THE 1999 TOPIC DETECTION AND TRACKING (TDT3) TASK DEFINITION AND EVALUATION PLAN

1. INTRODUCTION

The purpose of the TDT3 project is to advance the state of the art in Topic Detection and Tracking. The task domain is multilingual human language. This domain is to be explored and technology is to be developed in the context of an evaluation-guided R&D paradigm in which key technical challenges are defined and supported by common informative evaluations. This document defines the tasks, the performance metrics, and the evaluation procedures to be used to direct the research and to evaluate research progress and technical capabilities.

TDT3 addresses multiple sources of information, including both text and speech. These sources are namely newswires, radio and television news broadcast programs, and WWW sources. The source languages are English and Mandarin. The information flowing from each source is modeled as a sequence of stories. These stories provide information on many topics. The general technical challenge is to identify and to follow the topics being discussed in these stories.

2. TOPIC DEFINITION

In the initial TDT study, conducted during 1996 and 1997, the notion of a topic was limited to be an "event", meaning something that happens at some specific time and place. For example, the eruption of Mount Pinatubo on June 15th, 1991 is considered to be an event, whereas volcanic eruption in general is not. Events might be unexpected, such as an airplane crash, or expected, such as a political election.

In the second TDT project, TDT2, the definition of a topic was broadened to include, in addition to the triggering event, other events and activities that are directly related to it. This definition will be retained in TDT3. For the purposes of TDT3:

A topic is defined to be a seminal event or activity, along with all directly related events and activities.

A story will be considered to be "on topic" whenever it discusses events and activities that are *directly* connected to that topic's seminal event. So, for example, a story on the search for survivors of an airplane crash, or on the funeral of the crash victims, will be considered to be a story on the crash event. Obviously there must be limits to this inclusiveness. (For example, stories on FAA repair directives that derive from a crash investigation would not be considered to be stories on the crash event.) Topic boundaries are subject to interpretation, so the Linguistic Data Consortium (LDC) has created guidelines to improve agreement and consistency of topic labeling.[1]

[1] Determining the limits of TDT topics is often extremely difficult and arbitrary. The question is where to draw the line on including (or excluding) "related" events. The LDC have facilitated this task considerably by identifying certain general types of topics and creating specific boundary determination rules for each of those topic types. These rules are contained in a web-accessible document that LDC uses to instruct and guide annotation. This document may be found at the following URL: http://morph.ldc.upenn.edu/TDT/Guide/manual.front.html.

3. THE CORPUS

LDC[2] is providing two corpora to support TDT3 research. These are namely a new version of the existing TDT2 corpus that has been augmented with Mandarin language sources, and a new TDT3 corpus, which is to be used for evaluation. These corpora are collections of news, including both text and speech, from a number of sources in both English and Mandarin.

The TDT2 corpus has been augmented to support TDT3 research. The original TDT2 corpus spans the first six months of 1998 and was annotated for 100 topics. This corpus originally served to support both development and evaluation. (Jan-Feb was used for training, Mar-Apr for development testing, and May-Jun for evaluation.) The TDT2 corpus will now be devoted to research and algorithm development and will be supplemented with additional Mandarin sources for the same time period. The stories in the Mandarin sources will be annotated for a subset of 20 of the original 100 topics. For purposes of TDT3 development, the TDT2 corpus will not be officially divided into separate training and development test sets. All of the data in TDT2 may be used in both training and development test.

A separate TDT3 corpus will be supplied for evaluation. This corpus will span Oct-Dec 1998 and will include the TDT2 sources. The corpus will be annotated for 60 topics. Each of these topics will span both English and Mandarin sources. Annotation of the TDT3 corpus will be performed as for the TDT2 corpus. Each story in the corpus will be tagged according to whether it discusses each of the defined topics. These story-topic tags (Tag[story,topic]) assume a value of YES if the story discusses the target topic, or BRIEF if that discussion comprises less than 10% of the story. Otherwise the (default) tag value is NO. A complete transcription of audio source data will also be provided. These transcriptions will include non-news material in addition to news stories. (Non-news stories include commercials and list-type reports such as sports scores and financial data.) Accordingly, stories will be labeled either as NEWS or MISCELLANEOUS or, in case there exists no transcription for a story, as UNTRANSCRIBED. This story type label will be used in TDT3 task evaluation, with TDT3 tasks being scored only on NEWS stories.

The LDC provides multiple representations of the source data for the TDT2 and TDT3 corpora. For audio sources, there are three different representations:

1. the audio sampled data signal
2. a manual transcription of the audio signal
3. a transcription produced automatically by an automatic speech recognition (ASR) system.

For Mandarin sources, there are two different text representations:

[2] The Linguistic Data Consortium Phone: 215/898-0464
Email: ldc@ldc.upenn.edu Fax: 215/573-2175
URL: http://www.ldc.upenn.edu/

1. the original Chinese character source stream, which is either source text, or a manual or ASR transcription of speech.[3]
2. An English translation produced automatically from the original Chinese character source stream.

For complete details on the TDT2 and TDT3 corpora, refer to URL http://www.ldc.upenn.edu/TDT/.

4. THE TASKS

There are five TDT3 tasks. They are namely the segmentation of a news source into stories, the tracking of known topics, the detection of unknown topics, the detection of initial stories on unknown topics, and the detection of pairs of stories on the same topic.

4.1 THE STORY SEGMENTATION TASK

The TDT3 story segmentation task remains unchanged from TDT2. It is defined to be the task of segmenting the stream of data from a source into its constituent stories. Since original text sources are supplied in segmented form, this task applies only to audio sources. Segmentation of audio signals may be performed directly on the audio signal itself or on the various textual transcriptions of the audio signal. Segmentation is to be performed only in the original (untranslated) language of the source.

