Quality Control in Large Annotation Projects Involving Multiple Judges: The Case of the TDT Corpora

Stephanie Strassel, David Graff, Nii Martey, Christopher Cieri

Linguistic Data Consortium
3615 Market Street, Philadelphia, PA 19104, USA
{strassel, graff, nmartey, ccieri}@ldc.upenn.edu

Abstract

The Linguistic Data Consortium at the University of Pennsylvania has recently been engaged in the creation of large-scale annotated corpora of broadcast news materials in support of the ongoing Topic Detection and Tracking (TDT) research project. The TDT corpora were designed to support three basic research tasks: segmentation, topic detection, and topic tracking in newswire, television and radio sources from English and Mandarin Chinese. The most recent TDT corpus, TDT3, added two tasks, story link and first story detection. Annotation of the TDT corpora involved a large staff of annotators who produced millions of human judgements. As with any large corpus creation effort, quality assurance and inter-annotator consistency were a major concern. This paper reports the quality control measures adopted by the LDC during the creation of the TDT corpora, presents techniques that were utilized to evaluate and improve the consistency of human annotators for all annotation tasks, and discusses aspects of project administration that were designed to enhance annotation consistency.

1. Introduction

This paper will review the annotation effort that supported the development of the TDT2 and TDT3 corpora at the Linguistic Data Consortium, with particular emphasis on the quality control measures that were employed for each annotation task. In addition, the paper reviews general approaches to quality control that were adopted as part of the overall project design and project administration.

2. The TDT corpora

This section provides a brief overview of the content of the Topic Detection and Tracking (TDT) corpora. For a more detailed report, see Cieri et al. (2000). A discussion of the TDT research tasks and their evaluation appears in Wayne (2000).

The TDT corpora were created to supposrt the development of automatic techniques for detecting and tracking topically related material from a continuous stream of newswire or speech data.

The TDT2 corpus was designed to support three research tasks: story segmentation, topic detection and topic tracking. The data for TDT2 included two languages (originally just English; Mandarin was added later) and nine sources (television, radio and newswire), drawn on a daily basis for a six-month period from January through June, 1998. There were two annotation tasks: segmentation and topic-story labeling. During most of 1998, LDC annotators worked to segment hundreds of hours of audio and to label thousands of individual news stories.

TDT3 expanded on the TDT2 corpus, adding two new research tasks, story link detection and first story detection. The data was from a shorter time period (October through December 1998), but included a total of eleven news sources, and new kinds of annotation were added to support the new research tasks. The TDT3 annotation effort took place during the first three quarters of 1999.

The complex nature of the TDT corpora – their data, annotation tasks and staffing requirements – made inter-annotator consistency high priority, but a big challenge as well. Each corpus required a staff of 25-30 annotators and a management team of 3 fulltime staff members, plus countless hours of technical and administrative support. The multilingual nature of the project presented additional challenges for maintaining annotation quality and consistency. The remainder of this paper discusses each annotation task and the accompanying quality control procedures in turn, then presents a discussion of some general practices adopted to enhance inter-annotator consistency during corpus creation.

3. Some TDT basics

The TDT project is concerned with stories, events and topics. All research and annotation tasks relate to these concepts. Topic is defined in a very particular way for TDT research. For the purposes of TDT, a **topic** was defined as an event or activity, along with all directly related events and activities. A set of 100 topics, identified from a set of randomly-selected English seed stories, was chosen for the TDT2 corpus. Sixty topics were selected for TDT3 from both Mandarin and English seeds. Each topic has at its root a seminal event. While the more common vision of topic might be something like "hurricanes", a TDT topic would be limited to a specific hurricane-event, e.g. Hurricane Mitch. Within TDT, an **event** is something that happens at some specific time and place, along with its unavoidable consequences. An individual TDT broadcast news recording or newswire file consists of approximately 20 stories. A **story** as defined by TDT is a newswire article or a segment of a news broadcast with a coherent news focus. This particular construction of the concept of topic was a critical component of TDT annotation, as it allowed annotators to (potentially) identify *all* the stories in the corpus that discussed some pre-defined topic. The quality control measures that were adopted to support the TDT annotation tasks would not have been viable without these established definitions of topic, event and story.

4. Segmentation

4.1. The annotation task

LDC annotators produced the reference segmentation of broadcast news sources against which the evaluation systems were scored. For the most part, segmentation was a two-pass procedure. Annotators listened to the broadcast audio with the audio waveform and text on display. The text that corresponded to the broadcast audio was captured from closed-captioning when possible; otherwise the audio was transcribed by a professional transcription service.

In the first 2*RealTime pass, annotators assessed existing story boundaries in the transcripts and closed caption files, and added, deleted or moved story boundaries as necessary. In addition, annotators set and confirmed timestamps for all story boundaries. Finally, annotators classified all story units as either "news" or "not news". Non-news story units were things like commercials and long sections of reporter chit-chat.

