Quick Rich Transcription (QRTR) Specification for Arabic Broadcast Data

(XTrans-Format Version)

Version 3 – May 22, 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.

Each timestamp begins with a label identifying the segment type and subtype. There are ten possible labels, summarized in the table below. Each is explained more fully in the sections that follow.

[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)

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.

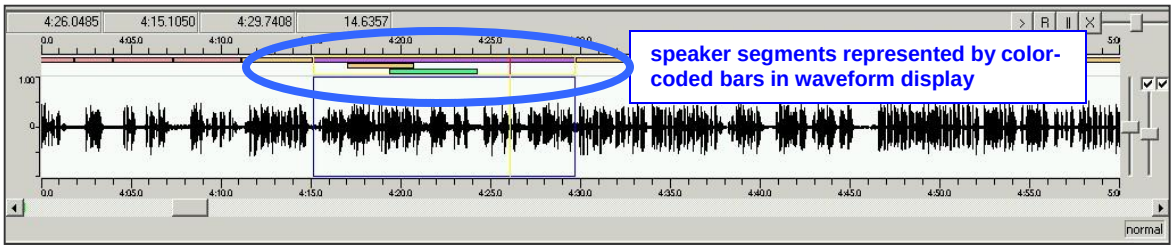
Very brief (under 0.5 seconds) periods of silence, music, background noise or other types of non-speech that occur while someone is speaking should simply be included within that SU segment, or split between two adjoining speaker SU segments. No other treatment is necessary. Lengthy segments of non-speech (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 a sub-sentence-level 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,

كما من جهة أخرى نرى الجوانب الإيجابية الموجوده في الوضع العربى.
On the other hand we do see the positive aspects of Arab society.

موضوع اليوم عن الطلاق.
Today's topic is divorce.

يمكن بدها تعمل ماجستيرأو تخلص دكتوراه.
Maybe she wants to get a master's or complete a doctorate.

لكن هي خطوة بالاتجاه الصحيح
A step in the right direction.

هي خطوة مهمة
An important step.

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 paying attention to the conversation. In QRTR, backchannels are treated as statement SUs. For example,

speaker1: بتعرف الوضع صار صعب عليه .
speaker2: مفهوم .
speaker1: كل شي تغيير بخلال أسبوعين .

speaker1: بتعرف الوضع صار صعب عليه .
speaker2: أهه .
speaker1: كل شي تغيير بخلال أسبوعين .
speaker2: مفهوم .

When a speaker chains together several backchannels in quick succession, annotators tag them as a single statement SU. For instance,

speaker1: هذه خطوة مهمة .
speaker2: إيه إيه مفهوم .

Long statements with multiple verbs are very common in Arabic. In these cases, annotators use a few rules of thumb detailed in this document, to determine whether the verb change requires a new statement SU. See **Section** 4.1.4 for additional guidelines on determining SU boundaries.

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.

دكتورة أمين هل الأطفال مثلا أكثر عرضة لوقع الصدمة من الكبار؟
Dr. