Quick Rich Transcription (QRTR) Specification for Chinese Broadcast Data

(XTrans-Format Version)

Version 3 – March 5, 2008

Linguistic Data Consortium

http://www.ldc.upenn.edu/GALE/Transcription

Table of Contents:

1   Introduction and Overview

The goal of quick rich transcription (QRTR) for broadcast news and broadcast conversation is to produce a verbatim, time-aligned transcript with minimal but useful markup. QRTR also identifies some salient structural features of the broadcast and provides speaker identification.

The elements of a quick rich transcript include:
- verbatim transcription
- time-aligned section boundaries, speaker turns and sentences (segmentation)
- section and sentence type identification
- speaker identification
- standard treatment of common spoken phenomena

Transcription begins with audio segmentation. This involves "timestamping" structural boundaries including sections (i.e., story transitions), speaker turns and sentence units (SUs). Speakers are identified by name where possible, or by a unique identifier, and other speaker traits like sex are noted. Once audio has been virtually segmented into smaller units, annotators transcribe the content of each segment. Special conventions are used to flag certain speech phenomena like disfluencies and mispronounced words. Quality control checks verify the completeness and accuracy of segmentation and transcription.

QRTR differs from Quick Transcription (QTR) in that each sentence unit is timestamped and labeled for its type. QRTR differs from careful transcription (CTR) in the amount of detail contained in the transcript markup, the number of features identified, the degree of accuracy and completeness of the transcript, the amount of time taken to complete the file, and the number of quality checks that are performed on the finished product.

Please see LDC's transcription website for links to guidelines for the various transcription tasks: http://www.ldc.upenn.edu/Projects/Transcription

2   Data

These guidelines pertain to data in the following genres:

- *Broadcast News (BN)* consisting of "talking head"-style news broadcasts from radio and/or television networks.
- *Broadcast Conversation (BC)* consisting of talk shows, interviews, roundtable discussions and other interactive-style broadcasts from radio and/or television networks.

Data is divided into files, which typically correspond to a recording of one broadcast from a single program. Files are typically 30 to 60 minutes in duration, though they may be of any length. Files come from a range of radio, television,

satellite and web broadcast sources from around the world. Each show is pre-designated as BN or BC based on its characteristic content. Note however that BN shows can sometimes contain stories that are conversational, while BC shows can include hard news reports.

3 Segmentation Task

3.1 Introduction

Transcription begins with segmentation. During the segmentation task, annotators virtually chop an audio recording into smaller units that correspond to certain features of the broadcast, for instance sentence units or speaker turns. Each segment must be timestamped – that is, time-aligned with the audio – to identify where the segment starts and ends. In most cases in broadcast audio, the end of one segment is also the beginning of the next. Segments are also classified by type and subtype. We identify three kinds of segments in the QRTR task: Sections, Turns, and Sentence Units. These are arranged hierarchically (sections contain turns, turns contain sentences).

It is suggested that annotators begin segmentation by identifying the most fine-grained segment type, sentence units (SUs). SU boundaries frequently occur at natural boundaries in the audio (pauses, breaths, speaker turns), which makes segmentation easier. This is not always the case, especially for complex or atypical SUs, and annotators will need to fine-tune some SU boundaries once they have completed transcription. As segments are created, XTrans will prompt the annotator to supply SpeakerID information, and the annotator will also indicate section (story and commercial) boundaries as encounter them. The sections that follow provide detailed information about each step of the process.

Annotators should note that segmentation in XTrans can be done with the keyboard only, with the mouse only, or with a combination of both. After you've become familiar with basic XTrans functionality, you will find that using only the keyboard is both faster and more intuitive than using the mouse.

3.2 Timestamping the Audio

Timestamps are required for all segments. In XTrans, annotators create a timestamped segment simply by marking the appropriate region of audio in the waveform display, then inserting the selected segment[1]. Timestamps are designated in seconds, rounded to the nearest thousandth of a second. Note that while XTrans does not show start/end timestamps within the transcript display, the waveform display includes a color-coded horizontal bar representing each segment, along with its start time, end time and duration.

---

[1] Detailed instructions for using the XTrans toolkit are available in "Using XTrans for Broadcast Transcription: A User Manual" distributed with the XTrans package and available from LDC's transcription website: http://www.ldc.upenn.edu/Projects/Transcription: (http://projects.ldc.upenn.edu/gale/Transcription/download_xtrans-linux-latest.php; http://projects.ldc.upenn.edu/gale/Transcription/download_xtrans-windows-latest.php)

Because broadcast speech recordings use a single audio channel, segments occur one right after the other, in direct succession and typically without intervening periods of unsegmented audio (silence). Small gaps in the succession of segments should indicate an untranscribed event, like a commercial, music, sound effects or background noise.[2] All speech and other material to be transcribed must be segmented.[3]

Timestamps should always be placed in between words, not inside of them or at the very edges of words where speech sounds could be truncated. Good places to insert timestamps are during pauses, breaths or other non-speech events, which typically occur at sentence unit (SU) boundaries. Finally, it is critical that the time and the audio event are properly aligned, so that the words transcribed within each segment match the speech associated with that segment.

3.3 What to Segment

All broadcast speech must be segmented and classified into sections (news reports, conversational segments or non-news). News reports and conversational segments must also be segmented into SUs, with speakerIDs added. Non-news sections like commercials should **not** be segmented into smaller units or labeled for speakerID, and they should not be transcribed.

Very brief (under 0.5 seconds) periods of silence, music, background noise or other types of non-speech that occur while someone is speaking should simply be included within that SU segment, or split between two adjoining speaker SU segments. No other treatment is necessary. Lengthy segments of non-speech (like sound effects) that interrupt a speaker's turn, or that come in between speaker turns, should be separated out and left unsegmented. Note that annotators should make an effort to leave SU segments intact; that is, avoid splitting a single SU into multiple segments even when it includes a lengthy pause.

3.4 Segmenting Overlapping and Simultaneous Speech

In broadcast audio, overlapping speech from two or more speakers is a relatively frequent occurrence. Although broadcast files contain a single audio channel, within XTrans each unique speaker in a file is assigned a separate *virtual* channel. Transcribers can simply create overlapping segments two or more distinct speakers using the normal XTrans functionality. Overlapping segments are represented in the waveform display as overlapping horizontal bars, as shown in the image below.

---

[2] Note that using the mouse for segmentation makes it easier to leave unintended small gaps in consecutive segments of continuous speech. Using the keyboard shortcuts for segmentation avoids this problem.
[3] The LAG (Listen All Gaps) feature in XTrans allows annotators to review all unsegmented material in a file.
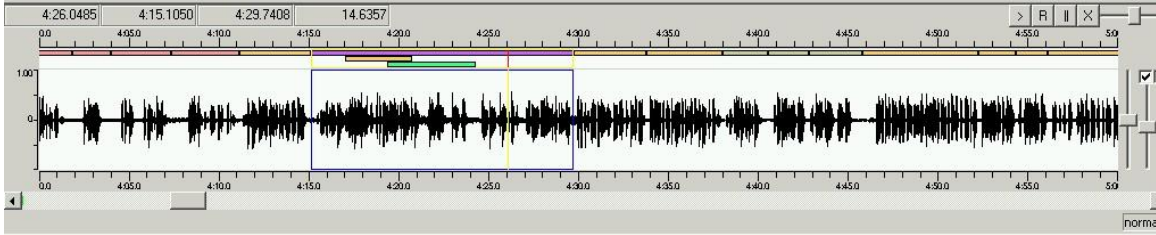
4 Sentence Units (SU)

Segmentation begins with identification of sentence unit boundaries. A sentence unit (SU) is a natural grouping of words produced by a single speaker. SUs have semantic cohesion – that is, they can have some inherent meaning when taken in isolation; and they have syntactic cohesion – that is, they have some grammatical structure[4]. In written language, sentences are usually designated by punctuation like periods or question marks. When creating SU boundaries for spoken language, our goal is to identify a semantically and syntactically cohesive group of words that constitute a reasonable sentence-like unit. Sentence units are the most basic kind of segment in the QRTR task. Each SU should be contained within its own segment. Segments should not contain multiple SUs, and single SUs should not be divided across multiple segments.