In the second 1.5*RealTime pass, annotators again listened to the entire audio file, and confirmed or adjusted existing story boundaries.

In addition, approximately 1% of all files were randomly selected for spot-checking by team leaders, who then corrected any errors and reported recurring problems back to the annotation staff.

4.2. Quality assurance measures

The most fundamental aspect of quality control for the segmentation task was the execution of the two complete, independent passes over the data. In the early stages of the TDT2 annotation effort, the second pass was not an exhaustive review of the entire audio file. Instead, annotators simply re-examined the story boundaries established by the first annotator and checked the existing timestamps. Because annotators were not listening to the entire broadcast, there was no way to catch story boundaries that had been missed by the first annotator. This was a serious problem for the closed-caption sources because some stories were only partially displayed in the text file and were therefore easily missed. Mid-way through the TDT2 project, exhaustive second passing was adopted and segmentation accuracy increased accordingly.

By having two separate annotators review the story boundaries and timestamps for each audio file, it was possible to minimize the occurrence of missed story boundaries, misclassifications of story content and inaccurate timestamps. Annotators conducting a second pass were familiar with the peculiarities of each source that if not examined closely might lead to errors.

The implementation of two complete passes was costly in terms of human effort: segmentation comprised a full quarter of effort needed to fully annotate the TDT3 corpus. However, the need for accurate and complete story boundaries and timestamps warranted this expenditure of labor.

In addition to the complete second passing of all segmentation files, all stories that were rejected as non-news during first and/or second pass segmentation were later reviewed by another annotator, and either confirmed or vetoed. Corrections were made where possible and stories returned to the pipeline for further annotation.

The final quality assurance measure adopted for the segmentation task was dual segmentation. An additional 5% of all broadcast files, balanced across sources and dates, were identified for dual segmentation by independent annotators. These stories received another complete first and second pass, and any resulting discrepancies were reconciled by team leaders.

The results of dual segmentation showed high rates of consistency among annotators. The files selected for dual segmentation contained a total of 1300 story boundaries. Of these, 203 displayed discrepancies between the two sets of annotations. These discrepancies took three forms. Over half of the discrepancies were the result of an insignificant "stylistic" difference in segmentation – for instance, whether to include brief reporter chit-chat at the end of one story or the beginning of the next.

Fifteen percent of the annotator discrepancies resulted from "judgement calls". Segmentation is not an exact science, and there is some ambiguity in the task. 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 were instructed to rely on audio cues (speaker changes, music, pauses) to inform their judgements, but some level of indeterminacy still existed.

Only a quarter of the segmentation discrepancies were the result of a distinct error on the part of one of the annotators (a missed story boundary or inaccurate timestamp). Human annotators compared favorably to system performance on the segmentation task when scored by NIST's evaluation software for both the TDT2 and TDT3 corpora.

5. Topic Labeling

5.1. The annotation Task

The vast majority of annotator effort in both TDT corpora went into topic labeling. This annotation task alone comprised a third of all annotator effort during TDT3. Using a custom-designed interface, annotation staff worked with the daily news files, reading each story sequentially and deciding how it related to the corpus topics. For each topic-story pair the annotator could render a decision of *yes*, *brief* or *no*, meaning that story was about the topic, mentioned the topic only briefly or was not about the topic. Any mention of a topic warranted a label of at least *brief*. Stories that were primarily about something else but discussed the target topic in more than 10% of their volume were labeled *yes*. This was in keeping with the premise that news stories could be about more than one topic. Annotators could also reject a story as non-news or as exhibiting some data formatting problem. A comment field allowed annotators to record their analysis of problem stories, or to note questions about a news report (e.g., "Is this one or two stories?"). The interface also allowed annotators to review their work and make changes. The labeling interface is shown below.

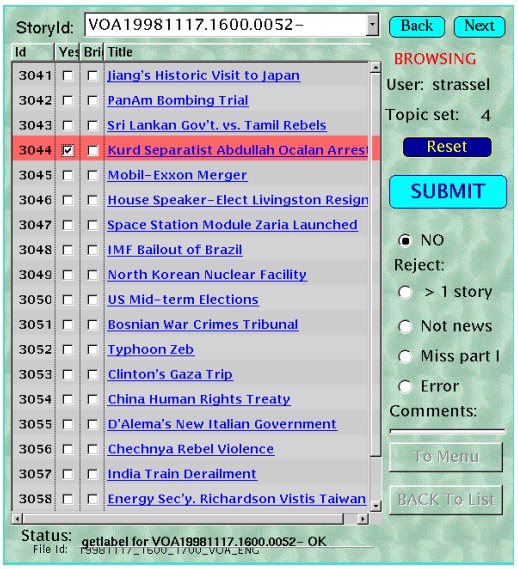


Figure 1: TDT Labeling Interface

For TDT2, LDC staff made five passes over the data. In each pass, annotators labeled a story with respect to 20 topics on average, resulting in a total of 100 topics. TDT3 consisted of 3 lists of 20 topics each, for a total of 60 topics. Before beginning each session of topic labeling, annotators were required to read through the relevant list of topic definitions (see section 8.3 for discussion of the topic lists).

