

# **Topic Detection and Tracking Annotation Guidelines**

## ***Task Definition to Support the TDT2002 and TDT2003 Evaluations in English, Chinese and Arabic***

Version 1.5 – April 5, 2005

Linguistic Data Consortium

Created by: Stephanie Strassel, <strassel@ldc.upenn.edu>

[www.ldc.upenn.edu/TDT](http://www.ldc.upenn.edu/TDT)

## Table of Contents

1	General Overview .....	3
2	Segmentation.....	3
2.1	Segmentation Overview.....	3
2.2	Definitions and Terminology .....	4
2.2.1	Defining “Story” and the Two Clause Rule .....	4
2.3	Annotator Training .....	5
2.4	Annotation Tools.....	5
2.5	Transcripts and Data Format.....	6
2.6	Annotation Tasks.....	6
2.6.1	Boundary Identification .....	6
2.6.2	Segment Classification .....	7
2.6.3	Classifying Story Types .....	8
2.6.4	Timestamping .....	10
2.6.5	Segmentation Example .....	11
2.7	Quality Control.....	11
3	Topic Annotation.....	12
3.1	Overview.....	12
3.2	Definitions and Terminology .....	13
3.3	Topic Selection .....	15
3.4	Topic Definition .....	16
3.5	Topic Research .....	17
3.6	Search Guided Topic Labeling .....	18
3.7	Quality Control.....	20
4	APPENDIX 1: Examples of selected general story types .....	22
5	APPENDIX 2: Examples of selected source-specific story types.....	28

# 1 General Overview

This document describes the process of corpus annotation for the Topic Detection and Tracking (TDT) project, specifically focusing on the annotation procedures adopted for the 2002 and 2003 TDT evaluations, using the TDT4 corpus. Corpus creation begins with the Linguistic Data Consortium collecting the multilingual raw material, newswires and other electronic text, web audio, broadcast radio and television to be included in the corpus. The next step is to produce text intermediaries for all audio material. The text intermediaries are used both in subsequent annotation tasks and in evaluation tasks. The text intermediaries are next segmented into individual stories that become the input for all subsequent annotation tasks. Before annotation can begin, LDC annotators select and define topics explicitly according to the procedure described below. With topics in hand, the team annotates the corpus with the aid of a search engine in an attempt to identify all stories that discuss each of the selected topics. Quality control measures are adopted at each stage of annotation. Once the corpus has been completely annotated, the text and the tables that encode annotators' judgments are formatted to accommodate the evaluation specification and released to the research community either directly or via the National Institute of Standards and Technology (NIST) who manage TDT evaluations.

## 2 Segmentation

### 2.1 Segmentation Overview

Segmentation refers to the identification of individual stories within a news broadcast by the insertion of boundaries where the topic of discussion changes. Newswire services deliver material with story boundaries marked, and while long newswire stories are occasionally split into two or more parts to ease transmission, LDC employs automatic processes to unite split stories prior to annotation. On the other hand, commercially-produced transcripts and closed-captioned text from broadcast sources require manual segmentation. Closed captioning services and commercial transcription agencies use established conventions for marking topic boundaries and speaker turn changes in news broadcasts, and these boundaries come with approximate timestamps which must be reviewed for accuracy. Annotators listen to the audio of the entire broadcast while viewing the corresponding waveform display and text intermediary and add, remove or re-position story boundaries as necessary. Annotators also classify each boundary as beginning a 1) news story, 2) teaser section, 3) miscellaneous text section, or 4) untranscribed or undertranscribed section. *News stories* are typically contiguous segments of a broadcast. *Teasers* consist of promotion for upcoming stories or brief reviews of top stories reported on more fully elsewhere within the same broadcast. *Miscellaneous text* segments include commercials and reporter chit-chat. If an audio segment contains no text or inadequate text to determine its topic, that segment is classified as *un(der)transcribed*. Annotators also perform a variety of quality control measures

to verify the accuracy of the segmentation and to measure inter-annotator variation.

## 2.2 Definitions and Terminology

A number of specialized definitions have been adopted for the Topic Detection and Tracking project. As our approach to segmentation has evolved through the various corpora, our use of terminology has also been refined to reflect current annotation practices.

### 2.2.1 Defining “Story” and the Two Clause Rule

For the purposes of all TDT corpora, a **story** is a topically contiguous section of news text or segment in a broadcast. During segmentation of the TDT-2 corpus, the definition of story had one additional requirement, known as the “two clause rule”: a story was defined as a section containing at least two independent declarative clauses on the same topic. Although this may seem straightforward, it in fact slowed progress and increased confusion as annotators counted clauses and argued over which were independent versus dependent. Some adopted a strict interpretation, while others would classify as a story a single independent clause with multiple dependent clauses if they felt it contained adequate content. Segments of potential interest that failed the two clause rule, e.g. brief stock reports, were considered non-news. The rule also introduced inconsistency in classification of particular types of news, depending on the presentation of the story within the broadcast. For instance, this stock report from the TDT-2 corpus

**<sr 1164.00>**

BACK ON WALL STREET, THE NASDAQ IS DOWN 14 AND THE DOW FINISHES UP.

is considered a story, since it contains two independent clauses, while this one,

**<sn 1020.42>**

BOTH THE DOW AND NASDAQ ARE UP TODAY IN HEAVY TRADING DESPITE LINGERING CONCERNS OVER THE SITUATION IN SOUTHEAST ASIA.

with only a single independent clause, is considered non-news. Previews or teasers about upcoming reports at the start of a broadcast or immediately before a commercial break also varied in their form, sometimes containing two independent clauses, other times not. For example, consider the inconsistent treatment of the contiguous upcoming story briefs in the following CNN report from TDT-2:

**<sn 27.710>**

I’m Elizabeth Arat. Russia sends a high-ranking diplomat to Iraq trying to find a peaceful solution to a worsening crisis.

**<sr 35.378>**

The World Health Organization has a new president and she's different from all previous presidents.

**<sn 41.239>**

We'll also hear about new words creeping into the Russian language. Those stories and more in the next hour on World Report.

Although all three story briefs occur within the same Upcoming Stories section of the broadcast, they were classified differently because of their clause count. The first and last segments contain only one independent clause on the news topic; the middle segment contains two clauses on the single topic, so this segment alone is categorized as a news story (<sr>).

Because of the difficulty in applying the two clause rule and the inconsistencies it introduced with respect to segment type classification, it was not employed in the construction of future TDT corpora. Instead, annotators creating the TDT-3 and TDT-4 corpora classified segments as news stories if they could easily render a judgment about the segment's relevance to a topic; story size became irrelevant.

### 2.3 Annotator Training

Annotators cut their teeth on files selected from the corpus to include a number of difficult situations, including files containing large sections of under-transcribed text, files containing drop-outs in the audio signal and files that require close attention to the actual content of the reports in order to find the story boundaries. Consider the following example:

```
<t 165.00>Reporter: have you ever seen anything like this before?  
<t 168.00>I've seen serious storms. This is the worst since I've  
been here in 1972.  
<sr 172.118>Maintenance workers are trying to remove slabs of ice  
from atop buildings and bridges.  
<b 182.00>Power was restored to much of the city overnight, but  
700,000 people outside the city are still in the dark.
```

From a brief reading, one might conclude that there is one story in this example, but there are in fact two different stories about two different ice storms, one in New York, and one in Montreal, a fact which becomes clear only upon careful listening and examining contextual cues. Subtleties of this kind are not uncommon and emphasize the need to use all one's resources, text, audio and waveform, during segmentation. Annotators are required to demonstrate their ability to establish accurate story boundaries and timestamps in difficult cases like these before they are permitted to work on actual segmentation files.

### 2.4 Annotation Tools

The creation of specialized tools for each phase of annotation allows TDT annotators to focus on the specific task at hand. The custom segmentation interface combines the Emacs text editor with the Xwaves audio program and adds further features to integrate the functions of each. Additional tools have

been developed for quality control passes. The annotation tools automatically log annotation judgments into an Oracle database.

## **2.5 Transcripts and Data Format**

While TDT newswire is collected as electronic text, the broadcast sources are collected as audio (and video where available). To support annotation and TDT research, LDC produces text intermediaries of all TDT audio. Text intermediaries come from a number of sources. Closed-captioning or commercial transcripts are used when they are available; otherwise, transcription agencies are contracted to provide closed-caption quality transcripts of the television and radio sources. Incoming transcripts and closed-caption files take a variety of formats, and must be converted into formats that are optimized for the annotation tasks. During segmentation, annotators work with a plain-text version of the transcript file, inserting timestamps and story boundaries directly into the text intermediary. The segmented transcripts are then converted to SGML-structured reference texts for use in subsequent annotation tasks, and into tokenized files and story boundary tables for use in the evaluation tasks. All examples of segmentation files cited in this document refer to the plain-text version of the transcripts used by the annotators during the segmentation task.

## **2.6 Annotation Tasks**

Segmentation comprises three distinct annotation tasks. Each will be described in turn below.

