Meeting Recording
Careful Transcription Guidelines

Linguistic Data Consortium

January 16, 2003

Version 1.2

Prepared by Linguistic Data Consortium
1/16/2004

Table of Contents

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 90-minute multi-site meeting data test set to be used in evaluating the output of automatic speech-to-text (STT) transcription systems. The goal of the transcription process is to provide an accurate, verbatim (word-for-word) transcript of the entire recording. The transcript will be time-aligned with the audio file, and additional features of the audio signal and speech will be identified using special markup. Multiple quality checks will be performed to produce final transcripts of very high quality.

The transcription guidelines in this document were developed to follow the existing DARPA/NIST Rich Transcription (RT) program transcription specifications. The existing RT guidelines address the transcription of news broadcasts and telephone conversations and are available from: http://www.nist.gov/speech/tests/rt/rt2003/spring/docs/RT_Transcription_V2.2.pdf

1.1 Data Content

The recordings in the multi-site meeting test corpus contain between 3 and 8 participants per session, and were recorded using a variety of microphones and cameras. Each speaker is recorded on not only an individual head mic channel, but also on 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.

1.2 Data format

The original transcripts will be produced in the Annotation Graph Rich Transcription format. If necessary, the files will be converted to NIST's Rich Transcription format before delivery.

1.3 Approach

Because the meeting data recordings contain a large amount of overlapping speech, it is necessary for annotators to access both the individual speaker recordings as well as the mixed recording in order to ensure highly accurate segmentation and transcription. The following approach will be adopted for transcription of the evaluation data:

1) using the individual speaker recordings, annotators produce accurate segment boundaries for each speaker
2) using the individual speaker recordings, annotators produce an accurate transcript for each speaker
3) using mixed recordings and merged segment and transcription files, annotators do multiple quality control passes to create additional markup and check for errors

2 Segmentation

2.1 Overview

The segmentation process begins with the creation of initial timestamps for the audio file. Timestamps indicate when different things are happening in the audio and so allow the transcript to be aligned with the corresponding audio file. Segmenting the file in this manner also makes

transcription of the audio easier, by permitting the transcriber to listen to small chunks of segmented speech at a time.

For careful transcription of the meeting data evaluation set, timestamps indicate both the start time and the end time of a turn or breakpoint boundary, to the nearest tenth of a second. Neither start time nor end time can overlap a previous timestamp of the same speaker; in fact, the segmentation tool will not allow this to happen.

Timestamps do not necessarily occur in direct succession, one after another. Instead, there may be intervening periods of silence on several channels while a speaker on another channel is talking, or there may be regions where several people are speaking simultaneously.

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.

2.2 Timestamp format

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>transcript

25.01 27.45 A:    transcript
25.55 29.24 D:    transcript
26.53 27.22 B:    transcript
33.41 35.30 E:    transcript
```

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

2.3 Placement of segment boundaries

Timestamps must occur at regular intervals within each audio file. At a minimum, timestamps must identify speaker turns (change of speaker) for all files. Therefore, annotators should insert timestamps around "chunks" of speech from a single speaker that are separated by periods of lengthy silence (i.e., greater than one half second). To facilitate transcription, annotators will also 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 happen 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". Transcribers take special care when inserting timestamps around words that begin or end with these sounds.

3 Transcription

3.1 Introduction

Once a file has been fully segmented and the speakers identified, it must be transcribed. Annotators must produce a verbatim (word-for-word) transcript of everything that is said within the file. 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.

3.2 Transcription Conventions

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. (Note that proper names are also labeled with a caret ^ symbol whose use is detailed in Section 3.2.4.4.)

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.

3.2.1.3 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), and *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 contracted: *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 Numbers

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

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

3.2.1.5 Hyphenated words and compounds

In general, annotators should be conservative about use of hyphens. For instance:

an overly complicated analysis **not** *an overly-complicated analysis*

However, in some cases, a hyphen is required:

anti-nuclear protests **not** *anti nuclear protests*

Compounds can be tricky. When in doubt, annotators should consult a dictionary and talk to their language team leader.

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
Dr. Spock
```

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

```
I went to the junior league game.
I went to the doctor and he suggested an herbal tea.
Hey mister, do you know how to get to the stadium?
```

3.2.1.7 Acronyms and spoken letters

Acronyms that are pronounced as a single word should be written in all capital letters, and preceded by the @ symbol:

```
@NASA
@AIDS
```

Acronyms that are normally written as a single word but pronounced as a sequence of individual letters should be written in all caps, with each individual letter preceded by a ~ tilde symbol:

```
~F ~B ~I
~C ~E ~O
```

Similarly, individual letters that are pronounced as such should be written in caps, with each letter preceded by a tilde:

```
I got an ~A on the test.
His name is spelled ~S ~I ~M ~P ~S ~O ~N.
```

3.2.1.8 Punctuation

Annotators should use standard punctuation for ease of transcription and reading. Acceptable punctuation is limited to periods and question marks at the end of a sentence, and commas within a sentence. Transcripts should not contain quotation marks, exclamation marks, colons, semicolons, dashes or ellipses in transcribing. Punctuation is written as it normally appears in standard writing, with no additional spaces around the punctuation marks.