The TDT custom labeling interface stepped annotation staff through the stories, recorded each annotator's progress and logged their decisions into an Oracle database. TDT2 annotators were encouraged to become "source specialists", working primarily with a single source of data. This strategy was adopted to enhance efficiency, since each source has it own peculiarities and it was believed that annotators would work more easily with a familiar source. This strategy had an unforeseen consequence, however. Certain topics are more prevalent in some sources than others, and annotators began to anticipate which topics were likely to come up in the source they were covering. When a topic was covered by an unlikely source, annotators frequently missed the story – not because they didn't read the story or understand its relevance to the topic, but because they'd forgotten the topic was on their list. Although later quality control passes did capture these missed stories, this strategy was abandoned in TDT3 in favor of automatic file assignment via the annotation interface, thus ensuring that all annotators worked with all sources during the course of the project. This automatic file assignment had the added advantage of alleviating some of the administrative burden of work assignment for the fulltime project managers.

5.2. Quality assurance measures

5.2.1. Dual annotation and discrepancy resolution

For both TDT2 and TDT3, between 5% and 8% of all news files received a complete second annotation by an independent annotator. During TDT2, this process required project managers to hand-select files for dual annotation and reassign them manually to independent annotators. Although the annotation staff were not told that the files had already been annotated, because of the timing and style of the assignment of these files, they often suspected that they had. The annotators did not know who had done the first round of annotation, and they did not have access to the original annotators' judgements. However, dual annotation file assignment was single-blind at best.

In TDT3, the automated file assignment via the topic labeling interface allowed for a double-blind assignment of dual annotation files. No one, not even project managers, knew which files had been selected for dual annotation or who they had been assigned to. The dual annotation was completely incorporated into the regular distribution of work. After topic labeling had been completed for a particular list, discrepancies between the two sets of topic labels were reviewed and inter-annotator consistency was measured. The topic labeling interface allowed team leaders to act as "fiat", checking discrepancies and correcting any errors. The kappa statistic was used to measure consistency of human annotation. Where a kappa of .6 indicates marginal consistency and .7 measures good consistency, kappa scores on TDT2 were routinely in the range of .59 to .89. Scores for TDT3 ranged from .72 to .86.

5.2.2. Precision

In precision, senior annotators reviewed all stories labeled as *yes* or *brief* to identify false alarms (stories erroneously labeled as on-topic). Working with a modified version of the labeling interface and examining one topic at a time, annotators read each story and either verified it as on-topic, or vetoed it. When possible, the precision check was performed by the same annotator who had conducted topic research for that topic (see section 8.3 below). During the precision check, annotators kept a sharp eye out for cases of "topic drift". This occurred when the definition of the topic did not remain constant for all annotators throughout the course of topic labeling. By referring back to the topic explication and rules of interpretation, topic research documentation and email archives discussing the topic (section 8.3 below), senior annotators were able to establish the proper scope of the topic and to exclude stories that should not have been included. All changes resulting from the precision check were then independently verified by team leaders.

The results of the precision pass showed a very low level of false alarms. For TDT3, the precision check resulted in a veto of only 2.5% of the original annotator judgements (213 of 8570 on-topic stories).

5.2.3. Recall and limited recall

In recall QC, senior annotators use a search engine to generate a relevance-ordered projection of the corpus with respect to a single topic. Queries used in recall QC could be the seminal article, a list of miscellaneous keywords, the topic explication itself or the union of all stories labeled as related to the target topic or a subset thereof.

For the TDT2 corpus, LDC staff conducted an exhaustive recall check over all 100 topics. The search engine returned a list of 1000 relevance-ranked stories. Annotators were required to skim through each story on the list of 1000 to find any potential misses. After reading 50 consecutive off-topic stories, annotators could evaluate

whether there were likely to be any on-topic stories further down the relevance-ranked list. If the 50 consecutive off-topic stories not even close to being on topic, annotators were permitted to move on to the next topic.

The amount of labor expended during the exhaustive recall check for TDT2 yielded a relatively small return, and for TDT3 it was decided that a non-exhaustive, limited recall would be performed. Rather than searching across the entire corpus to find missed stories, the search engine was employed to identify possible misses prior to the (chronologically) first on-topic story, since systems used the first four stories as training data for the topic tracking task.