4.2 THE TOPIC TRACKING TASK

The TDT3 topic tracking task has been changed only slightly from TDT2. It is defined to be the task of associating incoming stories with topics that are known to the system. A topic is "known" by its association with stories that discuss it. Thus each target topic is defined by one or more stories that are on (i.e., that discuss) the topic. To support this task, a small set of on-topic training stories is identified for each topic to be tracked. The system may train on the target topic by using all of the stories in the corpus, up through the most recent training story.[4] The tracking task is then to correctly classify all subsequent stories as to whether or not they discuss the target topic.

4.3 THE TOPIC DETECTION TASK

The TDT3 topic detection task remains unchanged from TDT2. It is defined to be the task of detecting and tracking topics not previously known to the system. It is characterized by a lack of knowledge of the topic to be detected. Therefore the system must embody an understanding of what a topic is, and this understanding must be *independent of topic specifics*. In the topic detection task, the system must detect new topics as the incoming stories are processed. The system must then proceed to associate input stories with those topics. Thus this process identifies a set of topics, as defined by their association with the stories that discuss them.

[3] The ASR transcription includes whitespace delimitation of words. Original source text and manual transcriptions do not.

[4] In TDT2, a list of stories certified as off-topic was provided as part of the topic training information. No such list of off-topic stories is provided in TDT3. In place of this list of stories known to be off-topic, the training portion of the corpus may only be presumed to be largely off-topic.

4.4 THE FIRST-STORY DETECTION TASK

The TDT3 first-story detection task is a revival of the on-line detection task of the TDT pilot study. It is defined to be the task of detecting, in a chronologically ordered stream of stories from multiple sources (and in multiple languages), the first story that discusses an event. This task may be viewed as essentially the same as the topic detection task. The only difference is in what the detection system outputs.

4.5 THE LINK DETECTION TASK

The TDT3 link detection task is a new TDT task. It is defined to be the task of determining whether two stories discuss the same topic. Thus, just as in the topic detection task, the system must embody an understanding of what a topic is, and this understanding must be *independent of topic specifics*. The link detection task, however, does not deal with topics explicitly. Thus links are not constrained to segregate stories into a set of orthogonal topics, and the assumption that each story discusses one and only one topic is avoided.

5. THE EVALUATION

In order to inform TDT research, to guide TDT technology development, and to assess TDT application potential, TDT task performance will be evaluated according to a set of rules for each of the five TDT tasks. All of the TDT tasks are cast as detection tasks. Detection performance is characterized in terms of the probability of miss and false alarm errors (P_{Miss} and P_{FA}). These error probabilities are then combined into a single detection cost, C_{Det}, by assigning costs to miss and false alarm errors:

$$C_{Det} \quad = \quad C_{Miss} \cdot P_{Miss} \cdot P_{target} \quad + \quad C_{FA} \cdot P_{FA} \cdot P_{non-target}$$

where

- C_{Miss} and CFA are the costs of a Miss and a False Alarm, respectively,
- PMiss and PFA are the conditional probabilities of a Miss and a False Alarm, respectively, and
- Ptarget and $P_{non-target}$ are the *a priori* target probabilities ($P_{non-target} = 1 - P_{target}$).

C_{Det} is the bottom-line representation of TDT task performance that is used to judge TDT systems. Unfortunately, the value of C_{Det} is also a function of application parameters. Specifically, C_{Det} is a function of the costs of detection errors and *a priori* target probabilities. Because of this, and in order to provide a more intuitively meaningful measure of system performance, C_{Det} will be normalized so that $(C_{Det})_{Norm}$ can be no less than one without extracting information from the source data. This is done as follows:

$$(C_{Det})_{Norm} \quad = \quad C_{Det} \ / \ \min\left(C_{Miss} \cdot P_{target}, C_{FA} \cdot P_{non-target}\right)$$

Thus the absolute value of $(C_{Det})_{Norm}$ is a direct measure of the value (i.e., relative cost) of the TDT system.

There are two reasonable methods of estimate detection error probabilities, called story-weighted and topic-weighted. The story-weighted method assigns equal weight to each decision for each story and accumulates errors over all topics. The topic-weighted method accumulates errors separately for each topic and then averages the error probabilities over topics, with equal weight assigned to each topic. **The topic-weighted method will be used exclusively in TDT3**, because detection error rates exhibit high variance among topics and because the number of on-topic stories also is highly variable for different topics.

Input data to the TDT systems comprises the various news sources. This source stream is presented to the processing systems in chronological order, and the various sources are presented together, interwoven so as to preserve chronological order. Source file sequencing will be controlled by means of a list of chronologically ordered source file names. Each source file will contain an uninterrupted sample of source data. It will be assumed that there is no temporal overlap between different source files.[5] TDT3 tasks require processing of all data, including both news and non-news stories. The computation of task performance, however, will exclude system output on non-news stories.

For all five TDT tasks, allowable information for the system to use will include knowledge of the source of the data and knowledge of the time of the stories. Specific evaluation rules for the five tasks now follow.

5.1 Story Segmentation Evaluation

Segmentation algorithms will be evaluated in terms of their ability to locate correctly the boundaries between stories. Only audio sources will be evaluated. Each file of source data is to be processed separately. Three choices for the type of source data to be processed are shown in Table 1.

Table 1. **Input data alternatives for the story segmentation task (audio sources only)**

Data Type
manual transcription
ASR transcription *(required)*
Original audio data

The test set over which segmentation is to be performed will include stories from all sources, both English and Mandarin. Thus evaluation will indicate the segmentation ability of a system as a function of the source language.