Transcribers pay close attention to changing subjects and ideas, in addition to connecting words and pauses, to determine where to insert the SU boundary.

We distinguish three types of SUs: statements, questions and incomplete sentences. After identifying the boundaries of an SU and creating a corresponding segment, annotators can use XTrans to assign the segment type. In general, the SU segment types are consistent with standard end-of-sentence punctuation used during transcription, as follows:

Expected Sentence-final Punctuation	SU Type	Symbol
period	end-of-sentence markup for Statement SUs	.
question mark	end-of-sentence markup for Question SUs	?
double dash	end-of-sentence markup for Incomplete SUs	--

Annotators will note that standard punctuation typically includes commas as well. For purposes of the QRTR task, we do not identify an SU (or sub-SU) unit that corresponds to a comma. Commas may be added into transcripts for human readability, but it should be understood that the existence of a comma does not imply the existence of a sentence unit. See **Section 7.1.2** for additional discussion of punctuation in QRTR transcripts.

The sections that follow provide specific rules for identifying SUs of each type.

---

[4] Note however that incomplete SUs may contain incomplete semantic and/or syntactic content.

4.1.1 Statement SUs

Statements are declarative sentences or fragments, and are usually punctuated by a period or exclamation point. For instance,

> Speaker1: 我们今天的客人高瑶就是这样一位成功的女性.
> Speaker1: 高瑶呢曾经在国际卫星组织担任过主管.
> Speaker1: 她现在是一家叫做 〈foreign lang="English"〉〈/foreign〉公司担任亚太事务总监.
>
> Speaker1: 他昨天被选为主席.
> Speaker2: 对，昨天.

Long statements with multiple verbs are very common in Chinese. In these cases, annotators should use their judgment about whether the verb change warrants a new statement SU.

See **Section 4.1.4** for additional guidelines on determining SU boundaries.

4.1.1.1 Backchannel SUs

A backchannel is a word or phrase that provides feedback to the dominant speaker, indicating that the non-dominant speaker is still listening. In QRTR, backchannels are treated as statement SUs. When a speaker chains together several backchannels in succession, annotators tag them as a single statement SU. For instance,

> Speaker1: 所以现在变成说王金平就变成非打不可 .
> Speaker2: **嗯.**
>
> Speaker1: 可是马英九马英九可能 −
> Speaker1: 干吗马王配啊?
> Speaker2: 文化女性 .
> Speaker1: 我找一个比我年纪的， 比你形象好的，能够听我的.
> Speaker2: **对对对.**

4.1.2 Question SUs

The question label should be used for a complete sentence that functions as an interrogative. The expected end-of-sentence punctuation for a question is a question mark. In Chinese, one of several sentence-final words may indicate a question SU, such as 吗, 呢, 吧. If these interrogative particles are not present, other words or phrases may signal the presence of a question SU, for instance 什么，哪里，谁，为什么 etc.

A tag question is a phrase added to the end of an utterance that invites the listener to give feedback. Tag questions usually do not stand alone as a question,

but rather form a complete question with the previous utterance. Rhetorical questions should also receive a Question SU label.

> Speaker1：那许老师你不是站在这些网友的对立面了吗?
>
> Speaker2：你怎么看连战的大陆行?
> Speaker2：你表示肯定呢，赞许呢还是支持呢?
>
> Speaker3：因为你要想你不断的加息, 不断的加税，苦了谁?
> Speaker3：苦了老百姓.
>
> Speaker4：你想不想去?
>
> Speaker5：这个细节呢, 你咋一看, 是很奇怪，很做作，对不对?

The question SU label should only be used when the utterance is clearly asking a question or functioning as a tag or rhetorical question. If you are unsure whether the SU is functioning as a statement or a question, you should label it as a statement.

4.1.3 Incomplete SUs

When an utterance does not constitute a grammatically complete sentence **and** does not express a complete thought, it is labeled as an incomplete sentence unit and marked with double dashes (--).