The results of the recall check for both TDT2 and TDT3 found, not surprisingly, that the rate of misses is higher than the rate of false alarms for human annotators. False alarms were easily caught and corrected since the ratio of on-topic stories was relatively low. The total number of stories in each corpus, on the other hand, was very high (approximately 75,000 for TDT2 and 43,000 for TDT3). For some topics, the "hit rate" was extremely low (in TDT2, some topics had no hits at all; in TDT3 each topic was guaranteed to have at least 4 hits in each language). The ratio of on-topic to off-topic stories made topic labeling akin to searching for a needle in a haystack, and some number of misses was inevitable.

5.2.4. Adjudication of sites' results

In order to offset the potential for missed stories even after recall QC, the LDC performed one additional quality assurance measure during topic labeling. NIST provided the LDC with the results of each research site's evaluation from the topic tracking task. The sites' systems were scored against the LDC's human-produced topic relevance tables, with the annotators' judgements taken as ground truth. Each system miss corresponded to a potential LDC false alarm, and each system false alarm was a potential LDC miss. The LDC adjudicated the systems' results as a final QC measure.

It would not have been possible to completely adjudicate all cases where LDC annotators differed from system performance. In the case of TDT3, NIST delivered results containing approximately 1.5 million topic-story tuples from 7 research sites. The effort needed to adjudicate all the cases of discrepancy would have exceeded the original corpus creation effort. Instead, the LDC reviewed cases where a majority of systems (i.e. 4 or more) disagreed with the original annotation. The adjudication effort for TDT2 was larger than that of TDT3 but still did not entail a complete adjudication of all discrepancies.

For both corpora, the number of LDC false alarms uncovered through the adjudication process was very low – for TDT3, less than 1% of system misses resulted in LDC false alarms. This was not surprising, given the complete precision check over all on-topic stories that eliminated most LDC false alarms. The rate of LDC misses identified during adjudication was higher than that of false alarms for both TDT2 and TDT3. However, even for the TDT3 corpus when no exhaustive recall check was performed, the rate of LDC misses was quite low: only 5% of stories reviewed during adjudication were actual misses (and as expected, the larger the number of sites reporting disagreement with LDC annotations, the higher the LDC miss rate).

6. Story linking

6.1. The annotation task

For the TDT3 corpus, a new form of non-exhaustive annotation supplemented the exhaustive topic-story labeling. Story link detection required the annotator to read a pair of stories and decide whether the pair "discussed the same topic". The use of *brief* was prohibited during the initial story link task – annotators were required to make a simple yes/no decision. The task involved a total of 180 seed stories, each of which was judged against 120 comparison stories (half of these were judged as relevant to the seed story by a search engine, the other 60 were chosen at random). A customized annotation interface displayed the seed story on one side of the display as a constant, while the compare stories were displayed on the other side of the screen in random order, until all 120 compare stories had been read and judgements made about each of them. Annotators made decisions for 21,500 story pairs in all. In contrast to the topic-story labeling task, the story link task did not involve the development of explicit topic definitions or topic descriptions. In fact, annotators were explicitly prohibited from establishing any pre-defined topic prior to judging the story pairs. The annotators were also prohibited from going back over their work and changing their judgements at a later point.

6.2. Quality assurance measures

The lack of a pre-defined topic led to unique challenges when evaluating inter-annotator consistency for the story link detection task. Unlike topic-story annotation, it was not possible to appeal to the topic explication or rules of interpretation in determining whether an annotator had accurately established a link between two stories. Answering the question "Do these two stories discuss the same topic?", with topic as a free variable, was a very different task than matching stories to a pre-established, clearly delimited topic description. Due to time and budgetary constraints, it was impossible to perform any dual annotation for the story link task, so no true measure of inter-annotator consistency (i.e., a kappa score) could be established. In consultation with the topic sponsors, modified versions of the precision and recall checks were applied to the story link task to provide some means of evaluating annotator consistency.

6.2.1. Precision

All *yes* links identified during the original annotation task were revisited, and either confirmed or changed to *brief* or *no*. The inclusion of *brief* during precision represented a modification of the original task: during story link annotation, annotators were forced to make a simple yes/no decision. The precision task was conducted by an independent annotator who had not participated in the story link task prior to doing precision. It was crucial for the integrity of the task that annotators not have any pre-conceived definition of the topic, so they were prohibited from discussing their work with the rest of the annotation team and worked in relative isolation. All changes made during precision were reviewed by team

leaders. Of the original 3942 *yes* story-pairs, the vast majority (83%) remained *yes*, just under 9% changed to *no*, with a similar number changing to *brief*. Despite the absence of a pre-defined topic, the story link annotators displayed a relatively high degree of consistency in linking stories.

6.2.2. Modified recall

In addition to the precision task, a modified form of recall was adopted to evaluate annotator behavior during the story link task. For each of the 180 story link seed stories, the "top 3 no" stories were reviewed by independent annotators. The top 3 *no* stories for each seed comprised the three highest-ranking compare stories (judged most relevant by the search engine using the seed story as a query) which were subsequently tagged as *no* during the original story link task. Three separate annotators performed this recall task, each viewing one of the three "top no" stories in comparison to that story's original seed. In addition, the two stories were simply presented as a story-pair, without the seed being explicitly identified as such. This differed from the original story link task where a single annotator viewed all 120 compare stories in succession against a single seed.

