Meeting Recording

Quick Transcription Guidelines

Version 1.3

Linguistic Data Consortium

January 16, 2004

1 Introduction

The Meeting Recording Transcription project aims to accurately capture the speech of multiple, sometimes simultaneous, speakers to support research in automatic speech recognition technologies. This document describes segmentation and transcription of the 15-hour NIST Meeting Pilot corpus. The corpus will be transcribed to a level of quality suitable for system training (i.e., “quick” transcription).

The goal of quick transcription (hereafter QTR) is to produce an accurate, time-aligned transcript as quickly and efficiently as possible. To this end a stripped-down transcription specification is required, which excludes special markup and multiple quality checks in favor of a single, focused transcription pass.

1.1 Data

The recordings in the Meeting Pilot Corpus contain between three and eight participants per session, and were recorded using a variety of microphones and cameras. Each speaker is recorded on both an individual head mic channel and several mixed channels that include all of the participants. Each speaker is identified by a unique speaker ID that is associated with the individual channel on which they were recorded.

2 Segmentation

2.1 Overview

Segmentation involves "timestamping" the audio file for each given speaker. This stage indicates structural boundaries like turns, utterances and phrases within the interview, and allows the finished transcript to be time-aligned with the corresponding audio file. Segment boundaries also make transcription of the audio easier, by allowing the annotator to listen to small chunks of segmented speech at a time.

2.2 Approach

In order to facilitate quick and accurate segmentation of the audio data, initial timestamps will be created automatically, using LDC's Segmenter program. For every participant’s recording, Segmenter inserts boundaries around clusters of utterances and leaves the periods of silence for that channel unsegmented.

This program groups together sentences or phrases spoken without an intervening pause and, based on a number of acoustic cues, divides phrases that contain a lengthy internal pause (more than .02 seconds) into separate segments. Periods of non-speech that contain breathing or other low-amplitude speaker-produced noise will remain unsegmented and untranscribed. Human annotators will then perform a one-time real time pass over the automatically-segmented files to verify the accuracy of the timestamps.

2.3 Speaker identification

Each timestamp of the audio recording is assigned a unique speaker ID, which corresponds to the channel on which that speaker was recorded. These speaker IDs are consistent for the duration of the segmentation file. A separate table records speaker gender and native/non-native status for each session.

After the individual speaker segmentation files have been automatically created, a post-processing step combines the individual files into a single, mixed segment file. Individual speaker IDs are retained in the mixed segment file.

This approach increases the precision of the automatic segmentation by eliminating the sections of overlapping speech present in the mixed channel recordings that present particular difficulty for Segmenter. It also ensures that speaker IDs remain accurate for every timestamp.

2.4 Timestamp format

For QTR, timestamps indicate both the start time and the end time of a turn or breakpoint boundary, and are accurate to the nearest millisecond. Within each individual speaker file, neither point can overlap a previous timestamp for the same speaker. In other words, there are no overlapping timestamps within an individual speaker file.

While the timestamps for a single speaker are arranged in chronological order, there may be intervening periods of silence on a given channel. Timestamps will adhere to the following format:

```
Start-time<TAB>end-time<TAB>SPKR-ID:(colon)<TAB>transcription

25.23 27.03 A: transcript
25.05 30.34 D: transcript
26.15 27.28 B: transcript
33.54 35.33 E: transcript
```

Note in the example that there are both intervening periods of silence between the timestamps and regions where several participants are speaking at the same time.

2.5 Placement of segment boundaries

Timestamps must occur at regular intervals within each audio file. At a minimum, timestamps will identify speaker turns (change of speaker) for all files. Therefore, annotators or automatic processes will insert timestamps around "chunks" of speech from a single speaker that are separated by periods of lengthy silence (i.e., greater than 0.2 seconds). Additionally, to facilitate transcription annotators may place breakpoints within particularly long turns. Because breakpoints are inserted for ease of transcription, their exact implementation is subject to the individual annotator's discretion. In general, breakpoints should be placed at natural breaks in speech, such as ends of sentences or phrases, breath groups or pauses. This typically means that breakpoints occur every three to eight seconds.