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 lots of hesitation sounds. Annotators should take particular care in sections of disfluent speech to transcribe exactly what is spoken, including all of the partial words, repetitions and filled pauses used by the speaker.

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. Each language has a limited set of filled pauses that speakers can employ. Annotators use the standardized spellings shown in the table below for filled pauses. 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.

All filled pauses are indicated with a % sign preceding the word.

English Filled Pauses
%ah
%eh
%er
%uh
%um

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.

```
32.45 35.01 A:    Yes, absolu- absolutely.
```

3.2.2.4 Mispronounced words

A plus symbol + is used for obviously mispronounced words (not regional or non-standard dialect pronunciation). Annotators should transcribe using the standard spelling and should not try to represent the pronunciation.

```
45.45 48.01 A:    He'll +probably I mean probably go with me
tomorrow.
```

3.2.3 Noise

3.2.3.1 Speaker noise

Speaker-produced noise is identified with one of the following five tags:

```
{laugh}                                 {breath}
{cough}                                 {lipsmack}
{sneeze}
```

3.2.3.2 Background noise

When there is noticeable background noise (not speaker noise) present during a span of speech, annotators employ the <noise> notation.

When the sound is instantaneous, like a short clap, paper rustle, door slamming shut or gunshot, the <noise> symbol is inserted next to the word during which the noise occurs. For instance:

```
67.45 69.01 A:    I'm not really sure <noise> what she said.
70.01 73.45 A:    Hey, did you hear that too?
```

If the sound is prolonged and spans several words in the transcript, the <noise> symbol is inserted before to the word where the sound begins, and </noise> is inserted after the word where the sound ends. For instance,

```
67.45 69.01 A:    <noise> I can't tell what's going on out there.
70.01 73.45 A:    It's really getting loud. </noise>
```

If the sound is very long, it might cross breakpoints or speaker turns.

Note: If the file contains persistent or overwhelming distortion, static or background noise, annotators should notify their language team leader.

3.2.4 Additional markup

3.2.4.1 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.

Sometimes it is 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:

```
70.01 73.45 A: 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.:

```
70.01 73.45 A: (( ))
```

3.2.4.2 Idiosyncratic words

Occasionally a speaker will make up a new word on the spot. These are not the same as slang words; rather, they're words that are unique to the speaker in that conversation. If annotators encounter an idiosyncratic word, they should transcribe it to the best of their ability and mark it with an asterisk *. For instance,

```
Do you dress like a *schlump yet?
Why she said *drr I don't know
```

3.2.4.3 Foreign languages

Portions of speech in another language are annotated using the <language text> convention to indicate the language and to transcribe the words that are spoken in that language. For instance:

```
And then I took all of the <German Sachen> to my room.
Oh, <Spanish gracias> he said.
```

If the annotator does not know the name of the language or what is being said, they should use the tag <foreign> in isolation.

```
Then there were a couple of <foreign> which I tried on.
```

3.2.4.4 Proper nouns

All proper nouns, including personal names, place names and the like, are marked with a caret ^. Common nouns that are functioning as names or titles are not marked. If the name contains more than one word, all words in the name are annotated with a caret.

```
^Osama ^bin ^Laden
^Sony
^Maria's Bar and Grill
He calls himself ~J ~R ^Jones.
Secretary of State ^Madeline ^Albright
Middle East
```

When annotators encounter a name whose spelling they're not sure of, they should spend a moment searching for the proper spelling. If they can't find the correct spelling after a moment or two, they should make their best guess and use a double caret ^^ instead of a single caret to mark the name, e.g.