Of the original 521 top 3 *no* stories, 72% remained *no* during recall. Quite a large number of the stories, 17.7%, changed to *yes* and another 10% changed to *brief*. These numbers seem very high on the surface– after all, the topic-story labeling task showed a miss rate of only 5% (as measured by adjudication). But in story link detection, the concept of topic is a free variable. Although all annotators understood the concept of topic in a general sense, they were free to define the specific topic for each seed story in any way they desired, and that definition was allowed to drift – for the individual annotator during the course of the story link task, and across annotators during the QC tasks.

7. First Story detection

7.1. The annotation task

The final annotation task for TDT3 consisted of identifying the first chronologically on-topic story for 180 topics: the 60 pre-defined target topics, plus 120 more whose seeds were chosen at random. For the 60 topics that were part of exhaustive topic-story labeling, first story annotation was a side effect of standard topic labeling. After exhaustive labeling, the stories were simply sorted in chronological order to identify the first story. For the 120 topics used only in first story detection, a search engine was employed to identify the first on-topic story. Annotators used the seed story as a query to identify an additional 4 on-topic stories. Then using these 4 stories as the query, they conducted a careful search of data that chronologically preceded the earliest hit. When the annotator was confident that the earliest on-topic story had been found, s/he could establish a title and brief description of the topic, then move on to another seed story. No further annotation was done on these 120 additional topics.

7.2. Quality assurance measures

Because the first story detection task employed a search engine to perform basic annotation, it was not feasible to employ the same search engine to perform QC or measure annotator consistency. Instead, a relatively simple approach to quality assurance was adopted. All first stories for the 120 non-target topics were reviewed by senior annotators for plausibility (were the stories truly "on-topic" for the given seed? Were they likely to be the first story, given the details of the topic?). Any recall check would have essentially replicated the original annotation task so recall was not performed. Rather, if the plausibility check revealed any doubt about the identified first story, the annotator would employ the search engine and conduct additional searches for the true first story, using keywords and dates to refine the original search. As a result of these QC measures, 2 of the 180 original first stories were changed. In both cases, the original first story was rejected not because it wasn't the *first* on-topic story, but because it was judged to be off topic entirely.

8. Other quality control practices

While each annotation task that was part of the TDT corpora generated its own form of quality control, the LDC adopted measures to enhance inter-annotator consistency that weren't linked to any particular annotation task. The creation of the TDT corpora required a very large annotation staff. For each project, a team of Mandarin- and English-speaking annotators (some monolingual, some bilingual) had to be assembled, trained and integrated into the project workflow.

8.1. Staffing

The majority of TDT annotators were college students and recent graduates who were in some way affiliated with the University of Pennsylvania. Because annotators worked part-time schedules, it was necessary to hire a large crew for each of the TDT projects in order to complete all annotation within the project's timeline. Preference in hiring was given to annotators with an interest in current events and the media. Annotators were required to be comfortable working with computers, because of the complex annotation interfaces used. Mandarin-speaking annotators were also required to be fluent bilinguals, because most of the project documentation existed in English, and because most of the project managers were not Mandarin speakers.

Because of the demands of the project, annotators were required to work a minimum of 10 hours per week, and were asked to commit to working at least 3 months. In reality some of the student workers were not able to fulfill this obligation, but most of the annotators did stay on for the duration of the project. Many of the TDT2 annotators in fact returned for TDT3.

The annotation staff was divided into a Mandarin team and an English team. Team leaders were native speakers of each language. Project administration in English, but frequently the Mandarin team would conduct meetings in Mandarin and create Chinese versions of the English project documentation to better serve the Mandarin annotation team.

Senior annotators (those who had participated in TDT from the beginning) were appointed to help train new staff, conduct meetings, answer questions and generally act as a liaison between the annotators and the project managers. The physical arrangement of the annotation

staff also encouraged collaboration and a team-based approach. Annotators worked in close proximity to one another, and were encouraged to discuss questions and problems with their team members.

8.2. Training

Annotator training was the single most important quality control measure adopted during the TDT corpus creation efforts. Inter-annotator consistency could not have existed for a project of this size without considerable attention to training. In addition, quality assurance measures were seen not only as a way to measure inter-annotator consistency or the inherent variability of the annotation task; these measures are also a way to improve human performance. Because quality assurance measures were integrated into the normal workflow of TDT annotation, their results could be fed back into continued annotator training, leading to improved consistency.

