

Guidelines for RT-04 Transcription

Linguistic Data Consortium

Version 3.1 – March 31, 2004

Table of Contents

1	Introduction	3
2	Segmentation	3
2.1	Overview	3
2.2	Broadcast News	3
2.2.1	Overview	3
2.2.2	Sections	4
2.2.2.1	Section types	4
2.2.2.1.1	Introduction	4
2.2.2.1.2	News stories <sr>	4
2.2.2.1.3	Non-news <sn>	4
2.2.2.2	Beginning of a section	4
2.2.2.3	End of a section	5
2.2.2.4	End of the file	5
2.2.3	Turns	5
2.2.3.1	Introduction	5
2.2.3.2	New speaker turns	5
2.2.3.3	End of a turn	5
2.2.3.4	Overlapping speech	5
2.2.3.5	Simultaneous speech	5
2.2.3.6	Short periods of non-speech	6
2.2.3.7	Extended periods of non-speech	6
2.2.4	Breakpoints	6
2.2.4.1	Overview	6
2.2.5	Summary of segment boundary symbols in broadcast news	6
2.2.6	Speaker Identification	7
2.2.6.1	Introduction	7
2.2.6.2	Speaker Type	7
2.2.6.3	Names and Identifiers	7
2.2.6.4	Native and non-native speakers	7
2.2.6.5	Examples of speaker IDs	7
2.3	Telephone Speech	7
3	Transcription	8
3.1	Introduction	8
3.2	Transcription Conventions	9
3.2.1	Orthography and spelling	9
3.2.1.1	Capitalization	9
3.2.1.2	Spelling	9
3.2.1.3	Contractions	9
3.2.1.4	Numbers	10
3.2.1.5	Hyphenated words and compounds	10
3.2.1.6	Abbreviations	10
3.2.1.7	Acronyms	10
3.2.1.8	Spoken letters	10
3.2.1.9	Punctuation	11
3.2.2	Disfluent speech	11
3.2.2.1	Introduction	11
3.2.2.2	Filled pauses and hesitation sounds	11
3.2.2.3	Partial words	11
3.2.2.4	Restarts	12
3.2.2.5	Mispronounced words	12
3.2.3	Noise	12
3.2.3.1	Speaker noise	12
3.2.3.2	Background noise	12
3.2.4	Additional markup	13
3.2.4.1	Hard-to-understand sections	13
3.2.4.2	Idiosyncratic words	13
3.2.4.3	Foreign languages	13
3.2.4.4	Proper nouns	13
3.2.4.5	Interjections	14
3.2.4.6	Summary of special symbols	16
3.3	Some general considerations	17
4	Second Passing	17
4.1	Introduction	17
4.2	Segmentation and Speaker ID Verification	17
4.3	Transcription verification	17
4.4	Automatic checks	18

1 Introduction

The goal of the transcription process is to provide an accurate, verbatim (word-for-word) transcript of the entire broadcast. 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.

2 Segmentation

2.1 Overview

The segmentation process begins with creation of initial timestamps for the audio file. Timestamps indicate when different things are happening in the audio, and so allow us to align the transcript with the corresponding audio file. Timestamps also make transcription of the audio easier, by allowing the transcriber to listen to small chunks of segmented speech at a time.

Timestamps must occur at regular intervals within each audio file. At a minimum, timestamps must identify

- **section boundaries** within Broadcast News files
- **speaker turns** (change of speaker) for all files

In addition, transcribers insert additional **breakpoints** within each speaker's turn. This helps break up long turns into more manageable units, and makes transcription easier.

Some things to consider when inserting timestamps of any kind:

- Timestamps must never occur in the middle of a word.
- Be careful not to clip off the end/beginning of a word when inserting a timestamp. This is trickiest with certain sounds, like "s", "f", "t", "k", "p". Take special care when inserting timestamps around words that begin or end with these sounds.

Good places to insert timestamps are

- at pauses
- at breaths
- at ends of sentences or phrases

2.2 Broadcast News

2.2.1 Overview

Broadcast news data comprises a single audio channel. Multiple speakers appear on this single channel, and each speaker is identified by name or unique speaker ID. More than one speaker may talk simultaneously, resulting in sections of overlapping speech.