Incomplete SUs frequently occur in two situations. When a speaker interrupts him/herself and then restructures the utterance and continues speaking on the same topic, an incomplete SU exists. In other cases, the speaker may trail off at the end of his/her turn and abandons the utterance completely, without restructuring it or continuing along the same lines. For instance:

> Speaker1：还有的人很多很多人就说，我的小时候就是这样过的,
> Speaker1：**没有这么—**
> Speaker1：嗯嗯我们小时候充满阳光嘛.

The other frequent case of incomplete SU occurs when one speaker's turn is cut short by an interruption from the other speaker, as in the following:

> Speaker2：**那陈水扁当然也** --
> Speaker1：陈水扁的个性是不甘示弱的啦.
> Speaker1：陈水扁就说呢就算李登辉嗯李登辉过去十二年也没有啊台独.
>
> Speaker1：他如果连这场仗都打的这个灰头土脸的话，说实在二千零八年，
> **我看他也就** --

Speaker2：到时候民进党就更厉害呀.
Speaker1：对.
Speaker1：没错没错.
Speaker1：他就必需要好好地去检讨去思考了.

Be careful not to confuse incomplete SUs with sentence fragments that express a complete thought (for instance a response to a question that is expressed as a phrase rather than a complete sentence.) Sentence fragments that express a complete thought and show no signs of being caused by an interruption or by the speaker simply trailing off, should be labeled as statement SUs.

4.1.4 Recognizing SU Boundaries

It can sometimes be difficult to determine where a sentence unit boundary exists and when to place two clauses within the same SU. Annotators should rely primarily on the meaning conveyed by the utterance and apply SU breaks in accordance with the rules described in these guidelines. However, annotators may sometimes rely on prosodic features like sentence intonation or pauses to determine where to place an SU boundary. In practice, SU boundaries tend to occur at the ends of fragments, simple sentences and complex sentences.

Speaker1：呃，不知道他们今天是不是可能不用上班了今天.
Speaker1：据说以前在中国 是这个三八妇女节这个妇女可以休息一天或者是半天，然后回家去做饭.
Speaker2：半天吧.

Speaker1：他家三口人.

Speaker1：我记得那个时候是啊当时被提名到最高法院的那个法官叫克莱姆斯.汤姆斯，是因为他的以前他的一个同事一个女同事告他性骚扰.

In Chinese we frequently see a subject introduced in the first sentence of a narrative and then dropped repeatedly from subsequent sentences. In such cases, annotators should rely on the context of the utterance to make a decision about where to put SU boundaries. In principle, we should treat the clauses with omitted subjects as one sentence; however, if the sentence segment is over 20 seconds, it might be okay to break the clauses where the subject is omitted. The examples that follow illustrate (1) a long utterance that should be broken in two, and (2) a shorter utterance that is treated as a single statement SU.

Speaker 1：黄沙水产市场表示，将与这两个养殖场签订保证书，规定凡在市场申请销售多宝鱼的经营业户必须出具五种证件，同时，要经过广州市农业局检验中心检验合格后才允许经营销售。

This utterance is quite long. It can be divided in two parts, as shown here:

Speaker1：黄沙水产市场表示，将与这两个养殖场签订保证书，规定凡在市场申请销售多宝鱼的经营业户必须出具五种证件，

Speaker1：同时，要经过广州市农业局检验中心检验合格后才允许经营销售。

Both units of the original long utterance can "stand alone" in this context.

However, in the next example, the utterance should be treated as a single segment (since the segment is short) even if the subject is omitted twice in the latter half of the sentence.

Speaker1：未来三周香港渔护署加紧巡视铜锣湾， 检查有没有野鸟不寻常死亡， 并加强监察售卖宠物鸟的地方。

Transcribers look for conjunctions in the utterance to identify a plausible division between two clauses that can "stand alone" semantically and syntactically.

Below are some additional rules of thumb annotators can follow in deciding where to put SU boundaries:

1. Rely on temporal expressions. If two clauses share a subject but have different temporal information, they should be treated as <u>separate SUs</u>.
2. When the subject changes from one clause to the next, create a new SU.
3. If subjects are overtly expressed in two adjacent clauses, treat the clauses as separate SUs.
4. For quotations, there is one SU break after the quoted material. However, if the quotation itself contains multiple sentences, break up the quote into separate SUs.