Segmentation must be performed as the data is being processed, but a certain amount of look-ahead will be allowed before boundaries must be output. Deferral of output is allowed because look-ahead presumably will improve segmentation performance. The deferral period, N_d, is a primary task parameter. For source data in text form, N_d is the maximum deferral period in words (for English) or characters (for Mandarin). For source data in audio form, N_d is the maximum deferral period in seconds. The deferral parameter values to be used in TDT3 are shown in Table 2. Decisions may not be deferred beyond the end of the current source file. (Deferral beyond the current source file is unnecessary because every source file begins and ends on a story boundary.)

Table 2. **Maximum deferral periods for the story segmentation task.**

Text		Audio
English (words)	**Mandarin** (characters)	**English & Mandarin** (seconds)
100	150	30
1,000	1,500	300
10,000 *(required)*	**15,000** *(required)*	3,000

Each research site is encouraged to study as many of these conditions as may be productively done. However, all sites that perform the segmentation evaluation are required to perform at least one evaluation under common shared conditions for the story segmentation task. These conditions are given in Table 3.

Table 3. **Required Story Segmentation Conditions**

Parameter	Value
Source language	Both English and Mandarin
Source data type	ASR transcription
Maximum decision deferral period	10,000 words for English 15,000 characters for Mandarin

Segmentation performance will be judged by determining how well computed story boundaries agree with reference boundaries.[6] This agreement will be judged with an evaluation interval that is swept through the input data. The evaluation interval is chosen to be long enough to include all computed boundaries that might reasonably be associated with a true reference boundary, but short enough to exclude unreasonable associations and multiple reference boundaries (i.e., whole stories). Evaluation is performed by sweeping the evaluation interval through the input source stream and judging the correctness of the segmentation at each position of the interval:

- If there is both a computed boundary and a reference boundary within the interval, then segmentation is judged correct.
- Likewise, if there is neither a computed nor a reference boundary within the interval, then segmentation is judged correct.
- However, a miss is declared if there is no computed boundary within an interval that contains a reference boundary,
- And a false alarm is declared when a computed boundary exists within an interval that doesn't contain a reference boundary.

[5] It is certain that data in different source files will overlap occasionally. The assumption of no overlap is made, however, because of the great simplification provided. The loss of strict temporal order is judged minor and insignificant, the time duration represented by each source file is a small fraction of a single day.

[6] The method used here derives from a method developed by Lafferty, et al. ("Text Segmentation Using Exponential Models", by Doug Beeferman, Adam Berger, and John Lafferty).

The duration of the evaluation interval depends on the type of input data being processed. The value to be used for TDT3 evaluation is shown in Table 4.

Table 4. **Evaluation interval for story segmentation evaluation**

Data Type	Interval length
Transcription (Mandarin)	75 characters
Transcription (English)	50 words
original audio data	15 seconds

Segmentation performance will be measured using the detection cost formula to combine P_{Miss} and P_{FA}. The evaluation cost parameters to be used for TDT3 evaluation are given in Table 5. Choice of P_{target} was based on an analysis of the stories in the TDT2 training corpus. (Note that P_{target} is also related to the length of the evaluation interval.)

Table 5. **Story segmentation evaluation cost parameters**

Parameter	Value
P_{target}	0.3
C_{Miss}	1.0
C_{FA}	0.3

5.2 TOPIC TRACKING EVALUATION

Tracking algorithms will be evaluated in terms of their ability to detect which stories are on-topic and which are not. Topics are to be tracked individually, and each topic is to be treated separately and independently. In training the system for a particular target topic, allowable information includes the training set and topic tags *for that target topic only*. During the evaluation of each target topic, no information is given on any other topic. Evaluation will be over the whole 3-month extent of the TDT3 corpus.

For each topic tracked, the primary system output will be a DECISION (YES/NO) for each story processed, indicating whether the story is judged to be on-topic or not. In addition to this primary output, the system must also produce a SCORE that indicates the confidence that the story is on-topic for the target topic. This SCORE may then be used to explore the trade-off between miss and false alarm errors.

There will be three alternative choices for the form of the source data to be processed. These are shown in Table 6.

Table 6. **Input data alternatives for the topic tracking task**

Source Data Types
text sources and manual transcription of the audio sources
text sources and ASR transcription of the audio sources *(required)*
text sources and the sampled data signal for audio sources

A primary task parameter is N_t, the number of stories used to define the target topic. The maximum value of N_t, $(N_t)_{Max}$, will be 4. All of these on-topic training stories are tagged YES for the target topic. The evaluation corpus will be divided into a different training set and test set for each target topic. The training set will comprise the first part of the corpus, up through the last training story tagged YES for the target topic. In addition, the training data will include story identification of the N_t stories that serve to define the target topic. The test set will comprise the remainder of the corpus that follows. Note that the target topic is defined using the *last* N_t on-topic training stories for the target topic. Note also that topic tags are supplied only for the N_t on-topic training stories and that there may well be on-topic stories in the training set that are not identified (i.e., that are outside the set of N_t on-topic training stories).[7]

Another important issue is which language(s) are to be used for topic training. For the TDT3 corpus, topic training may be performed on English sources only, on Mandarin sources only, or on both. Evaluation will be supported for three values of N_t for each of these three language conditions. The nine possible combinations are shown in Table 7. Note that the single-language training conditions prohibit knowledge of which cross-language stories are on-topic. Use of the cross-language portion of training corpus *without* this knowledge *is* allowed, however. Independent of the language training condition, the *test* set for each topic will include *all* sources for *all* languages. Evaluation of topic tracking performance will be conditioned on source language. This conditional analysis of tracking output will provide performance results as a function of language and will measure performance separately for cross-language and same-language training/test conditions.