### **2.6.1 Boundary Identification**

First, annotators must identify each transition between individual segments within a news broadcast. The transcript corresponding to each audio broadcast contains initial, putative segment boundaries provided in the caption stream or by the commercial transcription agency. These boundaries take the form of a section tag that appears on a separate line within the transcript preceding the section of text to which it corresponds. Each section tag also contains an initial timestamp corresponding to the putative start time of the following text. The tags follow this template: <s ##.###>, with the timestamp indicated in seconds elapsed since the start of the broadcast. Depending on the source of the initial transcript, timestamps may be displayed as round numbers, or to the tenth, hundredth or thousandth of a second. As annotators review and edit section tags, they standardize the format and expand all timestamps to the thousandth of a second.

The initial section tags provided in the transcripts are not always accurate. In some cases, boundaries have been inserted erroneously, dividing a single segment into two. In addition, not all boundaries between sections are identified. Annotators are therefore required to listen to an audio file in its entirety while viewing the corresponding text intermediary, rather than just listening to the

sections surrounding a segment boundary. They must then add or remove segment boundaries as necessary.

In addition to the putative segment boundaries, most text intermediaries also contain pre-existing <b> (break) and <t> (turn) tags that follow the same format as the <s> (section) tags. These <b> and <t> tags are inserted automatically by the closed captioning or transcription services, and have no bearing on the segmentation task. Annotators are not required to modify, remove or add <b> or <t> tags.

## **2.6.2 Segment Classification**

The second segmentation task involves classifying each segment according to its type. Within the TDT-4 corpus there are four distinct segment types:

### **2.6.2.1 News stories**

A news story is a topically contiguous segment of the broadcast. News stories may be of any length, even fewer than two independent clauses, as long as they constitute a complete, cohesive news report on a particular topic. Note that single news stories may discuss more than one related topic. When reports of similar content are adjacent to one another in a news broadcast, it is often difficult to tell where one story ends and the next begins. Annotators are instructed to rely on audio cues (speaker changes, music, pauses) to inform their judgments. Within the transcript, news stories are labeled <sr>.

### **2.6.2.2 Teasers**

Teasers consist of promotional spots for upcoming stories or very brief reviews of top stories reported on more fully within the same broadcast. Teasers are quite limited in their format and their placement within an audio broadcast. They typically occur at the beginning of a broadcast or preceding a commercial break, and are designed to capture and hold the listener's attention for stories that will be reported on later in the broadcast. Brief (one to two sentence) reviews of top headlines are also categorized as previews; these stories are also covered more fully elsewhere in the broadcast. Teaser sections are labeled as <st> within the transcript. Each teaser segment containing short (or variable-length) references to two or more distinct topics are not subdivided into individual teaser sections. If a broadcast begins with a teaser that mentions three distinct stories that are coming up later in the broadcast, these three are grouped together into a single <st> teaser segment; they are not rendered as a sequence of three distinct teaser segments divided by topic.

### **2.6.2.3 Miscellaneous text**

Miscellaneous text includes non-news segments like commercials, reporter chit-chat, station identifications, public service announcements, promotions for upcoming broadcasts and long musical interludes. In addition, in the TDT-2

corpus some teaser sections (namely, those consisting of fewer than 2 independent declarative clauses) were also labeled as miscellaneous text; in TDT-3 all preview segments were labeled as miscellaneous text. Miscellaneous sections lasting less than 8 seconds that indicate transitions between two news stories, such as a brief musical interlude, are not segmented as separate non-news sections; instead they are included within the preceding segment. Miscellaneous text sections receive a tag of <sn> within the transcript. If multiple miscellaneous sections follow one another within a transcript, they are grouped together with a single <sn> tag at the beginning of the section.

#### **2.6.2.4 Un(der)transcribed sections**

If an audio segment contains no text or inadequate text within the corresponding transcript such that it is difficult or impossible to determine its topic, that segment is classified as un(der)transcribed. Transcript sections that are incomplete but contain enough text to identify the main topic are classified as valid news stories. Un(der)transcribed sections of commercials, chit-chat, previews and the like, which would normally be considered miscellaneous text, receive the miscellaneous text classification. Un(der)transcribed sections are labeled with a tag of <su>.

### **2.6.3 Classifying Story Types**

The four segment categories described above encompass a number of different story types. Table 1 presents a comprehensive overview of the types of stories annotators routinely encounter in a broadcast file, along with the classification of each story type for the three TDT corpora. Specific examples of story types can be viewed in Appendix 1.

**Table 1. General story types and their classification within the TDT corpora**

Story Type	TDT-2	TDT-3	TDT-4
<b>Miscellaneous text</b>			
commercials	sn	sn	sn
reporter chit-chat	sn	sn	sn
musical interludes	sn	sn	sn
isolated anchor/program IDs	sn	sn	sn
local station ID	sn	sn	sn
public service announcements	sn	sn	sn
promo for upcoming programs	sn	sn	sn
lists of temperatures/ weather conditions	sn	sn	sn
lists of sports scores	sn	sn	sn
lists of stock quotes	sn	sn	sn
consecutive <sn>	cluster together		
consecutive <st>			
<b>Un(der)transcribed section</b>			
un(der)transcribed sections	su	su	su
<b>Preview/Teaser section</b>			
previews			
< 2 clauses	sn	sn	st
≥ 2 clauses	sr	sn	st
<b>General categories of news stories</b>			
news stories			
< 2 clauses	sn	sr	sr
≥ 2 clauses	sr	sr	sr
top news stories (recaps)			
< 2 clauses	sn	sr	st
≥ 2 clauses	sr	sr	st
weather reports	sr	sr	sr
sports reports	sr	sr	sr
stock reports	sr	sr	sr
interview segment	treat as single story		

### 2.6.3.1 Source-specific story types

In addition to the general story types that occur in most broadcast files, regardless of source or language, annotators also encounter source-specific peculiarities. Many of these special segment types defy the general segmentation principles adopted for TDT, and must be handled individually. Annotator training materials document these source-specific segment types and their treatment in TDT segmentation. Table 2 provides an overview of segmentation practices for each of the known source-specific segment types; examples of these types can be viewed in Appendix 2.

**Table 2. Source-specific segment types and their classification within the TDT corpora**

Some source-specific segment types	TDT-2	TDT-3	TDT-4
<b>PRI</b> GeoQuiz 1st part GeoQuiz 2nd part Musical Segment	sr sr sr	sr sr sr	sn <sup>1</sup> sn sr
<b>VOA_English</b> Editorials Communications World This Date In History Internet News	sr n/a <sup>2</sup> sn n/a	sr n/a sn n/a	sr variable <sup>3</sup> sn sn
<b>VOA_Mandarin</b> English lesson Historic Person of the Day	n/a n/a	n/a n/a	sn sn
<b>CNN</b> Play of the Day On the Scoreboard Cold and Flu Report	sn sn sn	sn sn sn	sn sn sn
<b>MNB, various Chinese sources</b> newspaper headlines review	n/a	sn	sn

## 2.6.4 Timestamping

The final segmentation task is to label each section boundary with a timestamp that corresponds to its start time. In most cases, the text intermediaries contain a putative audio timestamp corresponding to each putative segment boundary. Annotators review the pre-existing timestamps and boundaries while listening to the audio file and add, remove or re-position timestamps as needed. Annotators are instructed to pay particular attention to the accuracy of each segment boundary, and are given the following instructions for checking and modifying timestamps:

When you listen to the portion of the recording that starts at the time stamp for a boundary, you should hear all of the text that follows that boundary without any

<sup>1</sup> Some source-specific segment types classified as <sn> (GeoQuiz, newspaper headline reviews) occasionally occur not in isolation, but rather as a direct lead-in to a larger news story. In these rare cases, annotators treat the segment as a story introduction, and they are segmented as part of the larger news story rather than as a separate non-news section.

<sup>2</sup> N/a means that this segment type did not occur in this corpus.

<sup>3</sup> Communications World is an irregularly-occurring feature that focuses on radio technology around the world. Individual segments within this feature vary greatly in their content. Some features are legitimate news stories with a topically contiguous focus, and are classified as <sr> sections. Other features contain lists of short-wave radio frequencies or email messages from Communications World listeners, and are classified as non-news. Examples of each type can be viewed in the appendix.

clipping of the first word, and you should NOT hear any of the text that precedes the boundary. If the first word following the boundary is not fully audible, or if you hear text that comes before the boundary, the time stamp needs to be shifted. The raw transcripts provided for audio data come from commercial transcription services or closed captioning. Time stamps associated with all boundaries are created using tools that do not allow adequate precision, and must be considered essentially fictional.

### 2.6.5 Segmentation Example

What follows is a section of a CNN broadcast, pre- and post-segmentation. Observe the differences in the number and location of segment boundaries, the classification of each boundary, and the timestamps.