Broadcast news files are hierarchically arranged into Sections, Turns, and Breakpoints. All section boundaries are identified and timestamped within a broadcast file. However, only news stories are further segmented into turns and breakpoints. Commercials and other non-news sections are not segmented further (see **Section 2.2.2.1** below for definitions of each section type).

Within broadcast files, timestamps indicate only the start time of the section, turn, or breakpoint boundary. Because broadcast speech occurs on a single audio channel, timestamps occur one after the other, in direct succession and typically without intervening periods of unsegmented

audio. The end time of a timestamp is implied by the start time of the following timestamp. For instance, a series of timestamps in a broadcast file might look like this:

```
<sr 56.827>
<b 60.467>
<t 63.980>
<b 67.989>
<b 71.501>
```

Note that these timestamps occur in direct succession, with no large gaps.

2.2.2 Sections

2.2.2.1 Section types

2.2.2.1.1 Introduction

Each new part of the broadcast should be identified and labeled with the appropriate section label. There are two types of section boundaries:

- <sr> refers to news reports
- <sn> refers to non-news sections, including commercials

2.2.2.1.2 News stories <sr>

A news story is a topically contiguous segment of the broadcast. News stories may be of any length 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 story begins. Annotators rely on audio cues (speaker changes, music, pauses) to inform their judgments.

Promotional spots for upcoming stories or very brief reviews of top stories reported on more fully within the same broadcast are also news stories. These types of reports 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 news stories.

2.2.2.1.3 Non-news <sn>

Non-news segments include commercials, reporter chit-chat outside of the context of a story, station identifications, public service announcements, promotions for upcoming broadcasts, and long musical interludes. If multiple non-news sections follow one another within a transcript, they are grouped together with a single <sn> tag at the beginning of the section.

2.2.2.2 Beginning of a section

At the beginning of a new section, annotators insert the appropriate section label and timestamp. Because each section implies a new speaker turn, a speaker ID is also inserted at the start of each new section (see Section 2.2.6). For instance:

```
<sr 21.232> <<male, Lou_Waters>>
The last great explorer ^Jacques ^Cousteau has died in ^Paris at age eighty-seven.
<b 25.907>
{breath} Part of Early Prime is being preempted so that for the next half hour we can
remember one of the giants of the twentieth century.
```

2.2.2.3 End of a section

If the end of a section is directly followed by the start of another section, there is no need to specifically label and timestamp the end of the first section.

If the section is followed by a period of non-speech (music, sound effects or silence), annotators explicitly timestamp and label the end point of the section with <e>.

2.2.2.4 End of the file

Each file must end with a final timestamp, indicating where the audio recording for that program concludes. This timestamp should be labeled with <e> to indicate end.

2.2.3 Turns

2.2.3.1 Introduction

Within sections there are turn boundaries representing the start of each speaker's turn in the broadcast or telephone call.

2.2.3.2 New speaker turns

Every time there is a speaker change in the audio, this is indicated by inserting a <t> turn marker and timestamp. For Broadcast News, each <t> segment must also receive a speaker ID.

For Broadcast News files, annotators indicate <t> segments only within <sr> sections. They do not indicate turn boundaries within <sn> sections, since these sections are not transcribed.

2.2.3.3 End of a turn

If the end of one speaker's turn is directly followed by the start of another speaker's turn, there is no need to specifically label and timestamp the end of the first speaker's turn.

If a speaker's turn is followed by a period of non-speech (music, sound effects or silence), then annotators explicitly timestamp and label the end point of the speaker's turn with <e>.

2.2.3.4 Overlapping speech

Overlapping speech regions are marked with an <o> overlap tag at the beginning of the second speaker's interruption. Annotators will not transcribe any of the overlapping speech portions, but will leave the entire region blank. The section following an overlapping speech region will receive a new <t> turn tag, with speaker ID. For instance,

```
<t 90.66> <<male, Peter_Jennings>>  
And how do you perceive  
<o 91.21>  
<t 94.54> <<female, Paula_Zahn>>  
And that's just it, ^Peter.
```

If there is a large gap between the end of the overlapping speech region and the start of the next turn, mark the conclusion of the overlap with an <e> end section tag, and start the next turn with a <t> turn tag.