For TDT, training took several guises. When new staff members were hired, they had to be trained not only in the execution of the annotation tasks but also in the use of the specialized interfaces. Training involved a multi-tiered approach. Annotators learned first about the goals of the TDT project, the research tasks, and even the potential applications of TDT technology. By placing the project in a larger context, annotators had a better understanding of why they were asked to execute the tasks in a particular way, and they were more willing to adhere to the annotation guidelines when they knew why the guidelines were in place.

The annotation guidelines were available to new annotators online or in hardcopy; new hires were also given several hours of face-to-face training from a team leader or senior annotator. All training materials were made available in a multitude of formats: on the web, in email archives, hardcopy, even on video in some cases. The most important concepts from training (concept of topic, topic lists) were also put up on posters around the annotation area.

After this initial training, new annotators were required to undergo a probationary period in which they were given a number of test files to annotate. These practice segmentation and labeling files were deliberately chosen to present the new annotator with the full range of annotation challenges – not only in terms of decisions about the content of the news stories, but also in terms of using the annotation interfaces. After each test file was completed, team leaders would review any problems with the annotator. After a sufficient number of test files had been successfully annotated, the new staffer was allowed to begin annotation on "real" data. For the first few weeks of any new annotator's work, team leaders and senior annotators would monitor that person's progress, spot checking files and reporting any problems.

While there was a large initial investment of time and effort in training new annotators, the training process did not end at that point. Ongoing staff training was a regular part of the TDT corpus creation efforts. Because quality control measures were part of the regular workflow, it was possible to use the results of QC for further annotation training. Annotators were required to attend weekly staff meetings with their annotation teams. The purpose of these meetings was to convey changes in the annotation guidelines, to discuss problems that were revealed during QC and discuss solutions, and to give annotators a chance to make suggestions about the project. At some of these meetings, annotators were tested with "pop quizzes" (e.g., *Match the world leader to the topic*); at other times they were given a chance to blow off steam.

One thing that emerged during the annotation team meetings was that TDT annotation, and topic labeling in particular, was a very stressful job. The task was fairly tedious, and annotators even reported having nightmares about TDT topics and feeling overwhelmed by the sense that the news was mostly bad, and their jobs were mostly news. (The preponderance of stories about the Monica Lewinsky scandal during the TDT2 and TDT3 epochs created another kind of stress for the annotators, which most of them categorized as intense irritation.)

In an effort to reduce the negative psychological impact of TDT annotation, project managers implemented "downtime". Approximately one hour every other week was set aside for the team to do something other than annotation, as a stress reliever. The team chose the downtime activities each week, which ranged from painting, to having a walk in the park, to going to the local art museum.

8.3. Project resources

While the greatest annotation project resource were the annotation teams themselves, other resources were put into place to make annotation as efficient and consistent as possible.

8.3.1. Documentation

The first of these was the abundance of corpus documentation available to the annotators. The annotation guidelines themselves, available online and displayed within the annotation interfaces, spelled out in great detail every aspect of each annotation task. A "Frequently Asked Questions" list was also available online for each task, and annotators could add to this FAQ. Annotators were encouraged to document all their annotation questions, and the easiest way to do this was via a customized email system. Project managers implemented a group mailer, accessible only to TDT staff, which automatically sorted, distributed and archived each annotator's question and the reply. This was an invaluable resource for maintaining inter-annotator consistency, and annotators were encouraged to search the email archives to find the answers to their questions. During quality control tasks, in particular precision checking of on-topic stories, it was a simple matter to access the email archive for a particular topic and review all the questions that had been debated for that topic, and the answers that had been established.

8.3.2. Topic lists and rules of interpretation

The online topic lists themselves were an essential project resource, and a strong factor in providing overall quality control for topic-story labeling. Using the randomly-selected seed story, team leaders identified the story's seminal event, then created a baseline topic definition. The format of the definition was fixed for each topic. The topic title was a brief phrase that was easy to remember and immediately evoked the topic. Each topic was accompanied by a topic icon, which provided the annotator with a visual reminder of the topic. The seminal

event that contributed the topic was described in terms of who/what/when/where, and the topic explication that followed provided further details.

Each topic was also linked to a "rule of interpretation". Topics generally fell into a small number of larger categories: natural disasters, legal cases, elections, crimes, etc. For each of these categories, the corresponding rule of interpretation established what kinds of stories would be considered "on-topic" vs. "off-topic".

The topic lists were available in both English and Chinese, and annotators could also click on a link to view examples of on- and off-topic stories for each topic. A sample TDT3 topic definition follows:

3043.

Sri Lankan Gov't. vs. Tamil Rebels

中文

Seminal Event

WHAT: New wave of violence breaks out between Tamil rebels and Sri Lankan government
WHERE: Sri Lanka
WHEN: late 1998

Topic Explication

Since 1983, more than 54,000 people have been killed in Sri Lanka's civil war between the majority Sinhalese who control the government and military, and the Liberation Tigers of Tamil Eelam, who are fighting for a separate homeland for minority Tamils in Sri Lanka's north and east. The fall of 1998 brought a new wave of violence and terrorism in this ongoing war. Although peace talks looked likely in late 1998, the fighting had begun again by January 1999. **On topic:** Any stories covering acts of violence or terrorism in this conflict; investigations by external organizations (like Amnesty International); peace negotiations between the opposing sides.