Table 7. **Language training conditions for topic tracking evaluation**

	English	**Mandarin**	**Both Sources[9]**
N_t English (E) Mandarin (M)	1 (E)	1 (M)	1(E), 1(M)
	2 (E)	2 (M)	2(E), 2(M)
	4 (E) *(required)*	4 (M)	4(E), 4(M)

[7] This is different from the TDT2 tracking task definition, in which all training stories supplied to the system were identified as being either on-topic or not.

[9] The topic tracking test set begins after the last training story. Note that for mixed-language training, in which the number of training stories is controlled separately for different languages, it is quite possible that more than N_t training stories are encountered for one of the languages prior to the beginning of the test set. In this case, these on-topic stories are *not* labeled as such, despite their being in the training data.

The evaluation of topic tracking performance will support two story boundary conditions, as shown in Table 8. When story boundaries are not provided, tracking output must also identify that portion of the input source data to which the DECISION and SCORE apply. This identification is in terms of word/character indices (for text type sources) or time in seconds for sampled data signals. No restriction is imposed on the temporal quantization of such outputs. (E.g., outputs need not be synchronous with automatically determined story boundaries.) This is handled in evaluation by averaging the tracking output (DECISION and SCORE) over the true extent of the stories and then using these averaged values for evaluation.

Table 8. **Story boundary conditions for topic tracking evaluation**

Story boundary conditions
Reference story boundaries provided *(required)*
No story boundaries provided

The topic tracking system must process the input source data in chronological order, and tracking output must be made as each story is processed. Unsupervised adaptation is allowed, using information obtained as the input data are processed. No look-ahead is allowed, however. When story boundaries are not provided, tracking output must be completed by the end of the current source file.

Each research site is encouraged to study as many of the above conditions as may be productively done. However, all sites that perform the tracking evaluation are required to perform at least one evaluation under common shared conditions for the tracking task. These conditions are shown in Table 9.

Table 9. **Required Topic Tracking Conditions**

Parameter	Value
Topic training language	English
On-topic training stories	4
Source language	All
Source data type	text sources and the TDT3 automatic transcription of audio sources
Story boundaries	given

Topic tracking performance will be measured using the detection cost formula to combine P_{Miss} and P_{FA}. The evaluation cost parameters to be used for TDT3 evaluation are given in Table 10. Choice of P_{target} was based on an analysis of the stories in the TDT2 training corpus.

Table 10. **Topic tracking evaluation cost parameters**

Parameter	Value
P_{target}	0.02
C_{Miss}	1.0
C_{FA}	0.1

5.3 TOPIC DETECTION EVALUATION

Topic detection algorithms define topics by story affiliation. That is, each topic is defined as a set of stories. Thus topic detection algorithms will be evaluated in terms of their ability to cluster together same-topic stories. Evaluation will be over the whole 3-month extent of the TDT3 corpus.

The primary detection system output will be a TOPIC_ID for each story. This index will uniquely identify the topic that the story discusses. In addition to this primary output, the system must also produce a SCORE that indicates the system's confidence that the subject story is in fact on-topic for the identified topic. This score may then be used to explore the trade-off between miss and false alarm errors.

There will be three alternative choices for the form of the source data to be processed. These are the same alternatives as for topic tracking and are shown in Table 6.

Another important issue is whether the different languages are to be processed together or separately. For the TDT3 evaluation, three different conditions will be supported, as shown in Table 11.

Table 11. **Language conditions for topic detection evaluation**

Language Conditions
English only
Mandarin only
English and Mandarin together *(required)*

Topic detection may be performed both with and without knowledge of story boundaries, as shown in Table 8. When story boundaries are not provided, detection output must also identify that portion of the input source data to which the TOPIC_ID and SCORE apply. This identification is in terms of word/character indices (for text type sources) or time in seconds for sampled data signals. No restriction is imposed on the temporal quantization of such outputs. (E.g., outputs need not be synchronous with automatically determined story boundaries.) Story assignment is handled in evaluation by accumulating votes for all TOPIC_ID's over the true extent of each story (in terms of # of words/characters/seconds, as appropriate) and then assigning the dominant TOPIC_ID to the reference story. In case of ties, the voting will be retabulated using score-weighted votes.

The topic detection system must process the input source data in chronological order. However, the topic detection system may defer its identification of topics (and its association of stories with them) until a limited amount of subsequent source data is processed. This deferral period, N_f, is a primary task parameter and is the number of source files, including the source file being processed, for which processing may be completed before committing to and outputting story-topic associations. (The greater the deferral, presumably the better will be decisions regarding both the detection and identification of new topics.) The deferral parameter values to be used in TDT3 are shown in Table 12.

Table 12. **Maximum decision deferral periods for topic detection evaluation in number of source files.**

# of source files, including the one being processed
1
10 *(required)*
100

Each research site is encouraged to study as many of the above conditions as may be productively done. However, all sites that perform the topic detection evaluation are required to perform at least one evaluation under common shared conditions for the topic detection task. These conditions are given in Table 13.

Table 13. **Required Topic Detection Conditions**

Parameter	Value
Source Language	English and Mandarin together
Source data type	text sources and the TDT3 automatic transcription of audio sources
Maximum decision deferral period	10 source files
Story boundaries	given

Topic detection performance will be evaluated by measuring detection performance separately for each topic, in terms of miss and false alarm errors. Performance will be evaluated only on the set of predefined topics for which the TDT3 corpus has been annotated. (These are the only topics for which "truth" is known.) In order to tabulate detection performance, the TDT3 reference topics must be mapped to the appropriate system output topics to which they correspond. This mapping will be done automatically by choosing for each reference topic the system output topic that it matches best. The best matching system output topic is defined to be that which produces the lowest evaluation cost. This mapping process will result in a favorable and unavoidable bias in the results, because the mapping yields best (lowest cost) performance by construction, whereas some other mapping might be deemed more appropriate by other (subjective) criteria.