2.2.3.5 Simultaneous speech

When two or more speakers begin talking simultaneously, annotators will employ the <o> overlap tag for that region. The next section of non-simultaneous speech will be marked as a new <t> turn.

```
<o 189.01>  
<t 199.88> <<male, non-native, Jacques_Cousteau>>
```

And I loved the sea so much because of that.
 <t 203.39> <<female, speaker_1>>
 Y- yes, that's what I was going to ask you.

2.2.3.6 Short periods of non-speech

For a short (greater than 0.5 seconds but less than 5 seconds) period of silence, music, or other non-speech, a tag is inserted at the start of the non-speech section. The [[NS]] no speech marker is used to indicate that no speech occurs during this breakpoint. A new breakpoint is then inserted at the next region of speech. For example,

<b 123.456 >
 The crowd was furious.
 <b 124.567>
 [[NS]]
 <b 128.987>
 Calm was soon restored by the arrival of the riot police.

2.2.3.7 Extended periods of non-speech

For an extended (more than 5 seconds) period of silence, music or other non-speech, annotators insert an <e> end section tag at the start of the non-speech region. They then start the new <t> turn at the next speech region. For example,

<t 148.57>
 Gunfire filled the air.
 <e 154.50>
 <t 170.89>
 That sound greeted early morning visitors on Tuesday.

2.2.4 Breakpoints

2.2.4.1 Overview

Breakpoints are timestamps within a speaker turn. These internal timestamps are inserted to break up long speaker turns for ease of transcription. Annotators should insert breakpoints around breath groups, at ends of sentences or phrases, and at noticeable pauses. Breakpoints are also inserted around lengthy (greater than 0.5 seconds) non-speech events within a speaker's turn. This includes things like music, sound effects, and silence.

Because breakpoints are inserted for ease of transcription, their exact implementation is subject to the individual annotator's discretion. In general, breakpoints tend to occur every three to eight seconds.

2.2.5 Summary of segment boundary symbols in broadcast news

The table below summarizes the segment labels used during segmentation.

Label	Description
<sr>	start of news story section
<sn>	start of non-news section: commercials, etc.
<t>	start of non-initial speaker turn within section
	breakpoints within speaker turn
<e>	end of turn within section, followed by a non-speech region
<o>	start of overlap region (speaker one is interrupted by speaker two)

2.2.6 Speaker Identification

2.2.6.1 Introduction

In addition to identifying segment boundaries and timestamping them, annotators must also identify all of the speakers within a broadcast. Speaker name, type, and native/non-native speaker status are all recorded. If annotators are unable to determine the name of a speaker, they assign that speaker a unique numerical identification, and use the same speaker ID throughout the transcript file.

2.2.6.2 Speaker Type

There are four speaker types as follows:

- Female – used for adult females
- Male – used for adult males
- Child – used for children of either sex
- Other – used for speakers in unison, altered voices, unknown speaker sex, etc.

2.2.6.3 Names and Identifiers

Whenever possible, annotators record the proper name of the speaker. Examples of proper names include Jacques_Cousteau, William_Cohen, and Madeleine_Albright. Annotators must use the same spelling of proper names within a broadcast file, and wherever feasible across broadcast files as well.

If a speaker is not identified by name within a recording, a unique numerical index is used. Unnamed speakers are divided into Reporter and Speaker. Reporter is used for news anchors, interviewers, or reporters on the scene of a story. Speaker refers to anyone else who is not identified by name. The numerical IDs for Reporter and Speaker IDs cannot overlap; each successive anonymous speaker has a unique number, regardless of the category the speaker is assigned to. For example, the following sequence is entirely possible:

```
reporter_1
reporter_2
speaker_3
speaker_4
reporter_2 (the same voice as the previous reporter_2)
reporter_5
```

2.2.6.4 Native and non-native speakers

In addition to indicating speaker type and name/ID, annotators also indicate when a speaker is a non-native speaker. In English broadcast news, native is defined as a speaker of any North American English dialect. As native is the default, this is not explicitly marked. Non-native is used for speakers of other dialects of English, including British English or Indian English; non-native is also used to indicate people who are not native English speakers and have a discernable foreign accent. Examples of Speaker IDs include the following:

2.2.6.5 Examples of speaker IDs

```
<sr 1.402> <<male, Leon_Harris>>
<sr 158.244> <<female, Joie_Chen>>
<t 196.813> <<male, speaker_1>>
<t 498.314> <<female, non-native, speaker_3>>
<t 567.215> <<other, speaker_4>>
```

2.3 Telephone Speech

Unlike broadcast news, telephone speech is recorded on two separate channels. In most cases, each channel corresponds to a single speaker for the duration of the call. The two channels are labeled Speaker A (the local channel) and Speaker B (the remote channel).

