1),* pp. 31-57

Utiyama, M., & Isahara, H. (2001). A statistic model for domain-independent text segmentation. In *Proceedings of the 9th Conference of the European Chapter of the Association for Computational Linguistics*, pp.491-498.

van Kuppevelt, J. (1995). Discourse structure, topicality and questioning. *Journal of Linguistics, 31, pp109*-147.

White, M. D., (1998). Questions in reference interviews. *Journal of Documentation, 54,* 443-465.

Yaari, Y., (1997). Segmentation of Expository Text by Hierarchical Agglomerative Clustering. *Recent Advances in NLP,* 1997, 59-65, Bulgaria.

Youmans, G., (1991). A new tool for discourse analysis: The vocabulary-management profile. *Language, 67,* 763--789.

Appendix A: Question Categories

Question	Graesser's Type(s)		Continuation / Boundary
and then what happened	concept completion		C
and what happened then	concept completion		C
and what happened then	concept completion		C
and what happened then	concept completion		C
any special message for your own grandchildren	concept completion		C
how did they react to a new Jewish baby	concept completion		C
how long did you stay in Montreal	concept completion	quantification	C
how long were you at Pruskoff	concept completion	quantification	C
I wanted to ask you before you went into the ghetto and the German Jews were coming and Czechoslovakian Jews were coming what was the attitude of the community the Jewish community towards these people	concept completion		C
what Henry's family	concept completion		C
what is your message to future generations	concept completion		C
what was the town	concept completion		C
what year was that	concept completion		C
when was that	concept completion		C
where was her family during the war	concept completion		C
which which organization was	concept completion		C
which year was that	concept completion		C
why was that important to you	interpretation		C
tell me about meeting your wife	directive		C
how did it happen that you moved outside	enablement		C
how were you fed	enablement		C
how do you think that your whole experience having survived and seeing what your family has gone through has it impacted your life	judgemental		C
when you look back to your childhood George and the experience that that that you had how do you what	judgemental		C

Question	Graesser’s Type(s)		Continuation / Boundary
what type of feelings do you have when you think back about your childhood during the war before the war			
how did the whole thing feel	judgmental		C
how did you feel about the fact that other people were starving	judgmental		C
how did you feel when you found out right after the war what happened to these people in the camps	judgmental		C
how did you feel when you found out what happened really to those people that were taken from the ghetto to camps	judgmental		C
how do you feel about the fact that she wants you to have this tape	judgmental		C
how do you feel about this	judgmental		C
how do you feel that your father's tape had never been made	judgmental		C
will you react to your grandma's comments	judgmental		C
talk a little bit about the reunions with these Jews after (unintelligible)	Request	Concept completion	C
you want to talk a little bit about your pregnancy	Request	Concept completion	C
you're gonna tell us a little bit about that	Request	Concept completion	C
did you experience any anti-Semitism after the war	Verification		C
did you go back to Poland	verification		C
did you have any personal contacts with the Judenrat	verification		C
did you have found any of the immediate family	verification		C
did you know about the holocaust as growing up	verification		C
did you make any contacts in hiding with non-Jews	verification		C
did your family ever try to leave Hungary	Verification		C
did your sister move in with you after your mother left	verification		C
do you have some pictures or things that you'd like to share with us	Verification		C
have you know about your past your brother's birth	verification		C

Question	**Graesser’s Type(s)**		**Continuation / Boundary**
so you went through school in Budapest	Verification		C
so you went to New Jersey	Verification		C
was there animosity	verification		C
were you punished because you skipped	Verification		C
you were eighteen at this	Verification		C
what made you decide to leave Poland	causal antecedent		B
I'm sorry after the liberation, what did your family do	concept completion		B
what happened after your mother left	concept completion		B
what happened during the occupation when hitler invaded	concept completion		B
when were you realizing the danger of this anti-Semitism	concept completion		B
did you come from a religious or a secular home	disjunction		B
Alice how much did you know or were you aware of the Judenrat	quantification		B
Alice would you introduce your family to us	Request		B
before we take a look at your pictures George would you like to introduce your wife to us	Request		B
tell me how did it feel when you finally knew that you can come out and say you're a Jew in general while you were still in Poland	directive	judgmental	B
tell us about the trip to US	directive		B
who's trying to tell us about the story of your uncle	directive		B
would you tell me where were you born and what was your childhood like?	Request	Concept completion	B
did anybody in your family become a smuggler	verification		B
did you ever try to leave yourself	Verification		B
did you have any contact with the underground	verification		B
did you have any non-Jewish friends	verification		B
did you move into the ghetto together	verification		B
did your parents have any plans to go into hiding or to leave the country	verification		B
do you still have relatives that are living in Hungary	Verification		B