Evaluation of topic detection performance will be conditioned on source language. This conditional analysis of detection performance will thus provide performance results as a function of language. Language-conditioned analysis will be performed *after* topic mapping, however. Thus, since the reference topic is mapped only to one (the "best") system output topic, high detection performance will require that system output topics include topic stories from all languages.

Topic detection performance will be measured using the detection cost formula to combine P_{Miss} and P_{FA}. The detection cost parameters to be used for TDT3 evaluation are given in Table 14. Choice of P_{target} was based on an analysis of the stories in the TDT2 training corpus.

Table 14. **Topic detection evaluation cost parameters**

Parameter	Value
P_{target}	0.02
C_{Miss}	1.0
C_{FA}	0.1

It is assumed that each story discusses at most one topic. Stories that are tagged as discussing more than one of the target topics will not be used in computing topic detection performance.[10]

5.4 First-Story Detection Evaluation

The first story detection task is logically identical to the topic detection task.[11] The evaluation, however, focuses on the specific aspect of detection associated with novel information. To aid in first-story detection research, the TDT3 corpus will be augmented with first-story annotation for an additional 120 topics. Evaluation will be over the whole 3-month extent of the TDT3 corpus.

The primary first-story detection system output will be a DECISION (YES/NO) for each story processed, indicating whether the story is judged to be the first story of a topic or not. In addition to this primary output, the system must also produce a SCORE that indicates the confidence that the story is a first story. This SCORE may then be used to explore the trade-off between miss and false alarm errors.

The conditions for performing first-story detection are the same as for topic detection, except that the source language is limited to English. There are the same three alternative choices for the form of the source data, as shown in Table 6. There are the same story boundary conditions, as shown in Table 8. And there are the same decision deferral conditions, as shown in Table 12.

Each research site is encouraged to study as many of the above conditions as may be productively done. However, all sites that perform the first-story detection evaluation are required to perform at least one evaluation under common shared conditions for the task. These conditions are given in Table 15.

[10] The assumption that each story discusses only one topic is reasonable for the large majority of stories and greatly simplifies the definition of the detection task and its evaluation. Some stories will likely be tagged as on-topic for more than one of the target topics, however. These stories will be excluded in computing topic detection performance. Such stories will be retained as part of the story stream, however, and will not be eliminated from the corpus during topic detection processing.

[11] Knowing when a story is the first story on a topic implies knowing when a story is *not* a first story.

Table 15. **Required First-Story Detection Conditions**

Parameter	Values
Source Language	English only
Source data type	text sources and the TDT3 automatic transcription of audio sources
Maximum decision deferral period	10 source files
Story boundaries	given

First-story detection performance will be measured using the detection cost formula to combine P_{Miss} and P_{FA}. The evaluation cost parameters to be used for TDT3 evaluation are given in Table 16. Choice of P_{target} was based on an analysis of the stories in the TDT2 training corpus.

Table 16. **First-story detection evaluation cost parameters**

Parameter	Value
P_{target}	0.02
C_{Miss}	1.0
C_{FA}	0.1

5.5 LINK DETECTION EVALUATION

Link detection algorithms will be evaluated in terms of their ability to determine whether specified pairs of stories discuss the same topic (i.e., are "linked"). Evaluation will be over the whole 3-month extent of the TDT3 corpus.

The primary link detection system output will be a DECISION (YES/NO) for each pair of stories processed, indicating whether the stories are judged to be on the same topic or not. In addition to this primary output, the system must also produce a SCORE that indicates the system's confidence that the story-pair is linked. This SCORE may then be used to explore the trade-off between miss and false alarm errors.

There will be three alternative choices for the form of the source data to be processed. These are the same alternatives as for topic tracking and are shown in Table 6.

Another important issue is the choice of languages to be processed. For the new link detection task, TDT3 will support only English as a source language.

Table 17. **Language conditions for link detection evaluation**

Language Conditions
English only *(required)*

Link detection may be performed both with and without knowledge of story boundaries. For TDT3, however, evaluation will be supported only for systems that use the given reference story boundaries and produce output with respect to them.

The link detection system must process the input source data in chronological order. However, the link detection system may defer its identification of story links until a limited amount of subsequent source data is processed. This deferral period, N_f, is a primary task parameter and is the number of source files, including the source file being processed, for which processing may be completed before making story-story link decisions. (The greater the deferral, presumably the better will be the link decisions.) The deferral parameter values to be used in TDT3 are shown in Table 12.

Each research site is encouraged to study as many of the above conditions as may be productively done. However, all sites that perform the link detection evaluation are required to perform at least one evaluation under common shared conditions for the link detection task. These conditions are shown in Table 18.

Table 18. **Required Link Detection Conditions**

Parameter	Value
Source Language	English only
Source data type	text sources and the TDT3 automatic transcription of audio sources
Maximum decision deferral period	10 source files
Story boundaries	given

While link detection systems are expected to be capable of making link decisions for all pairs of stories, evaluation of all such decisions is not practical. Therefore evaluation will be limited to a subset of story pairs. This will keep system output files to a manageable level. System output must be limited to the specified story pairs. The story pairs for which output is required will be listed in chronological order, ordered primarily according to the second (i.e., the most recent or newest) story of the pair and secondarily according to the first (i.e., the oldest) story of the pair). This will facilitate coordination of system output with chronological processing and deferral requirements. The maximum deferral period is with respect to the source file containing the most recent of the pair of stories.