Initial Transcript	Completed, Segmented Transcript
<p><b>&lt;s 1277.0&gt;</b>            finally, errict rhett will miss the rest of the season with a torn ligament in his foot. jerome jurenovich, "cnn headline sports."</p> <p><b>&lt;s 1444.0&gt;</b>            al gore and george w. bush hope to use the presidential debates to pull ahead in the race. bruce morton looks at the impact of past debates that didn't run as smoothly as planned.            &lt;t 1455.0&gt;            have we had debate blunders? yes. did they matter? that's a harder question.</p> <p><b>&lt;s 1460.0&gt;</b>            in 1988, debate moderator bernard shaw of cnn asked democratic candidate michael dukakis, who opposed the death penalty, a provocative question.</p>	<p><b>&lt;sr 1264.716&gt;</b>            finally, errict rhett will miss the rest of the season with a torn ligament in his foot. jerome jurenovich, "cnn headline sports."</p> <p><b>&lt;sn 1278.456&gt;</b>            (untranscribed commercial)</p> <p><b>&lt;sr 1437.861&gt;</b>            al gore and george w. bush hope to use the presidential debates to pull ahead in the race. bruce morton looks at the impact of past debates that didn't run as smoothly as planned.            &lt;t 1455.0&gt;            have we had debate blunders? yes. did they matter? that's a harder question.            &lt;b 1460.0&gt;            in 1988, debate moderator bernard shaw of cnn asked democratic candidate michael dukakis, who opposed the death penalty, a provocative question.</p>

Figure 1. Segmentation Example

### 2.7 Quality Control

Beyond the practical annotator training described in Section 2.3 above, LDC employs some additional measures to ensure consistency in segmentation. First, annotators are provided with written documentation describing segmentation procedures. This local annotation guide is made available as an integrated webpage that annotators can access while performing the task. All questions and answers that emerge during segmentation are stored in a

searchable email archive, and annotators are required to participate in regular meetings to discuss problems and progress.

These general quality assurance measures can be coupled with a number of additional task-specific measures, including second passing, spot-checking, dual segmentation and the evaluation of the ratio of word tokens to time. The level of inclusion of these additional measures has varied from one TDT corpus to the next, depending on budget and timeline.

The TDT-4 segmentation effort adopts a focused approach to second passing. Annotators revisit those sections of the broadcast that are particularly tricky; namely segment boundaries, teaser sections and particular source-specific section types (those listed in Table 2 above). During second passing, annotators specialize in a particular broadcast source. Each broadcast source has its own peculiarities in terms of broadcast structure; as source-specialists annotators find it easy to recognize inconsistencies across broadcast files. More general second passing of additional material will take place as time and budget allow.

In addition to second passing, and as a regular part of the segmentation process, senior annotators conduct spot-checks on approximately 5% of all segmented material. As mistakes or inconsistencies are identified, these are discussed with the specific annotators involved and are corrected; inconsistencies are also discussed with the entire annotation team via e-mail and during weekly annotator meetings.

The final quality control measure, introduced towards the end of TDT-2 and adopted as a regular check during TDT-3 and TDT-4, is the measure of word tokens in the text intermediary of a segment per unit of time in the corresponding audio. Stories with an unusual ratio of text words to audio duration arouse suspicion of a segmentation error. These cases are reviewed and re-segmented as appropriate. Although this method produces a number of false alarms involving news stories with long musical interludes, it does prove helpful for identifying missed boundaries in some audio sources.

### **3 Topic Annotation**

#### **3.1 Overview**

After segmentation is completed, annotators turn their attention to topic annotation. Topic annotation encompasses the selection, definition, research and labeling of topics within the TDT corpus. In topic selection, annotators review a stratified random list of seed stories selected from the target language(s), identifying those stories whose seminal event is likely to be discussed in TDT sources across all languages. Once a seed story containing an appropriate seminal event has been identified, team leaders convert the seed story into a full-fledged topic, documenting factual information (what, who, where, when) for each topic as well as providing guidelines for interpreting a topic's

scope. Annotators contribute to topic definition by conducting topic research, which provides additional contextual information about each topic, including things like keywords, timelines and named entities. Once topic definition and research have been completed for a given topic, annotators begin topic labeling. Using the EZQuery search engine, annotators conduct multiple queries over the corpus to identify all on topic stories for a given topic. Quality control measures include a precision check to identify false alarms, adjudication of sites' results to identify misses, and dual annotation with discrepancy resolution to establish a measure of inter-annotator consistency.

### 3.2 Definitions and Terminology

The notions of event and topic are crucial to TDT annotation. A TDT *event* is defined as a **specific thing that happens at a specific time and place along with all necessary preconditions and unavoidable consequences**. For instance, when an U.S. Marine jet sliced a funicular cable in Italy in February 1998, the cable car's crash to earth and the subsequent injuries were all unavoidable consequences and thus part of the same event. For the purposes of TDT, a *topic* is defined as **an event or activity, along with all directly related events and activities**. It is important to highlight the difference between a TDT topic and the notion of topic in normal discourse. While one might normally think of a topic as something broad like "accidents", a TDT topic is limited to a specific collection of related events of the type accident, in this case a particular cable car crash.

To increase the consistency of judgments about what constitutes "related" events, annotators refer to a set of *rules of interpretation*. These rules state, for each type of *seminal event* -- crimes, natural disasters, scientific discoveries, scandals, etc. -- what other types of events should be considered related. This then informs the annotators' judgments about which stories are "on-topic". In the example above, stories about the investigation, the Marine pilot, the repercussions for his unit, the victims' families and their quest for justice are all on topic. The TDT-2 and TDT-3 annotation efforts included eleven topic types; TDT-4 adds a two more types, *Diplomatic/Political Meetings* and *Celebrity News*, to cover common topical themes that would previously have been included in the *Miscellaneous News* category. The thirteen types, with examples of each drawn from the TDT-3 Y2001 Training Topics, are:

1. **Elections**, e.g. 30030: Taipei Mayoral Elections

*Seminal events* include: a specific political campaign, election day coverage, inauguration, voter turnouts, election results, protests, reaction.

*Topic* includes: the entire process, from announcements of a candidate's intention to run through the campaign, nominations, election process and through the inauguration and formation of a newly-elected official's cabinet or government.

2. **Scandals/Hearings**, e.g. 30038: Olympic Bribery Scandal

*Seminal events* include: media coverage of a particular scandal or hearing, evidence gathering, investigations, legal proceedings, hearings, public opinion coverage.

*Topic* includes: everything from the initial coverage of the scandal through the investigation and resolution.

3. **Legal/Criminal Cases**, e.g. 30003: Pinochet Trial

*Seminal events* include: the crime itself, arrests, investigations, legal proceedings, verdicts and sentencing.

*Topic* includes: the entire process from the coverage of the initial crime through the entire investigation, trial and outcome. Changes in laws/policies as a result of a crime are not generally on-topic unless a clear and direct connection between the specific crime and the legislation is made.

4. **Natural Disasters**, e.g., 30002: Hurricane Mitch

*Seminal events* include: weather events (El Nino, tornadoes, hurricanes, floods, droughts), other natural events like volcanic eruptions, wildfires, famines and the like, rescue efforts, coverage of economic or human impact of the disaster.

*Topic* includes: the causal (weather/natural) activity including predictions thereof, the disaster itself, victims and other losses, evacuations and rescue/relief efforts.

5. **Accidents**, e.g., 30014: Nigerian Gas Line Fire

*Seminal events* include: transportation disasters, building fires, explosions and the like.

*Topic* includes: causal activities and all their unavoidable consequences like death tolls, injuries, economic losses, investigations and any legal proceedings, victims' efforts for compensation.

6. **Acts of Violence or War**, e.g., 30034: Indonesia/East Timor Conflict

*Seminal events* include: a specific act of violence or terrorism or series of directly related incidents (such as a strike and retaliation).

*Topic* includes: Direct causes and consequences of a particular act of violence such as preparations (including technological/weapons development), coverage of the particular action, casualties/loss of life, negotiations to resolve the conflict, direct consequences including retaliatory strikes. This topic type is difficult to define across the board, and can easily become extremely broad and far-reaching. As such, each topic of this type is treated individually and is defined in such a way as to sensibly limit its scope and make annotation manageable.

7. **Science and Discovery News**, e.g., 31019: AIDS Vaccine Testing Begins

*Seminal events* include: announcement of a discovery or breakthrough, technological advances, awards or recognition of a scientific achievement.

*Topic* includes: Any aspect of the discovery, impact on everyday life, the researchers or scientists involved, descriptions of research and technology directly involved in the discovery.

8. **Financial News**, e.g., 30033: Euro Introduced

*Seminal events* include: specific economic or financial announcements (like a specific merger or bankruptcy announcement); reactions to the event; direct impact on the economy or business world. General economic trends or patterns without a clear seminal event are not appropriate as TDT topics.

*Topic* includes: the specific event, its direct causes, impacts on finance, government interventions or investigations, public or business world reactions, media coverage and analysis of the event.

9. **New Laws**, e.g., 30009: Anti-Doping Proposals

*Seminal events* include: announcement of new legislation or proposals, acceptance or denial of the legislation, reactions.

*Topic* includes: the entire process, from announcement of the proposal, lobbying or campaigning, voting surrounding the legislation, reactions from within the political world and from the public, challenges to the proposal, analysis and opinion pieces concerning the legislation.

10. **Sports News**, e.g., 31016: ATP Tennis Tournament

*Seminal events* include: a particular sporting event or tournament, sports awards, coverage of a particular athlete's injury, retirement or the like.

*Topic* includes: training or preparations for a competition, the game itself, results.