Question	Graesser's Type(s)	Continuation / Boundary
so you were sick and you came home and people suspected you	verification	B
was it difficult to escape	Verification	B
were you interacting with your parents at this point	verification	B
were you thinking of escape	verification	B
you said that your family before the war was pretty observant and after the war did your father ever give you any explanation about why your family didn't observe	Verification	B

Appendix B: Question-classifier Program (written in Java)

```
import java.io.*;
import java.util.HashSet;
public class SegType {
	/**
	 * @param args
	 */
	public static void main(String[] args) {
		int[] matrix = new int[4];
		if (args.length != 1) {
			System.out.println("Usage: "+args[0] + " inputfile");
			return;
		}
		for (int i=0; i<matrix.length; i++)
			matrix[i] = 0;
		FileReader fr;
		System.out.println("Predict\tActual\tLine");
		try {
			fr = new FileReader(args[0]);
			BufferedReader reader = new BufferedReader(fr);
			String line;
			while (true) {
				line = reader.readLine();
				if (line == null) break;
				String[] tmp = line.split("\t");
				String label = tmp[0];
				line = tmp[1];
				if (!label.equals("C") && !label.equals("S")) {
				System.out.println("Unkonw type label for line:"+line);
				continue;
				}
				String[] parts = line.trim().toLowerCase().split(" ");
				HashSet set = new HashSet();
				for (int i=0; i<parts.length; i++) {
					set.add(parts[i]);
				}
				String result = null;
				//Judgemental Questions
				if ( (set.contains("how") && set.contains("do") &&
set.contains("feel"))
				|| (set.contains("how") && set.contains("does") &&
set.contains("feel"))
				|| (set.contains("how") && set.contains("did") &&
set.contains("feel"))
				|| (set.contains("comments") && set.contains("comment"))
						|| set.contains("judgement"))
				result = "C";
				//Enablement/Interpretation questions
				else if ( (set.contains("why") || set.contains("how"))
				&& (line.indexOf("how many")==-1)
				&& (line.indexOf("how much")==-1)
				&& (line.indexOf("how long")==-1)
				&& (line.indexOf("how often")==-1))
				result = "C";
```

```
                                        //Concept continuation question
                                        else if (
                                        (set.contains("when") || (line.indexOf("what year") != -1) ||
set.contains("where") || (line.indexOf("what happened") != -1))
                                        && (set.contains("then") || set.contains("this") ||
set.contains("that")))
                                        result = "C";
                                        //use of pronounce
                                        else if (set.contains("he") || set.contains("his") ||
set.contains("him") || set.contains("she") || set.contains("her") || set.contains("them") ||
set.contains("they") || set.contains("their"))
                                        result = "C";
                                        else if (set.size() <= 4)
                                                  result = "C";
                                        else
                                                  result = "S";
                                        //print results
                                        System.out.println(result+"\t"+label+"\t"+line);
                                        if (result.equals("C") && label.equals("C"))
                                                  matrix[0]++;
                                        if (result.equals("C") && label.equals("S"))
                                                  matrix[1]++;
                                        if (result.equals("S") && label.equals("C"))
                                                  matrix[2]++;
                                        if (result.equals("S") && label.equals("S"))
                                                  matrix[3]++;
                              }
                              System.out.println("\n\nConfusion Matrix:");
                              System.out.println("\t\tC\tS\tSum\t      <--(Real value)");
                              System.out.println("Predict
C\t"+matrix[0]+"\t"+matrix[1]+"\t"+(matrix[0]+matrix[1]));
                              System.out.println("Predict
S\t"+matrix[2]+"\t"+matrix[3]+"\t"+(matrix[2]+matrix[3]));
                              System.out.println("
Sum\t"+(matrix[0]+matrix[2])+"\t"+(matrix[1]+matrix[3]));
                    } catch (Exception e) {
                              // TODO Auto-generated catch block
                              e.printStackTrace();
                    }
          }
}
```