5   Identifying Section Boundaries

The QRTR task also calls for identification of section boundaries. A section is a topically contiguous segment of the broadcast. Sections begin at SU boundaries. At the beginning of each new section, annotators simply insert the appropriate section label. Consecutive sections of the same type should receive separate section boundary labels, except in the case of consecutive commercials and other untranscribed segments which should be grouped together as a single (untranscribed) section. All audio in a speech file must be assigned to a section.

We recognize three section types:

- **Reports** include typical "talking head" news broadcast, with an anchor reading the news. This may also include broadcasts from reporters in the field. News reports 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 broadcast, it is often difficult to tell where one story ends and the next begins. Annotators should rely on audio cues (speaker changes, music, pauses) to inform their judgments. When in doubt, **do not** create a new section boundary.

- **Conversations** include highly interactive segments of a broadcast, including roundtable discussions, interviews, call-in segments, debates and the like. Some conversation sections are quite long and can contain multiple topics. Annotators should create a new section boundary only at natural breaks in the flow of conversation, for instance, when there is a major shift in topic, or when a new panelist joins a roundtable discussion. If in doubt, the annotator should avoid creating a new conversation boundary.

It may sometimes be difficult to tell the difference between a report and a conversational segment. When in doubt, annotators should use **report**.

- **Non-news** text includes segments like commercials, station identifications, public service announcements, promotions for upcoming shows and long musical interludes. ***Note that non-news sections are not segmented, transcribed or further annotated in any way*** *(including speaker ID or SU segmentation).* Once a non-news section has been identified and labeled, it should be ignored for the rest of the transcription task. If multiple non-news sections follow one another within a transcript, they should be grouped together as a single section. This is different from multiple consecutive news or conversational reports, which should be separated into multiple sections.

6 Speaker Identification

In addition to identifying SUs and section boundaries, annotators also label the identity of speakers within a broadcast. Speaker IDs are required with each SU segment[5]. Each speaker label has three elements: speaker type (required), non-native status (optional) and speaker "name" (if available).

6.1 Speaker Type
All speakers must be assigned a 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, non-human (computer) voices, altered voices, unknown speaker sex, etc.

---

[5] The XTrans toolkit requires annotators to provide speaker ID for each SU annotation.

6.2   Names and Identifiers

All speakers must be identified by name. When the name is not known, annotators use a **unique** identifier for each speaker.

When names are known, they should be written out in full (family and personal name). For names with multiple spellings or transliterations, the most common variant should be used. If in common practice the name contains a middle initial or an appositive like "Jr.", these should be included and spelled out in full. All names must be written in English using the most common transliteration. Capitalization should follow standard conventions.

The spelling of speaker IDs must be consistent within a broadcast file, and wherever feasible across different broadcast files as well. It is also important that the spelling of names within a transcript match the spelling of the name in within the speaker ID label. For instance, if the transcript uses the transliteration "Osama bin Laden", then the speaker ID should also use "Osama", not "Usama".

When a speaker is not identified by name within a recording, the speaker should be labeled with a unique numerical identifier, e.g. speaker14. Each anonymous speaker is assigned a unique number that should be used for every instance of that speaker throughout the broadcast. Anonymous speaker IDs cannot be re-used for different speakers in the same file, regardless of gender or speaker type[6].

6.3   Native and Non-native Speakers

In addition to labeling speaker type and name, annotators also indicate when a speaker is non-native; that is, when they use a language variety other than the target, or when they speak the target language with a discernable foreign accent. Targets for the current task are

- o   Arabic – Modern Standard Arabic (MSA)
- o   Chinese – Mainland Mandarin Chinese
- o   English – American English

Speakers using other varieties/dialects of these languages, or speaking these languages with a heavy non-standard accent (for instance, Cantonese-accented Mandarin, or British English) should be marked as non-native.