Two things annotators must verify when inserting or double-checking timestamps of any kind are that timestamps never occur in the middle of a word and never clip off the end/beginning of a word. This latter consideration is trickiest, especially with certain sounds, like "s", "f", "t", "k", and "p". The human post-processing quality control pass specifically targets timestamps around words that begin or end with these sounds.

3 Transcription

3.1 Introduction

Once an audio file has been fully segmented and the speakers identified, it must be transcribed as accurately as possible. The words transcribed within each segment boundary must correspond exactly to the timestamps that have been created, so that the audio file is aligned with the transcript.

Annotators transcribe the file in its entirety, working with all channels at once, paying particular attention to overlapping speech regions.

3.2 Transcription conventions

The quick transcription task necessitates very few written conventions or special markup. The goal of the transcription process is to capture accurately the *content* words of the audio recording, and thus rapidly produce a readable transcript.

At minimum, standard written English spelling, capitalization (proper nouns and first words of sentences) and punctuation (periods, quotation marks, and commas) should be used for ease of comprehension and readability. All transcripts are spell-checked before they are considered finished.

3.2.1 Orthography and spelling

3.2.1.1 Capitalization

Capitalization in the transcripts is used to aid human comprehension of the text. Annotators should follow accepted standard written capitalization patterns, and capitalize words at the beginning of a sentence, proper names, and so on.

3.2.1.2 Spelling

Transcribers use standard orthography, word segmentation and word spelling. All files must be spell-checked after transcription is complete. When in doubt about the spelling of a word or name, annotators consult a standard reference, like an online or paper dictionary, world atlas or news website.

3.2.1.2.1 Mispronounced words

Annotators should not try to duplicate mispronounced words; instead, these words should be represented according to standard spelling.

3.2.1.3 Contractions

Annotators limit their use of contractions to those that exist in standard written English, and of course only when a contraction is actually produced by the speaker. Annotators must take care to transcribe exactly what the speaker says. The table below, while not comprehensive, shows some examples of how to transcribe common contractions.

Complete Form	Spoken As	Transcribed As	Incorrect
I have	I've	I've	
Cannot	can't	can't	
Will not	won't	won't	
you have	you've	you've	
Could not	couldn't	couldn't	
Should have	should've	should've	should of, shoulda
would have	would've	would've	would of, woulda
It is	it's	it's	Its
its (possessive)	Its	its	It's
Marvin (possessive)	Marvin's	Marvin's	
Marvin is	Marvin's	Marvin's	
Marvin has	Marvin's	Marvin's	
going to	Gonna	going to	Gonna
Want to	Wanna	want to	Wanna

Got to	Gotta	got to	Gotta

Note: Annotators should take care to avoid the common mistakes of transposing possessive *its* for contraction *it's* (it is); possessive *your* for the contraction *you're* (you are); *their* (possessive), *they're* (they are) and *there*.

Annotators should transcribe exactly what they hear using standard orthography. If a speaker uses a contraction, the word is transcribed as such: *they're*, *won't*, *isn't*, *don't* and so on. If the speaker uses a complete form, the annotator should transcribe what is heard: *they are*, *is not* and so on.

For non-standard contractions like "gonna" and "wanna" annotators should spell out the entire word: *going to*, *want to*.

3.2.1.4 Hyphenated words and compounds

No hyphenation is required for the QTR task. However, best practice suggests a conservative use of hyphens, especially with compound words, which can be tricky. When in doubt about the proper hyphenation of a word, annotators should do what is most natural.

3.2.1.5 Numbers

All numerals are written out as complete words. Hyphenation may be used for numbers between twenty-one and ninety-nine only.

```
twenty-two
nineteen ninety-five
seven thousand two hundred seventy-five
nineteen oh nine
```

3.2.1.6 Abbreviations

In general abbreviations should be avoided and words should be transcribed exactly as spoken. The exception is that when abbreviations are used as part of a personal title, they remain as abbreviations, as in standard writing:

```
Mr. Brown
Mrs. Jones
```

However, when they are used in any other context, they are written out in full:

I went to the *junior* league game.

The *doctor* suggested an herbal tea.