For tournament and championship events like the World Series or Superbowl, only direct precedents are considered on topic. Therefore, semi-finals and finals games leading up to the championship are on topic, but regular season play is not.

11. **Political and Diplomatic Meetings**, e.g., 30018: Tony Blair Visits China

*Seminal events* include: preparations for the meeting, the meeting itself, outcomes, reactions.

*Topic* includes: the whole process from the preparations and travel, the meeting itself, media coverage and public reaction, any outcome including legislation or policies adopted as a direct outcome of the meeting. Sources often report on one of a series of meetings between two officials or delegations; in these cases, only the current meeting part of the topic, although planning for a future meeting that is a direct outcome of the current meeting and is discussed as part of the current meeting will be considered on topic.

12. **Celebrity and Human Interest News**, e.g., 31036: Joe DiMaggio Illness

*Seminal events* include: most often involves the death of a famous person or other significant life events like marriage.

*Topic* includes the specific event, causes (such as illness in the case of a celebrity's death) or consequences (such as a funeral or memorial service), public reaction or media coverage, editorials and opinion pieces, retrospectives or life histories that are a direct consequence of the seminal event.

13. **Miscellaneous News**, e.g., 31024: South Africa to Buy \$5 Billion in Weapons

*Seminal events* include all specific events or activities that do not fall into one of the above categories.

*Topic* includes the event itself, direct causes and unavoidable consequences thereof.

This particular conceptualization of topic is a critical component of TDT annotation, as it allows annotators to potentially identify *all* the stories in the corpus that discuss some pre-defined topic. The topic definitions and rules of interpretation ensure that each annotator is working with the same understanding of the topic at hand and, at least in theory, that all annotators will identify the same stories as on-topic. With these basic concepts as a foundation, TDT annotators select, define and research topics prior to beginning topic labeling.

### 3.3 Topic Selection

In 2002, LDC selected 60 topics based upon a stratified, random sample of documents from the English and Chinese TDT4 sources. Forty of these topics were targeted for the year 2002 TDT evaluation, and the remaining 20 topics were held in reserve for the 2003 evaluation. Prior to the 2003 evaluation, 20 additional topics were selected from the TDT4 Arabic sources. These were

added to the 20 reserve topics from 2002, resulting in 40 evaluation topics for 2003. To summarize:

**2002 topics:** 20 English, 20 Chinese drawn from TDT4

**2003 topics:** 10 English, 10 Chinese, 20 Arabic drawn from TDT4

The stratified random topic selection method gives each month of data from each source an equal chance of contributing a topic, although no effort was made to ensure equal representation of each source/month in the final set of selected topics. Within any month of data from a source, stories were selected at random. Annotators reviewed randomized lists of 2000 seed stories for each language. Seed stories are identified as potential topics if they meet two criteria: 1) they contain a clear seminal event; and 2) they are likely to produce additional on-topic stories in all languages. There is no requirement that any topic produce a minimum number of on-topic stories beyond the initial seed story. While seed stories are selected from a single language, topics are defined multilingually and topic annotations are performed for each topic in all languages.

### 3.4 Topic Definition

Once a seed story has been identified as an appropriate TDT topic, it is converted into a full-fledged topic through the process of topic definition. Topic definition consists of a number of fixed components. The topic title is a brief phrase that is easy to remember and immediately evokes the topic. Each topic is accompanied by a topic icon, which provides the annotator with a visual reminder of the topic's content. The seminal event that contributed the topic is described by answering the questions what, who, when and where with regard to the event. The **Topic Explication** section provides a factual description of the topic's content. The **On Topic** section applies the rule of interpretation to the seminal event and spells out what other kinds of events should be included in the topic's scope. An optional **Notes** section warns the annotator of potentially confusing or difficult aspects of the topic, and may explicitly limit the scope of the topic for purposes of annotation. In addition, a series of links to additional information are provided for the annotator, including: seed story (with English translation if the seed is Chinese); Arabic and Chinese versions of the topic definition and keywords for each topic; the related rule of interpretation; and a link to the topic research document which provides further information about the topic. A sample topic description document appears below.

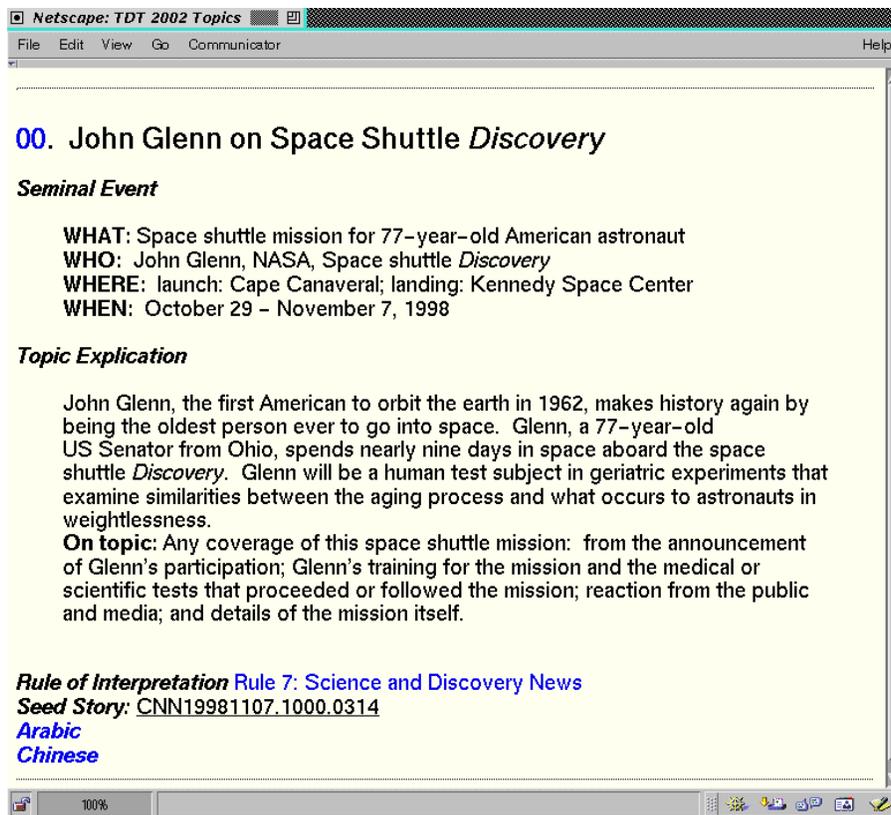


Figure 2. A sample topic definition

### 3.5 Topic Research

One of the biggest challenges for annotators is the task of keeping abreast of developments for a particular topic and understanding the scope of a topic in the corpus. Although the topic definitions spell out what sorts of stories might be considered on-topic, it is impossible to know in advance from having examined only one seed story how the topic might develop over time. In order to put the topics into a larger context, annotators conduct topic research, developing additional material like timelines, maps, keywords, named entities and links to non-TDT online resources for each topic. In addition, annotators use the topic research process to investigate how a given topic might be reported on across the sources, given the fact of media bias and the likelihood that different language sources might emphasize different aspects of the topic.

In search guided annotation, topic research feeds directly into topic labeling. Annotators submit parts of the topic research document as a query to the search engine during one stage of annotation. Topic research is a valuable resource not only for topic labeling, but also at later stages of quality control, when it provides a framework to monitor topic development and curb “topic drift”. Topic research is always accessible to annotators and is updated as the project and the topics evolve. A sample topic research page follows:

## Topic 00

### John Glenn on Space Shuttle 'Discovery'

Bethany Klein



*We're not stopping any time soon, so if you need to use the bathroom before we take off please raise your hand now.*

John Glenn, the first American to orbit the Earth in 1962, makes history again by being the oldest person ever to go into space. Glenn, at 77, spends nearly nine days in space aboard the space shuttle 'Discovery'. Glenn will be a human test subject in geriatric experiments that examine similarities between the aging process and what occurs to astronauts in weightlessness. Older people, for example, tend to lose bone and muscle mass, have trouble sleeping and experience decreased cardiovascular strength. That also happens to astronauts in space, but they soon recover on Earth.

#### Timeline

*6.21.1997:* Glenn volunteers for a space shuttle mission if NASA decides to study how weightlessness affects an older person.

*1.16.1998:* At NASA news conference, NASA administrator Daniel S. Goldin announces plans to send Glenn on the shuttle. Q and A session opens up concerning charges of political favoritism as a motivation for the decision. These questions, as well as questions regarding the validity and worth of Glenn's voyage, continue throughout the time period.

*8.13.1998:* Glenn takes a break from his Senate duties to be a full-time astronaut. From this point until the launch he trains almost continuously in Houston and Florida.

*10.9.98:* According to an ABCnews survey, nearly six in 10 say, whatever the reason (real scientific purposes or mainly as a p.r. gesture), it's good enough for Glenn to go on the mission.

*10.29.98:* Glenn and six other crew members meet with family members one last time before boarding the space shuttle discovery.

John Glenn is officially on his second trip orbiting the planet, a nine-day, zero-G journey that makes him the world's oldest astronaut. The afternoon launch was a success, despite a small panel falling off the shuttle Discovery during liftoff.