In the case of Chinese, nearly all speakers will be native speakers of some regional variety rather than native speakers of Putonghua. A native speaker of any Chinese dialect who is talking in Putonghua should be considered "native" for purposes of speakerID labeling. Do not mark native Chinese speakers as "non-native" when they are speaking Putonghua simply because you can detect

---

[6] Note that the LRS (Listen Random Segment) and LAS (Listen All Segments) functions in XTrans are helpful for verifying speakerID assignment.

a regional accent. Only speakers who are clearly **not** native speakers of Putonghua, or who speak Putonghua with a discernable **foreign language** accent, should be considered non-native.

See **Section 7.4.2** for additional discussion of Chinese dialects in broadcast transcripts.

7 Transcription

Quick-rich transcription requires annotators to produce a verbatim transcript of all speech within a file and to add minimal markup to capture salient features of the speech. Standard writing conventions, including orthography, spelling and punctuation, are used for ease of comprehension and readability. Transcripts must be produced in UTF-8 (Unicode) encoding. Transcripts should be spell-checked for common misspellings or typographical errors before they are considered complete.

7.1 Orthography and Spelling

7.1.1 Spelling

Transcribers should use standard Chinese orthography. <u>All files must be checked for typos after transcription is complete</u>. When in doubt about the orthography of a character or a proper name, annotators should consult a standard reference, like an online or paper dictionary, world atlas or news website.

7.1.2 Punctuation

Annotators should include standard punctuation for ease of transcription and reading. Acceptable punctuation is limited to the following:

Type	Usage	Symbol
period	end-of-sentence markup for Statement SUs	.
question mark	end-of-sentence markup for Question SUs	?
double dash	end-of-sentence markup for Incomplete SUs	--
comma	sentence-internal, used to aid readability	,

Transcripts should not contain quotation marks, exclamation marks, colons, semicolons, single (stand-alone) dashes, or ellipses in transcribing. Punctuation should be written as it normally appears in standard writing, with no additional spaces around the punctuation marks.

7.1.3 Numbers

All numerals should be written out as complete words instead of number characters. They should be written as spoken (using the <foreign lang="LANGUAGE"></foreign> tag as needed; see **Section 7.4.1** for more details).

written	pronounced	number character
一	yi	1
一百	yi bai	100
十	shi	10
二十一	er shi yi	21

7.1.4 Proper Nouns

No special markup is required for proper nouns. Note however that spelling of names should be consistent within the transcript, and should match the spelling of the name in within the assigned speaker ID.

For instance, if the speaker ID uses the transliteration "Hu Jintao" the transcript should use "胡锦涛" when that name is spoken, not "胡锦焘" or some other form.

7.1.5 Contractions and Acronyms

These phenomena rarely if ever occur in Chinese and no special guidelines apply.

7.1.6 Spoken Letters

Transcribers may come across English spoken letters. Letters pronounced as a sequence of individual letters should be written in English as they are pronounced marked with a tilde ~, with no space between the letters:

```
我今天考试的了个~A.
I got an A in today's quiz.

他的名字叫布什~BUSH.
His name is Bush, B-U-S-H.

~FBI
~CEO
```

7.1.7 Interjections

The following standardized spellings are used to transcribe interjections. Transcribers may find a lot of variations among speakers. The list below is for reference. Interjections do not require any special markup.

```
啊    a1        'surprise, praise'
啊    a2        'questioning'
啊    a3        'disbelief'
啊    a4        'answer; surprise; praise'
哎/嗳 ai1       'dissatisfaction'
哎呀 ai1ya1   'surprise; complaining'
```

哎哟	ai1yo1	'surprise; agony'
哎	ai2	'emphasis' (2)
嗳/哎	ai3	'disagreement; denying'
嗳/哎	ai4	'regret'
唉	ai4	'disappointment'
哈	ha1	'triumphant'
哈哈	ha1ha1	'triumphant'
咳	hai1	'sadness; regret'
嘿/嗨	hei1	'drawing attention'
呵/嗬	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'

7.2   Disfluent Speech

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. For purposes of QRTR, annotators should not spend too much time trying to precisely capture difficult sections of disfluent speech, but should make their best effort to transcribe what they hear after listening to the segment once or twice, and then move on.

