

**Guidelines for Quick Transcription of
Broadcast Conversations
Mandarin Chinese**

Version 1.1 – April 20, 2005

Linguistic Data Consortium

www ldc upenn edu/

Table of Contents

1	Introduction	4
1.1	Data	4
2	Segmentation	4
2.1	Overview	4
2.2	Broadcast Conversation	5
2.2.1	Overview	5
2.2.2	Sections	5
2.2.2.1	Section types	5
2.2.2.1.1	Introduction	5
2.2.2.1.2	News stories <sr>	5
2.2.2.1.3	Non-news <sn>	6
2.2.2.2	Beginning of a section	6
2.2.2.3	End of a section	6
2.2.2.4	End of the file	6
2.2.3	Turns	7
2.2.3.1	Introduction	7
2.2.3.2	New speaker turns	7
2.2.3.3	End of a turn	7
2.2.3.4	Overlapping speech	7
2.2.3.5	Simultaneous speech	8
2.2.3.6	Short periods of non-speech	8
2.2.3.7	Extended periods of non-speech	8
2.2.4	Breakpoints	8
2.2.4.1	Overview	8
2.2.5	Segment boundary symbols in broadcast conversation	9
2.2.6	Speaker Identification	9
2.2.6.1	Introduction	9
2.2.6.2	Speaker Type	9
2.2.6.3	Names and Identifiers	9
2.2.6.4	Native and non-native speakers	10
2.2.6.5	Examples of speaker IDs	10
3	Transcription	10
3.1	Transcription Conventions	10
3.2	Orthography and spelling	10
3.2.1	Capitalization	10
3.2.2	Spelling	10
3.2.2.1	Mispronounced words	11
3.2.3	Contractions	11
3.2.4	Numbers	11
3.2.5	Abbreviations	11
3.2.6	Acronyms and spoken letters	11
3.2.7	Disfluent speech	11
3.2.7.1	Introduction	11
3.2.7.2	Filled pauses and hesitation sounds	12
3.2.7.3	Partial words	12

3.3	Additional considerations	12
3.3.1	Noise	12
3.3.2	Hard-to-understand sections	12
3.3.3	Foreign languages	13
3.3.4	Interjections	13
3.3.4.1	Chinese Interjections	13
3.3.5	Speaker errors and non-standard usage	1
Appendix		2
3.4	Summary of Conventions	2

1 Introduction

Quick transcription is an approach to transcription that emphasizes speed and accuracy. The goal of quick transcription (hereafter QTR) is to produce an accurate, time-aligned transcript as quickly and efficiently as possible. To this end, the Linguistic Data Consortium has created a simplified version of more rigorous transcription conventions, excluding special markup and multiple quality checks in favor of a single, focused transcription pass.

This document will describe the Linguistic Data Consortium's quick transcription standards for English broadcast conversation (hereafter BC) data.

1.1 Data

The recordings in the Chinese Broadcast Conversation Quick Transcription collection contain between multiple speakers per broadcast program, on one audio channel. Each speaker is identified by a unique speaker ID throughout each audio file.

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
- **speaker turns** (change of speaker)

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", and "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 Conversation

2.2.1 Overview

Broadcast conversation data is recorded on 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. Overlapping speech regions occur frequently in BC data.

Broadcast conversation 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. <sr> sections are equivalent to report sections in Transcriber.
- <sn> refers to non-news sections, including commercials. <sn> sections are equivalent to non-trans sections in Transcriber.

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 Conversation, each <t> segment must also receive a speaker ID.

For Broadcast Conversation 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. The Transcriber tool has a built-in convention for transcribing overlapping or simultaneous speech. At the insertion of a new turn, transcribers will identify the second (or interrupting) speaker, and will transcribe the overlap as completely as possible.

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> <<male, Peter_Jennings + female, Paula_Zahn>>
1: the global warming situation
2: Well, because I don't --
<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> <<female, speaker_1 + male, non-native,
Jacques_Cousteau>>
1: ...
2: ...
<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, transcribers insert an <e> end section tag at the start of the non-speech region, 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 Segment boundary symbols in broadcast conversation

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)

Table 1: Description of broadcast conversation section tags

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 Chinese broadcast conversation, native is defined as a speaker of any standard Mainland Chinese dialect. As native is the default, this is not explicitly marked. Non-native is used for speakers of other dialects of Chinese, including Cantonese; non-native is also used to indicate people who are not native Chinese 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>>
```

3 Transcription

3.1 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 Chinese spelling 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 Orthography and spelling

3.2.1 Capitalization

There is no capitalization in written Chinese.

3.2.2 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.

3.2.2.1 Mispronounced words

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

3.2.3 Contractions

There are no contractions in Chinese.

3.2.4 Numbers

All numerals are written out as complete Chinese characters as they are pronounced, NOT in Arabic numerals:

一九九七
一千九百九十七
幺三三四五六

3.2.5 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 in these English examples:

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.

3.2.6 Acronyms and spoken letters

Acronyms and spoken letters should be capitalized.

AIDS
FEMA
CIA
~WWW dot ~ABC news dot com

3.2.7 Disfluent speech

3.2.7.1 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 one or two times, then move on.

3.2.7.2 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.

Chinese Filled Pauses
呵
呃
唔

Table 2: Chinese Filled Pauses

3.2.7.3 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:

```
<t 145.67> <<male, non-native, Tony_Blair>>  
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.:

```
<t 145.67> <<male, non-native, Tony_Blair>>  
(( ))
```

3.3.3 Foreign languages

Portions of speech in any language *other* than Mandarin are annotated using the <language> text </language> convention to indicate the language and to transcribe the words that are spoken in that language. For instance:

我要买一部新的<English> notebook</English>.
喂, <Cantonese> 不要搞 </Cantonese>.

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

那个东西真 <foreign> (()) </foreign>.

3.3.4 Interjections

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

3.3.4.1 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'
嗯/唔	n2/ng2	'query'
嗯	n3/ng3	'out of expectation'
嗯	n4/ng4	'answer' (3)
喔/噢	o1	'understanding'
喔/噢唷	o1yo1	'surprise'
哦	o2	'half belief half doubt'
哦/喔	o4	'understanding'
呸	pei1	'discarding; scolding'
哇	wa1/wa3	'surprise'
哟	yo1	'slight surprise'

3.3.5 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.

Appendix

3.4 Summary of Conventions

Category	Condition	Markup	Example	Explanation
Orthography and spelling	Numbers	Spelled out	一九九七 一千九百九十七 幺三三四五六	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 呵, 呃, 唔
	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	哼, 啊	Use standardized spellings

Table 3: Summary of transcription conventions