Glenn begins his geriatric research which includes sleeping nights in a cumbersome head net and body suit equipped with 23 sensors as well as being hooked up to a heart monitor with just seven electrodes and a miniature data recorder.

*11.4.98:* Glenn appears in an air-to-ground interview on Jay Leno's "The Tonight Show".

*11.7.98:* The shuttle crew members have a perfect landing at the Kennedy Space Center in Florida, capping a flight that returned the 77-year-old Glenn to space after 36 years. His heart and blood pressure were recorded during landing, and four hours of medical evaluation awaited him after he debarked.

*11.8.98:* Glenn goes to Johnson Space Center in Houston for three more weeks of additional medical tests. He and his fellow astronauts join family and friends there to celebrate their return.

Figure 3. A sample topic research document

## 3.6 Search Guided Topic Labeling

Early stages of the TDT project employed a brute-force approach to topic labeling, in the sense that every story was read and exhaustively labeled against every topic. Beginning with the TDT-3 2000 Evaluation, a search guided annotation strategy was adopted. Previous experiments had shown that search guided annotation could produce results as good as brute force annotation while reducing costs and the effect of annotator fatigue. In search guided annotation, individual annotators work with one topic at a time. After topic selection, research and definition are complete, one annotator employs the EZQuery search engine to make multiple passes over the English data, while other annotators make multiple passes over the Chinese and Arabic data in an attempt to find all stories discussing the topic at hand. Previous TDT efforts utilized a three-way relevance scale, in which stories were labeled with one of the following labels:

**YES:** more than 10% of the document is related to the topic

**BRIEF:** 10% or less of the document is related to the topic (includes passing mentions of the topic)

**NO:** the document does not mention or discuss the topic at all.

TDT-4 eliminates the **BRIEF** category in favor of a two-way relevance scale. **BRIEF** has always proven difficult for annotators to interpret and implement in a consistent manner. Although **BRIEF** was originally adopted as part of TDT-2 for the benefit of annotators, in fact the three-way decision is more difficult and creates more uncertainty than forcing a simple **YES/NO** decision.

Documents formerly labeled **BRIEF** are now split into either **YES** or **NO** depending on their content. For purposes of TDT-4, **YES** and **NO** are defined as follows:

**YES:** this story discusses the topic in a substantial way.

**NO:** this story does not discuss the topic at all, or only mentions the topic in passing without giving any information about the topic.

Formerly **BRIEF** documents are labeled **YES** if they present any substantive information, no matter how little, about a topic; they are labeled **NO** if they mention a topic only in passing without providing any information about the topic.

Although most **YES/NO** decisions are relatively straightforward, some decisions are difficult. Annotators are instructed to treat difficult cases as follows:

*If you're having trouble deciding between YES and NO, ask yourself whether you learned anything about the topic by reading the story, no matter how small and no matter if you've seen that same information before. If you learn something about the topic by reading the story, then it should count as YES. If you're still having trouble making up your mind, consult with your team leaders and post a message to the TDT mailer. When in doubt, a story will usually fall on the side of YES.*

Annotators may also choose to label a story with an optional **NOT EASY** label. All documents must receive either a **YES** or **NO** relevance label, but if an annotator struggles with a particular relevance judgment, s/he may add the additional **NOT EASY** label, which alerts team leaders of particular difficulties within a topic and triggers additional post-annotation quality control measures.

Search guided annotation implements four distinct stages, all of which utilize the EZQuery search engine. Stage one involves submitting the initial seed story, or a translation thereof, as a search query. The search engine returns a relevance-ranked list of documents and the annotator reads and labels each of these documents as **YES** or **NO**. Annotators keep reading through the list of stories until they've identified 5-10 additional on topic documents, or until they have reached "off topic threshold", defined as a 2:1 ratio of off topic to on topic stories, provided that the last 10 stories read are off-topic. If no additional on topic stories are identified in Stage One, annotators move directly to Stage Three.

In Stage Two, annotators submit all of the on topic stories identified in Stage One as a query to the search engine. They read and label the resulting relevance-ranked list as described above. Although annotators are required to perform only

one query at this stage, they are encouraged to perform multiple iterations in order to identify all of the on topic stories within the corpus. Regardless of how many queries they issue, annotators are obligated to read enough stories to reach the “off topic threshold” before moving to the third stage of annotation.

During Stage Three, annotators issue new queries using text selected from the topic research document and topic explication. As with the earlier stages, they review the relevance-ranked list of documents and reach the “off topic threshold” before progressing to the fourth stage of annotation. During the third stage annotators are required to issue at least one query and reach the “off topic threshold” at least once, but they are encouraged to experiment with different text queries until they have exhausted the fruitfulness of the third stage approach.

In Stage Four, annotators are encouraged to think creatively. By this point they have worked on the topic for at least several hours and as many as a few days. They have become topic experts and are well-positioned to draw on their specialized knowledge to find the remaining on topic documents that have not yet been identified. Annotator instructions during this stage are as follows:

*You are encouraged to use your specialized knowledge (drawn from topic research and the known on-topic stories you've already seen) to conduct additional manual searches through the corpus. These additional searches can be based on keywords, names, particular on-topic stories, etc. You might want to limit your search to stories that occur before, after, or within particular dates; you might want to focus on a particular source; you may think of some entirely new way to scan the corpus for additional stories related to your topic. Think creatively! If you come up with a novel way to search for additional on-topic stories, let us know. If you find additional information (names, places, dates, events) about your topic, revise the topic research page to include the new information.*

Annotation for a topic is complete when an annotator believes that s/he has found all the relevant documents within the corpus. Team leaders observe annotation progress and ensure that an appropriate number of searches have been conducted before confirming that a topic is complete, but the annotator, who by now is an expert on the topic at hand, is ultimately the best judge of a topic's completion point.

### **3.7 Quality Control**

The quality control measures for search guided annotation are multiple, including a precision check, adjudication of sites' results, and dual annotation with discrepancy resolution.

During Precision QC, senior annotators review all stories labeled as **YES** to identify possible false alarms, stories erroneously labeled as on topic. Working with a modified version of the labeling interface and examining one topic at a time, senior annotators read each story and either verify it as on topic, or change it to **NO**. When possible, the precision check is performed by the same senior

annotator who conducted topic research for that topic. During the precision check, annotators keep a sharp eye out for cases of “topic drift”, when the definition of the topic varies across annotators, by language, or over the course of topic labeling. By referring back to the topic explication and rules of interpretation, topic research documentation and annotator e-mail archives discussing the topic, senior annotators exclude stories outside the scope of the topic. Team leaders independently verify all changes resulting from the precision check. Along with reviewing all **YES** documents, senior annotators also review all judgments with the "Not Easy" label during the precision check.

In order to identify misses (on topic stories that are not identified as such), LDC relies on adjudication of research sites' results. NIST provides LDC with each research site's results for the topic tracking task. The sites' systems are scored against the LDC's human-produced topic relevance tables, with the annotators' judgments taken as ground truth. Each system false alarm is a potential LDC miss. It is not feasible to completely adjudicate all cases where LDC annotators differ from system performance; the effort needed to adjudicate all the cases of discrepancy would exceed the original corpus creation effort<sup>4</sup>. Instead, the LDC reviews cases where a majority of systems disagree with the original annotation and modifies the topic labels as required. In previous TDT corpus adjudication efforts, the probability of a system false alarm correlating to an annotator miss grew in proportion to the number of systems reporting disagreement with the original annotation.

As a final quality assurance measure, LDC performs dual annotation on 10% of all topics. Given the nature of search-guided annotation, one cannot derive measures of inter-annotator consistency from dual annotation. The relevance-ranked lists returned by the search engine depend upon their input and prior annotation. Where annotation results differ, normal variation in annotation practice is as likely an explanation as annotator inconsistency. However, LDC does perform dual annotation purely as a form of quality control. Dual annotation involves complete re-annotation of a topic by an independent annotator. It is part of the regular annotation workload, and annotators do not know which topics have been selected for dual annotation. Moreover, the annotators do not know who performed the first round of annotation, and they do not have access to the original annotators' judgments. After both rounds of topic labeling are completed for a particular topic, senior annotators compare the two sets of topic labels, resolve any discrepancies and measure inter-annotator consistency.

---

<sup>4</sup> For example, in the case of the TDT-3 1999 Evaluation topics, NIST delivered results containing approximately 1.5 million topic-story tuples from 7 research sites.

## 4 APPENDIX 1: Examples of selected general story types

*Note: All story boundaries reflect segmentation practices at the time of annotation rather than being indicative of current practices.*

### A. Commercial

19981227 1600 1630 CNN HDL.typ

<sn 729.00>

Okay, you probably think they look alike.

<b 806.00>

But with gateway you get basic help over the phone for as long as you own it.

<b 810.00>

After a month, you pay for it.

<b 811.00>

Here, you get a color printer.

<b 813.00>

Nope, not here.

<b 815.00>

This one you can trade in toward the purchase of a new one after two years.

<b 818.00>

Yours forever.

<b 821.00>

Still think they look alike?

<b 822.00>

We'll build you a PC 99 like this one for \$1,598 or \$45 a month.

<b 827.00>

Call 1-800-gateway and get more out of the box.