7.2.1   Filled Pauses and Hesitation Sounds

Filled pauses are non-word sounds 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. Instead, there is a restricted set of filled pauses for each language, with established spelling conventions.

For Chinese, filled pauses are limited to 呵, 呃 and 唔.

7.2.2   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. In Chinese, speakers almost always break off at the end of a character, so annotators should simply type the

character that they hear. If the transcriber is sure that the character is part of a word, rather than being spoken in isolation, he/she should put a dash – directly after the character. Note that this should be distinguished from a word which is repeated several times, which should not get any special mark-ups. For example:

> 非-非-非洲大草原有着各种各样的动物。
> `There are all kinds of animals in af- af- Africa.`

7.2.3 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. Just transcribe the word using the standard spelling, adding the plus sign + to signal that the word is pronounced incorrectly.

> 他**+胡** (`pronounced as pu`) 乱说了一通。

Keep in mind that this symbol should only be used for obviously mispronounced words. Dialect pronunciations or other common variants of words should not be marked as mispronunciations.

7.2.4 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 *. This is extremely rare in Chinese.

7.3 Speaker Errors and Non-standard Usage

Annotators should not correct grammatical errors, e.g. "I seen him" for "I saw him". The words must be transcribed as spoken. The same goes for non-standard usage or mis-used words. For instance, if the speaker says 出蓉芙水 but means *出水芙蓉*, transcribe it as 出蓉芙水.

Annotators should transcribe exactly ***what is spoken***, not what they expect to hear or what they consider "correct" speech.

7.4 Foreign Languages and Dialects

7.4.1 Foreign Languages

Portions of speech in any language other than the target language are annotated using the `<foreign lang="LANGUAGE"> text </foreign>` convention to indicate

the language and to transcribe the words that are spoken in that language.[7] Note that southern Chinese dialects are treated as foreign languages. These include Cantonese, Wu, Xiang, Gan, Minnan, and Hakka. For instance:

我要买一部新的<foreign lang="English"> notebook</foreign>.
喂,<foreign lang="Cantonese"> 多谢 </foreign>.

If the annotator does not know the name of the language or what is being said, they should insert "unknown" into the language tag:

那个东西真 <foreign lang="unknown"> </foreign>.

7.4.2 Dialects

Annotators will frequently encounter non-Putonghua dialect in the broadcast conversation programs, such as Henan Hua, Shandong Hua, Sichuan Hua, and Yunnan Hua. Portions of speech in northern Mandarin dialect should be surrounded with a special non-Putonghua marker:

<foreign lang="non-PTH"> text </foreign>

to indicate the use of colloquial dialect. The words should be transcribed as spoken, using standard Chinese orthographic conventions.

If the conversation switches back and forth between Putonghua and dialect, mark the dialect portions using the convention described above, as in

PTH text <foreign lang="non-PTH"> text </foreign> PTH text

SU segmentation is unaffected by the presence of non-PTH speech. A single SU segment may contain all PTH, all non-PTH, or a mix of both.

In the following example, the speaker is speaking Sichuan and Wuhan dialect respectively:

<foreign lang="non-PTH"> 这么热的天啥子也不能做.</foreign>
<foreign lang="non-PTH"> 吓老子一大跳.</foreign>

Putonghua pronounced with a dialectal accent should still be treated as Putonghua. Do not transcribe any accent features, for example, missing retroflex and confusion between lateral and nasal etc, but rather use the standard orthography. No special mark-up is needed, and dialect pronunciation should not be marked as mispronunciation. For example:

---

[7] Note that this convention is the convention in XTrans. The keybinding Ctrl+Shift+h will produce the tag: <foreign lang="English"> </foreign>. If the language is known and is transcribed, annotators update the language and insert text between the tags as appropriate.

```
Speaker says zi1 dao4, transcription should be kept as 知道
Speaker says: lan2fang1, transcription should be kept as 南方
Speaker says: feng4fang2, transcription should be kept as 凤凰
```

7.5   Background and Speaker Noise

Transcribers are not required to specially label background noise or sound effects. Note however the convention for indicating long periods of non-speech within or outside an SU segment (**Section 3.3**).