Link detection performance will be measured using the detection cost formula to combine P_{Miss} and P_{FA}. For the link detection task, P_{target} is the probability that a pair of stories chosen at random discuss the same topic. The evaluation cost parameters to be used for TDT3 evaluation are given in Table 19.

Table 19. **Link detection evaluation cost parameters**

Parameter	Value
P_{target}	0.02
C_{Miss}	1.0
C_{FA}	0.1

A IMPLEMENTATION DETAILS

A.1 SOURCE DATA FORMATS

Source data is organized into a sequence of files in chronological order. Each source file contains continuous news content from a single source. For newswire sources, the average length of a file is 20 stories. Newswire source files contain text data in the form of an untagged text stream. For audio sources, the file duration is one program, typically 30 or 60 minutes. Audio source files contain either sampled data that represents the audio waveform[12], or text data in the form of an untagged text stream.

There are four different kinds of text source files. These are namely text sources, manually transcribed audio, automatically transcribed audio, and a machine translation (into English) of the Mandarin transcripts. The format of the text source files depends on the kind of text represented and is shown in Table 20.

Table 20. **Text source files formats**

		Text Sources	**Audio Sources**	
			Manual Transcripts	**Automatic Transcripts**
English	**Original Language**	ASCII, 1 word per record	ASCII, 1 word per record	ASCII, 1 word per record
Mandarin	**Original Language**	GB, 1 character per record	GB, 1 character per record	GB, 1 word per record
	Machine Translation	ASCII, 1 word per record	ASCII, 1 word per record	ASCII, 1 word per record

The various evaluation tasks will have several input files supplied for the evaluation. The first is a task specific index file that contains the names and supplies the sequencing of the source files to be processed. The task specific index files are described in the following sections. The second index file contains auxiliary side information that systems are permitted to take advantage of. The auxiliary information index file is an ASCII text file with one record per file. The format of this file is given in Table 21.

Table 21. **Auxiliary information index file**

Record Structure	Record 1-N: <Filename> <Source> <Language> <DateStamp>
Field Descriptions	**Filename**: The root source file name, without file type extensions or directory information. (e.g. 19980612_1931_2033_NYT_NYT)
	Source: Name of the data source. (e.g. ABC_WNT, CNN_HDL)
	Language: Language (either ENGLISH or MANDARIN) of the original transcript.
	DateStamp: Date and time when the data collection began for this source. The format is: "YYYYMMDD HH:MM:SS" (e.g. "19980612 12:30:14")

[12] Sampled data files are structured as SPHERE files. SPHERE software and documentation is available at the following URL: http://www.itl.nist.gov/div894/894.01/software.htm.

A.2 STORY SEGMENTATION I/O FORMATS

Input data: The segmentation test will be directed by an index file containing a list of source files to be segmented. Systems must process the source files in order of occurrence. For the segmentation task, there will be three different index files, corresponding to audio data that is manually transcribed, automatically transcribed, and untranscribed audio. (Machine-translated source files will not be supported for the segmentation task.) The first data record in the index file will contain:

Record Structure	Header Record: \# <Task> <PointerType>
Field Descriptions	**Task**: An indication of the TDT task to be run. For story segmentation the field will contain **SEGMENTATION**.
	PointerType: The type of boundaries to be output by the system. The possible values are **RECID** for segmentation of source data in text form, or **TIME** for segmentation performed directly on the audio signal

Each subsequent data record in the index file will identify a source file to process. These records will have only field:

Record structure	Record 2…: <Source_file>
Field Descriptions	**Source_file**: The file name of the source data file to be processed. The source data file is either an untagged text stream file, or the original audio data file, depending on the selected condition.

Output data: Segmentation systems under evaluation must record segmentation decisions in an output file, one record for each hypothesized story boundary. The first record in this file will contain three fields that specify information that applies globally to the whole file, and the balance of the file will contain boundary decision records. These 3 fields will contain:

Record Structure	Header Record: <System> <N_d> <PointerType>
Field Descriptions	**System**: An alphanumeric character string that uniquely identifies the system being tested.
	N_d: The maximum allowed deferral period, in words for English text files, in characters for Mandarin text files, and in seconds for audio files.
	PointerType: The type of boundary indicated in the balance of the file. The possible values are **RECID** or **TIME.**

Each subsequent data record in the file will identify a hypothesized boundary. These records will have two fields and will contain:

Record Structure	Record 2…: <Source_file> <Boundary>
Field Descriptions	**Source_file**: The file name of the source data file to be processed. The source data file is either an untagged text stream file, or the original audio data file, depending on the selected condition.
	Boundary: a hypothesized boundary. For text files, Boundary is the index number of the first RECID in the hypothesized segment, in the range {1, 2, . . .}.[13] For audio files, Boundary is the time of the beginning of the segment, in the range {0.0, . . .}. (It isn't necessary to output the beginning of the first segment.) The hypothesized Boundary points must occur in chronological order.

A.3 TOPIC TRACKING I/O FORMATS

Input data: The tracking test will be directed by a set of index files, one per test topic. Each index file will contain a list of topic training stories, followed by a list of source files for which the target topic is to be tracked. The index file structure is as follows. First comes a header record:

Record Structure	Header Record: \# <Task> <PointerType> <Topic=N>
Field Descriptions	**Task**: An indication of the TDT task to be run. For topic tracking the field will contain **TRACKING**.
	PointerType: The type of boundaries to be output by the system. The possible values are **RECID** for segmentation of source data in text form, or **TIME** for segmentation performed directly on the audio signal.
	Topic=N: Declares the topic id number for which the test is to be run. (This is for documentation only. Allowable information for topic training is restricted to the indices of the training stories that follow.)