Occasionally, multiple speakers appear on a single channel. If this happens, the additional speakers are identified by the channel they appear on, plus a number to distinguish them from the main speaker. For example,

Speaker A1: second speaker on channel A
Speaker A2: third speaker on channel A
Speaker B1: second speaker on channel B

and so on.

For Telephone Speech files, annotators apply turn boundaries to the entire speech file, working with one channel at a time. Section boundaries do not apply to telephone speech data.

Within telephone speech files, timestamps indicate both the start time and the end time of a turn or breakpoint boundary. Timestamps do not necessarily occur in direct succession, one after another. Instead, there may be intervening periods of silence on one channel while the other channel's speaker is talking. Therefore, the end time of each timestamp must be explicitly indicated, and cannot be extrapolated from the start time of the following timestamp. For instance, a series of timestamps in a telephone speech file might look like this:

25.66 27.15 A:
33.53 35.76 A:
37.00 42.34 A:

Note that there are intervening periods of silence between the timestamps.

During segmentation, annotators work with one audio channel at a time. When the two channels have been completely segmented, they are interleaved for purposes of transcription. For instance,

24.24 25.53 B:
25.66 27.15 A:
27.25 30.71 B:
31.00 33.30 B:
33.53 35.76 A:

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.

For Broadcast News files, annotators transcribe only <sr> sections. They do not transcribe any section which has been labeled <sn>, non-news.

For Telephone Speech files, annotators transcribe the file in its entirety, working with both channels at once.

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 beginnings of sentences, proper names, and so on. (Note that proper names are also labeled with a caret ^ symbol, the use of which 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.¹

Complete Form	Spoken As	Transcribed As	Incorrect
I have	<i>I've</i>	I've	
cannot	<i>can't</i>	can't	
will not	<i>won't</i>	won't	
you have	<i>you've</i>	you've	
could not	<i>couldn't</i>	couldn't	
should have	<i>should've</i>	should've	should of, shoulda
would have	<i>would've</i>	would've	would of, woulda
it is	<i>it's</i>	it's	its
its (possessive)	<i>its</i>	its	it's
Marvin (possessive)	<i>Marvin's</i>	Marvin's	
Marvin is	<i>Marvin's</i>	Marvin's	
Marvin has	<i>Marvin's</i>	Marvin's	
going to	<i>gonna</i>	going to	gonna
want to	<i>wanna</i>	want to	wanna
got to	<i>gotta</i>	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.

¹ All contractions are expanded to their original form, as a post processing step. All cases are reviewed and expanded manually, to reflect both the contracted form of the word and the resulting expansion. Certain contractions receive more attention than others: for example, in the case of **Marvin's**, an annotator will carefully review the contraction and restore the word to its original construction – Marvin is, Marvin has, or Marvin's (possessive).

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
St. John's Cathedral

However, when the complete forms of these commonly-abbreviated words are used in any other context, they are written out in full. For example:

I went to the *junior* league game.
I went to the *doctor*, and he gave me an aspirin.
Hey *mister*, do you know how to get to the stadium?

3.2.1.7 Acronyms

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

@NASA
@AIDS

3.2.1.8 Spoken letters

Abbreviations that are normally written as a single word, but are 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.

3.2.1.9 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, single (stand-alone) 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	Arabic Filled Pauses	Chinese Filled Pauses
%ah	%ah	%呵
%eh	%E	%呃
%er	%M	%唔
%uh	%uh	
%um	%hm	
	%hum	

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.

3.2.2.4 Restarts

Speaker restarts are indicated with double dash --. Annotators use this convention for cases where a speaker stops short, cutting him/herself off before continuing with the utterance.

<b 850.021>
I thought he -- I thought he was there.

<b 94.271>
The thi- -- the thing we're worried about is

3.2.2.5 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.