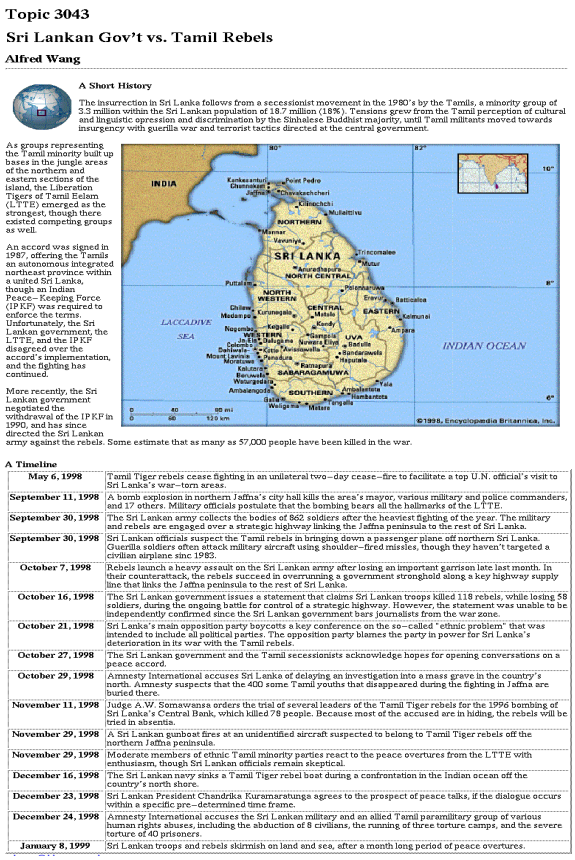
Rule of Interpretation Rule 6: Ongoing Violence or War

Figure 2: Sample TDT Topic Definition

Annotators were required to consult the topic list each time they began an annotation session, and the topic labeling interface required them to view the topic list before beginning annotation, and confirm that they'd read through it. The topic lists and rules of interpretation ensured that each annotator was working with the same understanding of the topic at hand and, at least in theory, that all annotators would identify the same stories as on-topic.

8.3.3. Topic research

One of the largest challenges to the annotators was the task of keeping abreast of developments for a particular topic. Although the topic definitions spelled out what sorts of stories might be considered on-topic, it was 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 performed topic research, providing additional material like timelines, maps, keywords, named entities, and links to online resources, for each topic. An example of a typical topic research document appears below:



Topic 3043

Sri Lankan Gov't vs. Tamil Rebels

Alfred Wang

A Short History

The insurrection in Sri Lanka follows from a secessionist movement in the 1980's by the Tamils, a minority group of 3.3 million within the Sri Lankan population of 18.7 million (18%). Tensions grew from the Tamil perception of cultural and linguistic opression and discrimination by the Sinhalese Buddhist majority, until Tamil militants moved towards insurgency with guerilla war and terrorist tactics directed at the central government.

As groups representing the Tamil minority built up bases in the jungle areas of the northern and eastern sections of the island, the Liberation Tigers of Tamil Eelam (LTTE) emerged as the strongest, though there existed competing groups as well.

An accord was signed in 1987, offering the Tamils an autonomous integrated northeast province within a united Sri Lanka, though an Indian Peace-Keeping Force (IPKF) was required to enforce the terms. Unfortunately, the Sri Lankan government, the LTTE, and the IPKF disagreed over the accord's implementation, and the fighting has continued.

More recently, the Sri Lankan government negotiated the withdrawal of the IPKF in 1990, and has since directed the Sri Lankan army against the rebels. Some estimate that as many as 57,000 people have been killed in the war.