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

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

7.6   Hard to Understand Regions

Sometimes an audio file will contain a section of speech that is difficult or impossible to understand. In these cases, annotators should 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 area of uncertain transcription with double parentheses:

```
Speaker1: 新政策不允许(()).
```

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. For example:

```
Speaker1: (())
```

Do not simply skip the region without attempting to transcribe it first.

7.7   Final Pointers
1. **Transcribe what you hear, not what you think is correct.**
2. Do not add words if they are not in the audio, and do not delete words that are spoken, even if they are ungrammatical.
3. Do not try to normalize dialectal words.
4. Do not attempt to transcribe accent features. Use standard orthography.
5. Do not skip words that are hard to understand. Use (()).

8 Summary of Conventions

Category	Condition	Markup	Example	Explanation
Orthography and spelling	**Numbers**	Spelled out as complete words	yi, yi bai, shi, er shi yi	Write out in full; use foreign language tags as needed
	Punctuation	Comma, question mark, period, double dash	, ? . --	Limited to these four symbols.
	Individual letters	~tilde in front of letters	~I before ~E ~YMCA	written in English as they are pronounced, marked with a ~tilde, no space between letters
Disfluent speech	**Filled pauses**	No markup	啊, 呃 and 唔	Limited to these 3 words; use standardized spellings
	Partial words	-	非-非-非洲	Speaker-produced partial words are indicated with a dash. Transcribe as much of the word as you hear.
	Incomplete utterance	--	I think he was -- I thought he was there.	Used when the speaker stops short and abandons the utterance completely, restarting with a new sentence.
	Mispronounced words	+	他+胡 (pronounced as pu)	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**	{ }	{cough} {laugh} {sneeze} {lipsmack}	Sounds made by the talker. Limited to these four. ***NOT required markup for QRTR***
	Non-speaker noise	Not used		***NOT required markup for QRTR***
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. ***This is very rare in Chinese***. **NOTE:** Do not use for non-standard dialect terms or misused words.
	Foreign language	<foreign lang="language"> </language>	<foreign lang="English"> Hello. </language>	This is used to indicate foreign speech. If the word is unknown, leave it out. If the language is unknown, write "unknown". <u>DO NOT leave the "Language" definition blank</u>. **NOTE**: Do not use this convention for foreign borrowings that are common in the target language, e.g. *apropos*.
	Interjections	no special markup	唉, 啊, 哦/喔	Use standardized spellings

Appendix 1: Recommended Strategy

There are many different ways to interact with XTrans to create a time-aligned transcript. The following is a synopsis of LDC's recommended strategy for creating broadcast transcripts with XTrans. Note that most of these functions are keyboard rather than mouse-based commands. For quick transcription, it is strongly recommended that transcribers choose keyboard over mouse-based functions as much as possible. This takes a little getting used to but you will find it much faster and easier to use the keyboard only rather than switching between keyboard and mouse (and it's easier on your wrists!). Consult the XTrans user manual for additional information.

Quick Guide for Quick Transcription

1. open audio file	`File > Open audio file`
2. open new transcript file	`File > New`
3. associate audio and transcript	`Edit > Blindly associate transcript to audio`
4. begin playback and mark segment start	`Alt+M`
5. stop playback and mark segment end	`Alt+M`
6. insert segment	`Ctrl+N (Ctrl+Insert on *nix)`
7. assign speaker information	`dialog box (use tab & arrow keys to select options)`

8. create next segment (repeat 4-7). To create segment for same speaker, first select speaker in speaker panel then repeat steps 4-6.

9. assign section boundary	`Ctrl+I Ctrl+S`
10. assign SU type	`Ctrl+I Ctrl+U Ctrl+___`
11. transcribe the segment[8]	
12. save your work frequently	`Alt+F Alt+S`

13. repeat steps 4-12

14. save and exit

---

[8] Some transcribers prefer to fully segment the file the go back and transcribe it; while others prefer to transcribe as they segment.