<b 1120.021>
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:

<b 1116.183>
I'm not really sure <noise> what she said.
<b 1120.021>
Hey, did you hear that? It sounded like a car backfiring.

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,

<b 1116.183>
<noise> I can't tell what's going on out there.
<b 1120.021>
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:

```
<b 1116.183>
And she told me that ((I should just leave))
```

If an annotator is truly mystified and cannot make a guess 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. For example:

```
<b 1116.183>
(( ))
```

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, but rather are 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.² 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.

² The foreign convention is changed, as a post-processing step, to show the foreign language value more clearly.

```
<foreign language="German">Sachen</foreign>
<foreign language="Japanese"></foreign>
```

Or, if the language is unknown, then that, too, is noted, and appears as such:

```
<foreign language="unknown"></foreign>
```

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

When annotators encounter a proper name whose spelling they are not sure of, they should spend a moment or two searching for the proper spelling. If after a short while they cannot find the correct spelling of the proper noun in question, they should make their best guess and use a double caret ^^ instead of a single caret ^ to mark the name. For example:

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

Arabic Interjections

ah	ha	mhm	yA
Ah	Hay	O	yaa
aha	hE	OhO	yO
ayyO	hi	uh	yOO
Eyy	ih	wAw	yuu

Chinese Interjections

啊	a1	'surprise, praise'
啊	a2	'questioning'
啊	a3	'disbelief'
啊	a4	'answer; surprise; praise'
哎/噯	ai1	'dissatisfaction'
哎呀	ai1ya1	'surprise; complaining'
哎哟	ai1yo1	'surprise; agony'
哎	ai2	'emphasis'(2)
噯/哎	ai3	'disagreement; denying'
噯/哎	ai4	'regret'
唉	ai4	'disappointment'
哈	ha1	'triumphant'
哈哈	ha1ha1	'triumphant'
咳	hai1	'sadness; regret'
嘿/嗨	hei1	'drawing attention'
呵/喏	he1	'surprise'

哼	hng5	'dissatisfaction'
噃/唔	n2/ng2	'query'
噃	n3/ng3	'out of expectation'
噃	n4/ng4	'answer' (3)
噃/噢	o1	'understanding'
噃/噢/唷	o1yo1	'surprise'
哦	o2	'half belief half doubt'
哦/噃	o4	'understanding'
哂	pei1	'discarding; scolding'
哇	wa1/wa3	'surprise'
哟	yo1	'slight surprise'

3.2.4.6 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 thought he -- 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. <i>apropos</i> .
	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 try to 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.

Annotators should not try to 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 Second Passing

4.1 Introduction

Second passing is used as a quality control measure to ensure the accuracy of segmentation, transcription (including markup), and speaker identification. After the initial file has been fully segmented and transcribed, a new annotator listens to the entire broadcast or telephone call while viewing the corresponding transcript, and makes adjustments to the timestamps or transcription as needed. Second passing entails a mix of manual and programmatic checks on the transcript files. The particular types of checks conducted during second passing are described below.

4.2 Segmentation and Speaker ID Verification

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

For telephone speech, annotators listen to the entire call to ensure that all speech for each channel is captured within a turn segment and that no speech remains outside of a segment boundary.

For broadcast news speech, annotators listen to the entire broadcast, including <sn> (commercial) sections, to ensure that no news or teaser section has gone unsegmented or untranscribed. Annotators also verify that the correct section labels (<sr>, <sn>) have been used, that the overlapping speech convention has been correctly applied, and that each new turn or segment boundary has a corresponding speaker ID.

Each speaker ID is further checked to ensure that the same speaker name is spelled consistently within each file, and that the same speaker type and native/non-native speaker status has been recorded for each instance of that speaker.

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 particular attention to a handful of areas that are particularly difficult to transcribe, in particular 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 also employ a series of programmatic checks known as a syntax check. The syntax check scans the file for common segmentation and transcription errors as well as errors of data formatting. The checks include:

Timestamping

- timestamps are out of linear order
- timestamps without text data
- timestamps followed by non-empty lines
- multiple section, turn or breakpoint boundaries on a single line

Speaker IDs

- missing or badly formatted speaker IDs
- missing or badly formatted speaker type

Transcription

- 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 makes any necessary corrections.