### B. Reporter Chit-Chat

19980105\_1130\_1200\_CNN\_HDL.typ

<sr 659.020>

v.w., as they introduce their newest car in some sense, something that you've seen before.

<b 665.00>

the new beetle, 1998 new beetle.

<b 669.00>

sticker price, \$15.2 without the options.

<b 677.00>

it will be produce introduced at the north american international auto show.

<b 682.00>

one of the most prestigious in the world.

<b 685.00>

back to you, lyn.

<t 686.00>

when we were in college, jeff one of the things we liked about the bug it didn't cost much.

<b 691.00>

sounds like this will be more competitively priced?

<t 694.00>

reporter: true.

<b 694.00>

i guess \$15.2 doesn't go as far as it used to back when we were in college.

<b 700.00>

it's as affordable as it can be.

<t 703.00>

thank you very much, jeff, for the live report.

<sn 701.407>

the hour's top stories are two minutes away.

### C. Isolated anchor/program IDs

19980812 1800 1900 VOA ENG.typ

<sn 3148.728>

<turn Curry>

And this is VOA News Now, I'm Neil Curry with Theresa Erikson at 22:53 Universal Time.

### D. Local Station IDs

19981021 2000 2100 PRI TWD.typ

<sn 47.142>

<turn Unknown>

You're listening to 91 FM, WHYI FM Philadelphia, your member supported NPR station serving Pennsylvania, New Jersey, and Delaware.

**E. Program Sponsor Announcements**

19981019 2000 2100 PRI TWD.typ

<sn 233.336>

This in NPR news.

<turn Unknown>

Support for National Public Radio comes from NPR stations and Amgen, discovering effective medicines, fighting disease and improving lives through biotechnology.

<turn Unknown>

Programming on 91 FM is supported by WHYI store of knowledge, offering people of all ages thousands of things to pick up, play with and explore, with locations in Cherry Hill, Willow Grove, the Court at King of Prussia and the Depford Mall.

<turn Unknown>

The World is made possible in part by Merck, pharmaceutical research dedicated to preventing disease and improving health. Merck, committed to bringing out the best in medicine.

**F. Pledge Drives**

19981019 2000 2100 PRI TWD.typ

<sn 475.864>

<turn JOE CASUTTO PUMACO>

Here's the number to call to make a pledge to 91 FM, 215 923-1234, that's 215 923-1234. Stick around. We've got the World coming up in just a moment. Hello, I'm Joe Casutto Pumaco reminding you that we're in the middle of our October membership drive at 91 FM and we want to have your membership pledge now on the line at 215 923-1234. You make the programs possible on this noncommercial public radio station. Pledge now, 215 923-1234. Thanks very much for your support.

**G. local forecast**

19981002 2000 2100 PRI TWD.typ

<sn 183.665>

<turn Unknown>

Good evening from 91 FM, WHYI. Here's the forecast from the Franklin Institute. Tonight, mostly clear and cool, low 46, and in the upper 30s in some northern and western suburbs. Saturday, sunshine, then increasing afternoon clouds, high 68. Sunday, clouds and some sun, a few showers are possible. High again, 68.

<b 200.0>

Coming up on the program, THE WORLD, we'll hear about fashion designers and companies which use their popularity to promote political and social causes. Also, the daily geography quiz. You're tuned to THE WORLD on 91 FM, WHYI.

**H. Public Service Announcement**

20001119 2000 2100 VOA ENG.typ

<sn 3374.204>

Next, an international public service announcement.

<t 3595.8>

On June 21st 1998, Elana Gonzalez was reported missing from her home in Englewood, New Jersey. Four days later, Gonzalez's former boyfriend Freddie Bostamante was charged with her murder. That same month, Bostamante fled to Guatemala. Freddie Bostamante is an Hispanic male born on April 17th 1966 in Guatemala. He is one meter 75 cm tall and weighs 77 kilograms. He has black hair and brown eyes. Bostamante was a resident-alien of the United States. His alien registration number is 042442161. He speaks Spanish and English. Bostamante has worked in construction. He may be dangerous and suicidal. If you have any information concerning Bostamante, you should contact the nearest US embassy or Consulate or call Detective Brian Cullinan at the Burgeon County New Jersey prosecutor's office, Homicide squad, at 201-646-2300. All reports will be investigated and the identities of all informants will be kept confidential.

<t 3599.8>

That was an international public service announcement. For more information on international fugitives, visit the international crime alert web site at [www.ibt.gov/fugitives](http://www.ibt.gov/fugitives).

**I. Promo for upcoming programs**

19981026 1830 1900 ABC WNT.typ

<sn 1684.025>

Tomorrow we'll take a look at getting to space on the cheap.

<b 1689.00>

That is our report on "World News Tonight." later on "Nightline," why the Marines are fighting to save a struggling company.

<b 1697.00>

I'm Peter Jennings.

<b 1698.00>

Have a good evening.

<b 1700.00>

**J. Lists of temperatures/weather conditions**

VOA19980111.2300.1200

<sn 1200.042>

<turn ANNOUNCER HENSON>

Thanks, Les. Les Carpenter will have a full 20-minute business report right after the news at zero hour universal time on many of these same frequencies and satellite channels. It is 2320 universal time.

Taking a look at the weather for some cities in our region. Perth sunny, breezy and hot, the high 33. Wellington mostly cloudy, breezy and cool, high 16. Hong Kong mostly cloudy, high 20. Shanghai periods of clouds and sunshine, high 8. Warsaw partly sunny, high 8. Beijing snow and cold, high minus 1. Hanoi mostly cloudy, high 21. And Seoul, cloudy, high 1. The weather in Berlin should be pleasant. The highs should be about 10. Cairo partly sunny, windy, and cool, high about 10. And here in Washington, D.C., back to seasonable January weather after a week of uncommonly warm and wet days. It was sunny and cool, high of 8 on Sunday. Monday, cloudy with a high of 10. The weather forecasts are a service of VOA today. Accuweather prepares them.

**K. Lists of sports scores/results**

19980311 1130 1200 CNN HDL.typ

<sn 1384.197>

On the scoreboard -- sabres and the Isles in a deadlock.

<b 1386.00>

And the kings with a four-goal first.

<b 1388.00>

<sn 1387.916>

NBA -- houston wins.

<b 1390.00>

So does Washington.

<b 1391.00>

Antonio McDyess with 19 for Denver.

<b 1392.00>

Anthony Keith James, "CNN Headline Sports."

<b 1394.00>

**L. Un(der)transcribed sections**

19981110 1830 1900 NBC NNW.typ

<su 678.924>

On Wall Street tonight --

**M. Previews**

< 2 clauses

19981006 1600 1630 CNN HDL.typ

<sn 1642.509>

Coming up -- President Clinton is calling on world financial leaders to take urgent steps to stop the global economic crisis from spreading. That and more in two minutes.

**≥ 2 clauses**

19981123\_1600\_1630\_CNN\_HDL.typ

[<sn 491.983>](#)

The extended national weather forecast is next.

<b 494.00>

And then -- it has been one month since the shooting death of a Buffalo, New York, doctor who performed abortions.

<b 499.00>

We'll have an update on the murder investigation.

<b 503.00>

**All that and more just ahead.**

**N. Top news stories (recaps)**

**< 2 clauses**

19981027\_1130\_1200\_CNN\_HDL.typ

[<sr 929.210>](#)

Checking our top stories -- the center of powerful hurricane Mitch is 90 miles north of the coast of Honduras, moving west at six miles an hour.

<b 939.00>

Hurricane warnings are in effect for parts of Mexico's Yucatan peninsula and for coastal Belize, Guatemala and Honduras.

[<sr 944.130>](#)

Thousands of Yugoslavian military and police forces are leaving Kosovo just hours ahead of a U.N. deadline.

<b 951.00>

A White House spokesman says over 90% of the Serbian security forces have left Kosovo or returned to their garrisons.

[<sr 958.059>](#)

Ailing Russian President Boris Yeltsin has checked into a rest home near moscow.

<b 964.00>

He's reportedly there to recover from high blood pressure and exhaustion.

[<sn 966.099>](#)

These stories and more in 15 minutes.

**O. Weather reports**

19980818\_1130\_1200\_CNN\_HDL.typ

[<sr 492.998>](#)

I'm meteorologist Dave Hennen with your "Headline News weather."

<b 495.00>

It's going to heat up in parts of the plains today, Kansas and back in through Oklahoma will see temperatures around 100 or a little bit above.

<b 502.00>

Os will be widespread from the northern plains, back into the southeast, as the cooler weather continues to filter into the northeast.

<b 510.00>

The forecast could see some rain in the northeast again after heavy rains yesterday in the northeast.

<b 514.00>

By tomorrow, though, high pressure will be building in, bringing some drier weather.

<b 517.00>

Scattered showers and storms will continue along the gulf coast.

<b 521.00>

And back through the upper midwest, we'll see more in the way of showers and thunderstorms right on through tomorrow.

<b 527.00>

The extended forecast, on Thursday, look for warm temperatures to continue in the southwest, and back through the southern plains.

<b 533.00>

It's going to be cool in the northeast, though.