Then come $(N_t)_{Max}$ records containing the topic training stories, in chronological order: (Only the last N_t of these stories are to be used for training.)

[13] For English text sources, the RECID's are equivalent to words. For manually transcribed Mandarin text sources, RECIDs are equivalent to characters. For ASR transcribed Mandarin sources, RECIDs are equivalent to words.

Record Structure	Record 2…($(N_t)_{Max}$ +1): # Topic_training_story <Story_ID> <Source_file> <Begin> <End>
Field Descriptions	**Story_ID**: A character string story identifier. (This is the TDT3 corpus "docno".)
	Source_file: The file name of the source data file containing the training story. The source data file is either an untagged text stream file, or the original audio data file, depending on the selected condition and the source.
	Begin: The word index (or time, if the source file contains sampled audio data) of the beginning of the story.
	End: The word index (or time, if the source file contains sampled audio data) of the end of the story.

Finally come the records that identify the source files to process. These records will have two fields:

Record Structure	Record ($(N_t)_{Max}$ +2)…: <Source_file> <Begin>
Field Descriptions	**Source_file**: The file name of the source data file to be processed. The source data file is either an untagged text stream file, or the original audio data file, depending on the selected condition.
	Begin: the word/character index (or time, if the source file contains sampled audio data) of the point in the source at which processing is to begin. [14]

Output data: The Topic Tracking task is to hypothesize points in the source stream where the target topic is discussed. Topic tracking systems will perform this task by outputting information about these hypothesized points to a file, one record for each putative discussion of the target topic. The first record in this file will contain five fields that specify information that applies globally to the whole file. These 5 fields will contain:

Record Structure	Header Record: <System> <Boundaries> <N_t> <Topic> <PointerType>
Field Descriptions	**System**: An alphanumeric character string that uniquely identifies the system being tested. (E.g., CDM_P05-8.v37)
	Boundaries: either **YES** or **NO**, where **YES** indicates that story boundaries are supplied to the system being tested and **NO** indicates that they are not.
	N_t: The number of stories used to train the system to the target topic.
	Topic: An index number in the range {1, 2, . . . ~100} which indicates the target topic being tracked.
	PointerType: The type of boundaries to be output by the system. The possible values are **RECID** for textual source data or **TIME** for source data in audio form.

Each subsequent data record in the file will identify the beginning point in the source stream of a judgment about whether the target topic is being discussed, along with an associated decision and confidence. This decision and confidence will apply to all subsequent source data until the point specified by the next output data record. These records will have 4 fields and will contain:

Record Structure	Record 2…: <Source_file> <Pointer> <Decision> <Score>
Field Descriptions	**Source_file**: The filename of the source file being processed.
	Pointer: Indicates where in the source file the subject discussion commences.[15] For textual source data, Pointer is the beginning RECID in the source file (in the range {1, 2, . . .}). For source data in audio form, Pointer is the beginning time, in seconds.
	Decision: Either **YES** or **NO**, where **YES** indicates that the system believes that the source being processed does in fact discuss the target topic. **NO** indicates not.
	Score: A real number that indicates how confident the system is that the source being processed discusses the associated topic. More positive values indicate greater confidence.

[14] Processing begins at the very beginning of almost all source files. However, when the source file includes the end of the training data, tracking doesn't begin until after the last topic training story.

[15] Output records must be ordered so that each successive record's value of Pointer is greater than that of its predecessor, thus indicating the termination of the range of the predecessor record's Decision and Score.

A.4 TOPIC DETECTION I/O FORMATS

Input data: The topic detection test will be directed by an index file containing a list of source files for which topics are to be detected and tracked. Systems must process the source files in order of occurrence. The index files will follow this format:

Record Structure	Header Record: # <Task> <PointerType>
Field Descriptions	**Task**: An indication of the TDT task to be run. For topic detection the field will contain **DETECTION**.
	PointerType: The type of boundaries to be output by the system. The possible values are **RECID** for source data in text form or **TIME** for audio data.

Each subsequent data record in the file will identify a source file to process. These records will have only one field:

Record Structure	Record 2…: <Source_file>
Field Descriptions	**Source_file**: The filename of the source data file being processed.

Output data: The Topic Detection task is to detect topics and then to hypothesize points in the source stream where they are discussed. Topic Detection systems will perform this task by recording information about these hypothesized points in a file, one record for each putative discussion of a topic, written in ASCII format. The first record in this file will contain four fields that specify information that applies globally to the whole file. These four fields will contain:

Record Structure	Header Record: <System> <Boundaries> <N_f> <PointerType>
Field Descriptions	**System**: An alphanumeric character string that uniquely identifies the system being tested. (E.g., CDM_P05-8.v37)
	Boundaries: Either **YES** or **NO**, where **YES** indicates that story boundaries are supplied to the system being tested and **NO** indicates that they are not.
	N_f: The maximum deferral period allowed before a decision must be made.
	PointerType: The type of boundaries to be output by the system. The possible values are **RECID** for text stream segmentation or **TIME** for audio segmentation.