A Timeline

May 6, 1998	Tamil Tiger rebels cease fighting in an unilateral two-day cease-fire to facilitate a top U.N. official's visit to Sri Lanka's war-torn areas.
September 11, 1998	A bomb explosion in northern Jaffna's city hall kills the area's mayor, various military and police commanders, and 17 others. Military officials postulate that the bombing bears all the hallmarks of the LTTE.
September 30, 1998	The Sri Lankan army collects the bodies of 862 soldiers after the heaviest fighting of the year. The military and rebels are engaged over a strategic highway linking the Jaffna peninsula to the rest of Sri Lanka.
September 30, 1998	Sri Lankan officials suspect the Tamil rebels in bringing down a passenger plane off northern Sri Lanka. Guerilla soldiers often attack military aircraft using shoulder-fired missles, though they haven't targeted a civilian airplane sinc 1983.
October 7, 1998	Rebels launch a heavy assault on the Sri Lankan army after losing an important garrison late last month. In their counterattack, the rebels succeed in overrunning a government stronghold along a key highway supply line that links the Jaffna peninsula to the rest of Sri Lanka.
October 16, 1998	The Sri Lankan government issues a statement that claims Sri Lankan troops killed 118 rebels, while losing 58 soldiers, during the ongoing battle for control of a strategic highway. However, the statement was unable to be independently confirmed since the Sri Lankan government bars journalists from the war zone.
October 21, 1998	Sri Lanka's main opposition party boycotts a key conference on the so-called "ethnic problem" that was intended to include all political parties. The opposition party blames the party in power for Sri Lanka's deterioration in its war with the Tamil rebels.
October 27, 1998	The Sri Lankan government and the Tamil secessionists acknowledge hopes for opening conversations on a peace accord.
October 29, 1998	Amnesty International accuses Sri Lanka of delaying an investigation into a mass grave in the country's north. Amnesty suspects that the 400 some Tamil youths that disappeared during the fighting in Jaffna are buried there.
November 11, 1998	Judge A.W. Somawansa orders the trial of several leaders of the Tamil Tiger rebels for the 1996 bombing of Sri Lanka's Central Bank, which killed 78 people. Because most of the accused are in hiding, the rebels will be tried in absentia.
November 29, 1998	A Sri Lankan gunboat fires at an unidentified aircraft suspected to belong to Tamil Tiger rebels off the northern Jaffna peninsula.
November 29, 1998	Moderate members of ethnic Tamil minority parties react to the peace overtures from the LTTE with enthusiasm, though Sri Lankan officials remain skeptical.
December 16, 1998	The Sri Lankan navy sinks a Tamil Tiger rebel boat during a confrontation in the Indian ocean off the country's north shore.
December 23, 1998	Sri Lankan President Chandrika Kuramaratunga agrees to the prospect of peace talks, if the dialogue occurs within a specific pre-determined time frame.
December 24, 1998	Amnesty International accuses the Sri Lankan military and an allied Tamil paramilitary group of various human rights abuses, including the abduction of 8 civilians, the running of three torture camps, and the severe torture of 40 prisoners.
January 8, 1999	Sri Lankan troops and rebels skirmish on land and sea, after a month long period of peace overtures.

Figure 3: Sample TDT Topic Research Page

Topic research was a valuable resource not only for initial topic annotation, but also at later stages of quality control, as it provided a framework to monitor topic development and curb "topic drift".

All TDT documentation was made available to annotators in a multitude of formats, and the materials were kept up to date as the project and the topics evolved. Most of this material, including the annotation guide and topic lists, can be viewed from the LDC's TDT website: http://www.ldc.upenn.edu/Projects/TDT.

8.3.4. Tools and technical resources

The general approach taken during the creation of the TDT corpora was to conduct multiple, complete passes over the data. Each pass was designed to support one specific annotation task. Because annotators were required to concentrate on one task at a time, they were able to develop an efficient approach to annotation. This efficiency was greatly aided by the use of specialized tools for each pass. LDC programming staff created custom interfaces for each annotation task, and each interface was simple to use and difficult to break. Each tool was constructed with the input of annotation team leaders, and the tools were revised and updated as project needs changed. The annotation tools also automatically logged annotation judgements into an Oracle database, which kept track of multiple judgements about each story in the corpus, and always allowed for both "test" and "fiat"

users, which enhanced their ability to be used in training and QC.

9. Conclusions

The creation of the TDT corpora is the largest data annotation project undertaken by the Linguistic Data Consortium to date. The massive amount of data, complex annotation specifications and large annotation staff required attention to quality control, and particularly inter-annotator consistency, at every stage of the game. By integrating quality control measures into the regular flow of work, by emphasizing ongoing annotator training and documentation of decisions, and by taking a specialized approach to each annotation task, it was possible to create two richly annotated corpora containing nearly 10 million human decisions while maintaining a high level of quality, consistency and accuracy.

10. References

Cieri, C., D. Graff, M. Liberman, N. Martey, S. Strassel, 1999. The TDT-2 Text and Speech Corpus. Presented at the DARPA Broadcast News Workshop, Washington, DC., February 1999.

Cieri, C., D. Graff, M. Liberman, N. Martey, S. Strassel, 2000. Large Multilingual Broadcast News Corpora for Cooperative Research in Topic Detection and Tracking: The TDT2 and TDT3 Corpus Efforts. In Proceedings of Language Resources and Evaluation Conference, Athens, Greece, May 2000.

Strassel, S., D. Graff, N. Martey, C. Cieri, 2000. The TDT3 Text and Speech Corpus. Presented at the Topic Detection and Tracking Workshop, February 2000.

Wayne, C., 2000. Multilingual Topic Detection and Tracking: Successful Research Enabled by Corpora and Evaluation. In Proceedings of Language Resources and Evaluation Conference, Athens, Greece, May 2000.