<b 535.00>

Scattered showers and storms most likely along the gulf coast, right on into Friday.

<b 553.00>

For more weather information, you can log into CNN.com for the four-day forecast for over 6,100 cities.

**P. Sports reports**

19980207 1830 1900 ABC WNT.fdcch.typ

<sr 1262.364>

<turn JOHN\_FRANK, ABC\_News>

Thanks, Aaron. Well, it is finally here -- opening day of the winter Olympics in Nagano, Japan. But the competition began years ago with the battle to host the games. Nagano won that battle but as today's events reveal, the town may have gotten more than it bargained for. Here's ABC's Mark Litke.

<turn MARK\_LITKE, ABC\_News>

They've been preparing for six years, but only now are the people of Nagano feeling the full impact of the Olympics games. This quiet, traditional town, best known for its temples and apples, is suddenly bursting at the seams with more dignitaries, more foreign visitors, more cultural oddities than it's ever witnessed before. "We're just simple country people," he says, "so we're amazed with all this." Amazed but also proud. The opening ceremony was rich with their local culture and tradition. Millions around the world watched the parade of athletes in their new stadium, watched the Olympic flame being lit by one of their young champions. But even with all the pride and excitement, the famed hospitality and patience of this city will be put to a severe test in the days ahead. Heavy security is already having a big impact here. Roads are clogged, traffic diverted, nerves frayed. Olympics commercialism has overtaken the streets as official sponsors vie for attention. And today, the first anti-Olympic protest march, criticizing the high cost of the games, billions spent despite the country's shaky economy. Even some who wanted the Olympics are upset because they couldn't get tickets for popular events. The closest many got to the opening ceremony -- these banks of TV screens set up around the city. Of course, grumbling aloud is considered impolite in Japan, so most Nagano-ites will simply endure all the inconveniences and hope that once the games are over, their city won't have lost any of its quiet, traditional charm. Mark Litke, ABC News, Nagano.

**Q. Stock reports**

19980202 2100 2200 VOA WRP.typ

<sr 1586.534>

<turn ANNOUNCER\_ARTERY> Well, we had a spectacular day. The Dow Jones Industrial Average closed at 8,107, that was up 201 points about two and a half percent. And the Standard and Poor's 500 Index closed at a record high 1,001, that was up 21 points. Analysts say there were there main reasons for the rally. One was the strong rebound in Asian stock markets. That rebound spilled into Europe and then on here to New York. And second, many traders now believe that not much will come of the latest scandal allegations against President Clinton. And finally, there was genuine euphoria over the proposed merger of Glaxo Wellcome and Smith Kline Beecham, two British-based drug companies. That combination in a deal worth almost 70 billion dollars would create the largest pharmaceutical company in the world.

<turn ANNOUNCER\_CLARK> Breck, this is Susan Clark. Let me just, if I could just ask you a question. Conventional wisdom in the past couple of weeks with the problems on currency markets and stock markets around the world has been that a lot of money moves to drug company stocks because drug companies still have a product to sell that people need no matter how poor they are. Does this have a lot to do with today's pattern and today's results or not much?

<turn ANNOUNCER\_ARTERY> Not with today's. You are right that money has been moving into pharmaceutical stocks for that reason. Today's activity though, was on speculations, speculation that there will be further take overs and mergers in the industry. The cost of research is very high and its much more efficient to have a larger company than a smaller one. It's worth noting too, that even this combined Glaxo-Smith Kline, if it comes about, will still have less than eight percent of the world pharmaceutical market, so obviously there is plenty of room for further consolidation. And a couple of other brief items, the Governors of the U.S. Central Bank start a two-day meeting tomorrow, but analysts are unanimous in saying the Central Bank is going to leave short term interest rates alone. And the Government reported construction spending in the U.S. rose by one-tenth of a percent in December, although spending on residential construction soared by 1.4 percent. This is Breck Artery.

<turn FEMALE\_ANNOUNCER> Thanks very much, Breck, appreciate it.

**R. Program updates, listener comments, listener question & answer**

19980522\_1800\_1900\_VOA\_TDY.typ

<sn 629.288>

Let's begin with singer/songwriter Pete Seeger's song about times and seasons. From 1962, Pete Seeger with a song inspired by an ancient Biblical text, "Turn, turn, turn - To Everything there is a Season." A song about time and seasons and changes as we begin this farewell edition of VOA Saturday coming to you from Washington on the Voice of America. That's because, if you haven't already heard, beginning at zero hour universal time next Friday, May 29, VOA's English language service will undergo a radical transformation.

<turn ANNOUNCER\_ERICKSON>

All music programs broadcast in English will be eliminated. Instead, music programming will only be heard by listeners who can tune in local radio stations that broadcast VOA's satellite transmissions. Special English broadcasts will continue to be heard on shortwave, but on different frequencies from the ones now in use. And most of the remaining VOA English programs, like VOA Today, VOA Saturday, and VOA Sunday, as well as stateside Studio 38, Critic's Choice, and New Horizons will no longer be heard.

**S. Interview segments**

19981211\_2100\_2200\_MNB\_NBW.typ

<sr 1022.077>

With us this evening from Washington, Alan baron, who was chief minority counsel during the Senate campaign finance investigation.

<b 1031.00>

He also served as special impeachment counsel of the house at a time when two federal judges were impeached.

<b 1036.00>

With us from our MSNBC studios, Jay severin republican strategist, MSNBC political analyst. Jay, I'd like to begin with you.

<b 1045.00>

Remember the PrimeTime speech the president gave, his first apology. He was no sooner finished with today's speech, and it was said again about the following quotes -- I have been condemned by my accusers with harsh words. It's hard to hear yourself called deceitful and manipulative. The charge was that what he couldn't help was the part of him that wanted to make this about him instead of falling on his sword.

<t 1072.00>

I think that's right, Brian.

<b 1074.00>

Keep in mind the context of this is this is a presidential address. This is something that the power of the presidency has that no one else has. You should husband it and use it wisely. It was used today I think in a blunder. It was use NT a way that was not carefully crafted. Couldn't conceivably have helped any republicans move his way, perhaps quite the opposite. And here we are only hours later, not asking how well a presidential strike like this worked, but why and what are the reasons it hasn't worked because I believe that was the immediate consensus. This lacked the elements that were required to do the only reason you do it, which is to move people to you. I think with one not so deft move the president may have wasted some equity here, made it harder for republicans to come over to him, and maybe even harder for democrats to defend him.

<t 1128.00>

But Alan baron, somewhat amazingly there is already talk tonight that, well, this time didn't quite work, perhaps we'll tweak the message and try again when he lands in the Middle East. Is that believable?

<t 1141.00>

I think the president will continue to try to persuade that middle group, those moderate republicans that seem to exist in at least people talk about that group ooze - as existing to try to persuade them that this is really the wrong thing, that he really is contrite, that he really acknowledges his wrongdoing, and that we should not dumb down the impeachment process in order to make the point that he did something wrong.

<b 1169.00>

He's acknowledged that. I think he will continue to try to do it. I'm not sure anything's going to work at this point. The 800-pound gorilla is out of the cage and nobody knows how to stop it.

## 5 APPENDIX 2: Examples of selected source-specific story types

*Note: All story boundaries reflect segmentation practices at the time of annotation rather than being indicative of current practices.*

### A. PRI GeoQuiz Part 1

19981022 2000 2100 PRI TWD.typ

<sr 2374.503>

We are Africa bound for today's GeoQuiz and scaling the heights of the Sahara as we truck along a mountain range in the northwest corner of Chad. This range is might imposing with its highest peak rising more than 11,000 feet above the desert floor. But it's the ancient paintings and not the peaks that have made these mountains famous. Prehistoric rock paintings and carvings were discovered here. Scenes of elephants and hippopotami grazing on open plains have intrigued scientists for decades. UFO fanatics get excited about them, too, since some human figures are dressed in what sure do look like space suits. To get here safely, you may need more than a space suit. Visitors have to battle temperatures as high as 120 degrees, and the local inhabitants work hard to keep travelers out. So what is the name of Chad's mountain range that seems utterly alien to this world? The answer is just around the corner.

### B. PRI GeoQuiz Part 2

19981022 2000 2100 PRI TWD.typ

<sr 3009.135>

Time now for the answer to our Geography Quiz today. We're in Northwest Chad looking for the highest mountain range in the Sahara Desert. This is a place of volcanoes, hot springs, and ancient art depicting elephants and other water loving animals. The answer is the Tibesti Mountain Range.

### C. PRI Musical Segment

19981016 2000 2100 PRI TWD.typ

<sr 3026.719>

<turn Cahn>

The U.S. trade embargo against Cuba has gotten a bit looser during the past few years. And as a result, there's more native Cuban music to be listened to here in the states. A lot of it sounds familiar. As "The World's" Marco Worman tells us, that's because many of today's Cuban musicians are finding the old ways to their liking.

<turn World>

Tourist wandering down to New Orleans eat up Dixie Land jazz with a spoon. Old jazz musicians and even young ones who keep tunes like "Basin Street Blues" alive thrive on that Nostalgia. The same longing for the old tunes has emerged for Cuban music. Jesus Alimani, the trumpeter and leader of Cubanismo says he and his bandmates have been baffled by their acclaim in the west.