Each subsequent data record in the file will identify a topic, the point in the source stream that discusses it, and a measure of the confidence in the identification. This decision and confidence will apply to all subsequent source data until the point specified by the next output data record. These records will have 5 fields and will contain:

Record Structure	Record 2…: <Topic> <Source_file> <Pointer> <Decision> <Score>
Field Descriptions	**Topic**: An index number in the range {1, 2, . . .} which uniquely indicates the topic.
	Source_file: The filename of the source data file being processed.
	Pointer: Indicates where in the source file the subject discussion commences. For textual source data, Pointer is the beginning RECID in the source file (in the range {1, 2, . . .}). For source data in audio form, Pointer is the beginning time, in seconds.
	Decision: Either **YES** or **NO**, where **YES** indicates that the system believes that the source being processed does in fact discuss the target topic. **NO** indicates not.
	Score: A real number that indicates how confident the system is that the source being processed discusses the associated topic. More positive values indicate greater confidence.

A.5 FIRST-STORY DETECTION I/O FORMATS

Input data: The first story detection test will be directed by an index file containing a list of source files for which the first story of new topics are detected. Systems must process the source files in order of occurrence. The file format is identical to the detection index file, with the only exception being the designation of the task. The index files will follow this format:

Record Structure	Header Record: # <Task> <PointerType>
Field Descriptions	**Task**: An indication of the TDT task to be run. For first-story detection the field will contain **FIRST_STORY**.
	PointerType: The type of boundaries to be output by the system. The possible values are **RECID** for source data in text form or **TIME** for audio data.

Each subsequent data record in the file will identify a source file to process. These records will have only one field:

Record Structure	Record 2…: <Source_file>
Field Descriptions	**Source_file**: The filename of the source data file being processed.

Output data: The first story detection task is to determine whether or not each processed story is the first story of a new topic. The first story detection system will perform this task by outputting one record for each putative decision to a file. The first record in this file will contain four fields that specify information that applies globally to the whole file. These four fields will contain:

Record Structure	Header Record: <System> <Boundaries> <N_f> <PointerType>
Field Descriptions	**System**: An alphanumeric character string that uniquely identifies the system being tested. (E.g., CDM_P05-8.v37)
	Boundaries: Either **YES** or **NO**, where **YES** indicates that story boundaries are supplied to the system being tested and **NO** indicates that they are not.
	N_f: The maximum deferral period allowed before a decision must be made.
	PointerType: The type of boundaries to be output by the system. The possible values are **RECID** for text stream segmentation or **TIME** for audio segmentation.

Each subsequent data record in the file will identify a putative decision, the point in the source stream of that decision, and a measure of the confidence in the decision. This decision and confidence will apply to all subsequent source data until the point specified by the next output data record. These records will have 4 fields and will contain:

Record Structure	Record 2…: <Source_file> <Pointer> <Decision> <Score>
Field Descriptions	**Source_file**: the filename of the source data file being processed.
	Pointer: Indicates where in the source file the new subject commences. For textual source data, Pointer is the index number of the specified word, in the concatenation of all story texts for the source file (in the range {1, 2, . . .}). For source data in audio form, Pointer is the specified time, in seconds.
	Decision: Either **YES** or **NO**, where **YES** indicates that the system believes that the source being processed does in fact discuss a new topic. **NO** indicates not.
	Score: A real number that indicates how confident the system is that the source being processed discusses a new topic. More positive values indicate greater confidence.

A.6 LINK DETECTION I/O FORMATS

Input data: The story link detection test will be directed by an index file containing a list of source files to process followed by a list of story pairs for which story link decisions are to be made. Systems must process each story pair independently, and in order of occurrence. The index files will follow this format:

Record Structure	Header Record: # <Task>
Field Descriptions	**Task**: An indication of the TDT task to be run. For link detection the field will contain **LINK_DETECTION.**

Next is the list of source files to process, in chronological order:

Record Structure	Record 2…N_{source}+1: # source_file <Source_file>
Field Descriptions	**Source_file**: The file name of the source data file to be processed. The source data file is either an untagged text stream file, or the original audio data file, depending on the selected condition.

Each subsequent data record in the file will identify a pair of story ids to process. The first story id will be the first (i.e., oldest) story, and the second story id will be the second (i.e., most recent) story. The list will be sorted by the first column, and then by the second column. Thus the second (and most recent) story in each subsequent record will be in monotonically increasing chronological order. These records will have only two fields:

Record Structure	Record N+3…: <Story_ID_1> <Story_ID_2>
Field Descriptions	**Story_ID_1**: The character string story identifier of the first story. This identifier is the source file and the TDT3 corpus docno, separated by a colon.
	Story_ID_2: The character string story identifier of the second story. This identifier is in the same format as for the first story, namely the source file and the TDT3 corpus docno, separated by a colon.

Output data: The story link detection task is to determine whether or not each story pair is on the same topic, or "linked". The story link detection system will perform this task by outputting one record for each story pair presented in the index file. The order of the decisions in the output file *must* match the index file exactly. The first record in this file will contain two fields that specify information that applies globally to the whole file. These two fields will contain:

Record Structure	Header Record: <System> <N_f>
Field Descriptions	**System**: an alphanumeric character string that uniquely identifies the system being tested. (E.g., CDM_P05-8.v37)
	N_f: The maximum deferral period allowed before a decision must be made.

Each subsequent data record in the file will identify a putative decision on a story pair, and a measure of the confidence in the decision. These records will have 4 fields and will contain:

Record Structure	Record 2…: <Story_ID_1> <Story_ID_2> <Decision> <Score>
Field Descriptions	**Story_ID_1**: The character string story identifier of the first story. (This is the TDT3 corpus "docno".)
	Story_ID_2: The character string story identifier of the second story. (This is the TDT3 corpus "docno".)
	Decision: Either **YES** or **NO**, where **YES** indicates that the system believes that the story pair is linked. **NO** indicates not
	Score: A real number that indicates how confident the system is that the story pair is linked. More positive values indicate greater confidence.