3.2.1.7 Acronyms and spoken letters

Acronyms and spoken letters should be capitalized. Acronyms that are pronounced as words should be written as such, and should not be distinguished by any particular symbol.

```
AIDS
FEMA
```

On the other hand, acronyms that are pronounced as a series of letters should be preceded by a ~ tilde. Individual spoken letters should be marked in the same manner.

```
~CIA
~RESPECT
I got a ~C on that paper.
~WWW dot ~ABC news dot com.
```

3.2.2 Disfluent speech

3.2.2.1 Introduction

Regions of disfluent speech are particularly difficult to transcribe. Speakers may stumble over their words, repeat themselves, utter partial words, restart phrases or sentences, and use hesitation sounds. Annotators should attempt to accurately transcribe as many of these utterances as possible. For purposes of QTR, annotators should not spend too much time trying to accurately capture difficult sections of disfluent speech but should make their best effort to transcribe what they hear after listening to the segment two or three times, then move on.

3.2.2.2 Filled pauses and hesitation sounds

Filled pauses are non-lexemes (non-words) that speakers employ to indicate hesitation or to maintain control of a conversation while thinking of what to say next. The spelling of filled pauses is not altered to reflect how the speaker pronounces the word (e.g., typing AH for a loud "ah" or ummmm for a long "um".) For English, this set includes *ah*, *eh*, *er*, *uh*, *um* but may be extended to include other common filled pauses.

3.2.2.3 Partial words

When a speaker breaks off in the middle of the word, annotators transcribe as much of the word as can be made out. A single dash - is used to indicate point at which word was broken off.

```
      Yes, absolu- absolutely.
```

3.3 Additional considerations

3.3.1 Noise

Neither background noise nor speaker noise will be transcribed.

3.3.2 Hard-to-understand sections

Sometimes an audio file will contain a section of speech that is difficult or impossible to understand. In these cases, annotators use double parentheses (( )) to mark the region of difficulty.

It may be possible to take a guess about the speaker's words. In these cases, annotators transcribe what they think they hear and surround the stretch of uncertain transcription with double parentheses:

```
      340.23   342.12   F: And she told me that ((I should just leave.))
```

If an annotator is truly mystified and can't at all make out what the speaker is saying, s/he uses empty double parentheses to surround the untranscribed region. Where possible, this untranscribed region gets its own timestamp, e.g.:

```
      340.23   345.88   F: (( ))
```

3.3.3 Speaker errors and non-standard usage

Annotators should not correct grammatical errors, e.g. "I seen him" for "I saw him" should be transcribed as spoken. The same goes for misused words: annotators should transcribe what is spoken, not what they expect to hear.

4 Appendix

4.1 English Interjections

ach
duh
eee
ew
ha
hee
hm
huh
huh-uh
jeepers
jeez
mhm
mm
nah
okay
oof
ooh
uh-huh
uh-oh
whew
whoa
whoops
woo-hoo
wow
yay
yeah
yep
yup

4.2 Summary of Conventions

Category	Condition	Markup	Example	Explanation
Orthography and spelling	Numbers	Spelled out	twenty-five, one oh nine, one hundred thirty-seven	Write out in full; no hyphenation necessary for twenty-one through ninety-nine.
	Standard contractions	Transcribe as spoken	can't, I'm	If you hear a contraction used, write it as a contracted form.
	Non-standard contractions	Not used	going to, want to	Do not use non-standard contractions. Write the words out in full.
	Punctuation	Question mark, period, comma	? , .	Limited to these symbols.
	Pronounced Acronyms	Capitalized; no special markup	NAFTA, NATO	Write letters out as a word.
	Individual letters	Capitalized; marked with ~ tilde	~I before ~E, ~FBI	Individual letters spelled out.
Disfluent speech	Filled pauses	No special markup	uh, um	Most common include ah, eh, er, uh, um
	Partial words	-	absolu-	Speaker-produced partial words are indicated with a dash.
Other markup	Semi-intelligible speech	((text))	They lived ((next door to us)).	The transcriber's best attempt at transcribing a difficult passage.
	Unintelligible speech	(( ))	(( ))	This indicates an entirely unintelligible passage.
	Interjections	No special markup	Uh-huh, yeah, mhm	Use standardized spellings