<turn Cubanismo>

Well, with even the -- because who would believe that at the end of the century, we're going to be playing the -- the music that was created in the beginning.

<turn Worman>

During the band's two tours of the United States this year, Americans flocked to dance and listen. But don't think that all Cuban music sounds like Cubanismo.

<turn Peter\_Watras>

It's really great music, but it's not particularly the music that the majority in Cuba listen to.

<turn Worman>

Peter Watras writes about music for The New York Times.

<turn Watras>

It's doesn't mean that it's not good music, either. Just that in Havana, especially, the public is really unforgiving about -- about falling out of date. And so, you have a pile of musicians who were working on -- have essentially forgotten there. I mean, just the same way that, you know, that happens here in the United States with pop music, too.

<turn Worman>

But Cubanismo isn't wheeling out some lame cha, cha, cha that you might see Bugs Bunny dancing to in an old cartoon. The music is complex and, at times, digs deep in to Jesus Alimani's roots of Congolese animism. Alimani belongs to the Brotherhood of Abaqua, an African tradition that influences several of the numbers in Cubanismo's repertoire.

<turn Alimani>

Well, it's basically music that is more understandable for the Cubans because it's something that is a really strong part of our history, of our religion. And of course, it's something that has been really close to me because it's something that I've been watching since I was born, because all the -- where I used to live, we go to, like, three or four different brotherhoods.

<turn Alimani>

The response, obviously, is very positive because people can dance and can sing with us and so that's basically what we -- have in mind before we did the recording -- like, -- people dancing with this music.

<turn Worman>

Tours of Israel, Turkey, Lebanon and the United States this year, alone, mean Jesus Alimani and the bad have made people dance. In fact, it's hard not to, when Cubanismo is playing. For "The World," this is Marco Worman.

#### **D. VOA Communications World**

##### **News story example**

<sn 1981.799>

From the Voice of America, Communications World, a review of electronic media with Kim Andrew Elliot.

<t 1989.7>

Welcome to Communications World for the weekend of December 16th 2000. Some people believe that Short Wave is dead. However, a company in Florida is now using Short Wave to keep in touch with trucks traveling around the United States. I'll have an interview and in the audience participation section, a brief history of non-commercial radio in the United States. We begin with media news.

<sr 2021.982>

The Supreme Court of the Russian Republic of Dagastan has upheld the lower courts guilty verdict against radio free Europe Radio Liberty reporter, Andre Babitski. Mr. Babitski was convicted of carrying a false passport. Arifi Arial President Thomas Dine, said the Court decision is a clear attack on media freedom. He said Arifi Arial will support Mr. Babitski's appeal to the Russian Supreme Court. In February Mr. Babitski was arrested in Dagastan for carrying false documents which Arifi Arial says were planted on him. He was fined 300 dollars but that fine was suspended under an amnesty agreement.

##### **Non-news example**

<sn 2975.755>

Let's begin this section with a brief meeting of the Communications World family to discuss our plans for the next few weeks. The holiday season will bring a lot of few programming changes at VOA News Now and these will affect this program. On December 24th and the 31st special broadcasts of radio plays on VOA News Now will preempt Communications World on VOA News Now at 0533 and 1333 Universal time. However, on Monday January 1st New Year's Day I will have two live shows on VOA News Now, during which I will call Communications World listeners. These will be at 1433 to 1458 and 2133 to 2158 Universal time. If you would like me to call you during either of these broadcasts, please let me know. Include your phone number and whether you wish to be called during the 1433 or 2133 broadcast. Thanks to those of you who have already volunteered. I hope some additional transmitters will be on the air so that the coverage area of these broadcasts can be expanded. On Saturday January 6th another radio play, so Communications World will be preempted again at 0533 and 1333. This will be the week that we feature the best of New Year's Eve radio broadcasts, so I will again try to get some supplemental frequencies. International and domestic radio around the world can be very interesting and amusing on New Year's Eve. It is a tradition on Communications World to play excerpts of some of these broadcasts on the Saturday after New Year's and now listeners can send excerpts of this audio to me as streamed audio files attached to emails.

#### **E. VOA This Date in History**

19981015 1700 1759 VOA ENG.typ

<sn 1676.462>

CROSBY: It's Thursday, October 15, 1998 and on this date in 1892, mechanical voting machines were used for the first time in an election in the United States. It happened in the town of Lockport, New York. Mechanical voting machines have been developed to eliminate, as much as possible, fraud, error, and carelessness on the part of voters and election officials. The machines also sped the process of counting votes. And on this date in history on October 15, 1950, President Harry Truman met with General Douglas MacArthur at Wake Island in the Pacific. They were there to discuss U.S. policies in the Korean War. The president called the meeting after General MacArthur, the commander of UN forces in Korea, persisted in issuing statements about the conduct of the Korean War that contradicted Mr. Truman's stated policies. As the constitutional Commander-In-Chief, the

president believed he had the authority to reestablish controls over General MacArthur, a very proud man, who was once one of the most popular military leaders of the time. Mr. Truman and the General had left Wake Island in agreement, but that did not seem to be the way it worked out.

#### **F. VOA Internet News**

20001012 2100 2200 VOA ENG.typ

<sn 1656.482>

It's coming up on 01:28 Universal time and that means many of our local FM and medium wave affiliates are about to take a two-minute break. From Washington you are listening to the Voice of America.

<b 1687.889>

Where can book lovers find the ultimate Sherlock Holmes fan club? Why elementary, my dear Watson, on the web of course. I am Charles Bowen with the Internet news. Today's report, a site called Sherlockian.net, a portal to everything the web has to offer about the world's greatest detective Sherlock Holmes. Maintained as a labor of love since 1994 by web master Chris Redman, this is your ultimate resource for all you ever wanted to know about our master of deduction. Categories include the original Sherlock Holmes stories, writer Arthur Conan Doyle, major Sherlockian sites, actors and films, libraries and books, even parodies. To link directly to this site visit the Internet news homepage at netnewstoday.com on the worldwide web. The Internet news is a service of the George Washington University in Washington DC, an International Center Of Learning providing educational opportunities to meet today's global challenges. This is Charles Bowen for the Voice of America.

#### **G. CNN Cold & Flu Report**

19981008 1600 1630 CNN HDL.typ

<sn 761.498>

Welcome to another year of CNN's "cold and flu report."

<b 766.00>

With the season of sniffles and sneezes starting up again, CNN will bring you a new report each week about colds, flu and the latest news on how to deal with them.

<b 774.00>

October and November are the months to get your yearly flu vaccine.

<t 779.00>

Typically, these vaccines consist of killed influenza viruses.

<b 783.00>

And when you get injected with them, your body develops protective antibodies against influenza.

<t 789.00>

Reporter: Flu vaccines don't provide perfect protection, but they do greatly reduce the odds of catching the flu.

#### **H. CNN Play of the Day**

19980106 1130 1200 CNN HDL.typ

<sn 1335.261>

IT'S THE SPURS AND MAGIC FROM THE O-RENA.

<b 1345.00>

WITH TIME TICKING AWAY IN THE THIRD, ORLANDO'S DERRICK HARPER PUTS UP AND NAILS THE PRAYER FROM JUST INSIDE HALF-COURT.

<b 1349.00>

HARPER LOVES IT, SO DOES RONNIE SIKLIE.

<b 1350.00>

ANOTHER LOOK AS HARPER DIALS LONG DISTANCE AND CONNECTS WITH THE "PLAY OF THE DAY."

#### **I. MNB newspaper headline review**

19981202 2100 2200 MNB NBW.typ

<sn 3227.671>

Our first look at the morning newspaper headlines.

<t 3373.984>

Mike T: the wealthiest man in the world today gave away a small portion of his sizable personal wealth, something he's vowed to do more of as he grows older. Microsoft chairman Bill Gates and his wife Melinda announced they will give away more than \$100 million to help immune Yzerman children around the world. It's a cost effective way to go about health care. Could reduce childhood deaths by 1/3. Microsoft is a part owner of MSNBC.

<b 3399.389>

Mike Espy is on the front page of most newspapers tomorrow morning. The house speaker and other republican leaders in the house reaffirmed their support for the embattled Henry Hyde and said they are prepared to watch this thing go past January if need be.

<b 3426.00>

"Dallas morning news," it has happened so many times it's lamb cliché. Violence yet again derailing the peace process in the Middle East. This time the West Bank withdrawal. And this was why. An Israeli soldier pulled from his car, beaten with stones by a Palestinian mob. This will not please the prime minister of Israel. There is increased trouble, and now increased tension.

<b 3449.00>

"U.S.A. today" instead of waiting for something terrible to happen, airlines are going to download data that in effect spells out every moment, every minute of an airline flight while the flight is in progress, and afterwards will download after landing. They can look back over the data to see what if anything went wrong and any indications of future trouble.

<b 3471.00>

Also "Miami Herald" a big makeover plan in the White House. Of course; it could take 20 years like everything else in Washington. \$300 million. The plans for the White House, adding a recroom underground parking, a visitor's center, and a new briefing room. The current one having been in disrepair for some time. It is home to rats and, of course, other journalists.