```
^^Rafjanii ^Agrawal
```

These names will be reviewed again during second passing and the spelling, verified.

3.2.4.5 Interjections

The following standardized spellings are used to transcribe interjections. Interjections do not require any special symbol.

English Interjections

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

3.2.5 Summary of special symbols

Category	Condition	Markup	Example	Explanation
Orthography and spelling	**Numbers**	Spelled out	twenty-five, one oh nine, one hundred thirty-seven	Write out in full; dashes 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	Comma, question mark, period	, ? .	Limited to these three symbols.
	Pronounced acronyms	@	@NAFTA	Write letters with all caps, no space between letters.
	Individual letters	~	~I before ~E ~Y ~M ~C ~A	Individual letters spelled out, capitalized, each with ~
Disfluent speech	**Filled pauses**	%	%ah, %uh	Limited to small list for each language; use standardized spellings

	Partial words	-	absolu-	Speaker-produced partial words are indicated with a dash. Transcribe as much of the word as you hear.
	Speaker restart	--	I think -- I thought he was there.	Used when the speaker stops short and then repeats themselves or abandons the utterance completely, restarting with a new sentence.
	Mispronounced words	+	+probably	Mispronounced word (a speech error). **NOTE**: Do not use this symbol to indicate non-standard but common regional/social dialect pronunciations. Transcribe non-standard pronunciation variants or mispronounced words using standard orthography.
Noise conditions	**Speaker noise**	{ }	{breath} {cough} {laugh} {sneeze} {lipsmack}	Sounds made by the talker. Limited to these five.
	Non-speaker noise	<noise> </noise>	<noise> What's that sound? </noise>	Use <noise> for instantaneous sounds. Use <noise> text </noise> for ongoing sounds. This convention should be used for any background noise, static, distortion or other non-speaker noise.
Other markup	**Semi-intelligible speech**	((text))	They lived ((next door to us)).	This is the transcriber's best attempt at transcribing a difficult passage.
	Unintelligible speech	(( ))	(( ))	This indicates an entirely unintelligible passage.
	Idiosyncratic words	*	*poodleish	Speaker uses a "made-up" word. **NOTE:** Do not use for non-standard dialect terms or misused words.
	Foreign language	<language text>	<French merci> <foreign>	This is used to indicate foreign speech. If the word is unknown, leave it out. If the language is unknown, merely write <foreign>. **NOTE**: Do not use this convention for foreign borrowings that are common in the target language, e.g. *apropos*.
	Proper names	^	^Osama ^bin ^Laden ^Mariani's Bar and Grill Secretary of State ^Albright	Use caret symbol for each word of proper name. Do not use the caret for common nouns that are part of a title or name.
	Interjections	no special markup	uh-huh, yeah, mhm	Use standardized spellings

3.3 Some general considerations

Annotators should not correct grammatical errors or non-standard usage, 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.

Annotators should not imitate a speaker's non-standard pronunciation. Standard spelling should be adopted for non-standard pronunciations. Obviously mispronounced words (as opposed to non-standard pronunciations) should be marked with the plus + symbol.

4 Quality Control

4.1 Introduction

Second passing is used as the primary quality control measure to ensure the accuracy of segmentation and transcription including markup. After the initial file has been fully segmented and transcribed, a new annotator listens to the entire file while viewing the corresponding transcript, making adjustments to the timestamps or transcription as needed.

Second passing entails a combination of manual and programmatic checks on the transcript files. The particular types of checks conducted during second passing are described below.

4.2 Segmentation Verification

Second pass annotators verify that each timestamp matches the corresponding transcript exactly by playing each timestamp in turn, confirming that the audio and transcript for that segment are an exact match, and making any necessary corrections. Annotators also check that the timestamp has been placed in a suitable location, e.g. between phrases or sentences, and that the timestamp does not chop off the start or end of any word.

Annotators should listen to the entire mixed recording to ensure that all speech for each channel is captured within a turn segment and that no speech remains outside of a segment boundary.

4.3 Transcription Verification

During the transcript checking phase of second passing, annotators examine the transcript in detail, checking for accuracy, completeness and the consistent use of transcription conventions. Annotators pay especial attention to a handful of areas that are particularly difficult to transcribe, notably, unintelligible speech sections and areas of speaker disfluency. Any proper names whose spelling could not be verified during the initial transcription process are corrected and standardized within the file. Finally, annotators conduct a spell check on the file.

4.4 Automatic checks

In addition to the manual checks described above, annotators employ a series of programmatic inspections of the data known as a “syntax check”. The syntax check scans the transcript file for common transcription errors as well as inaccuracies of data formatting, including:

- timestamps without text data
- timestamps followed by non-empty lines
- foreign language convention badly formatted
- unintelligible speech convention badly formatted
- illegal character used in transcript
- bad spacing around punctuation
- digits are not spelled out

Annotators run the syntax checker as a final pass over the data before completing a file. The syntax checker outputs an error report. The annotator reviews each error in succession, and can "zoom" to the problematic section of the transcript to make any necessary corrections.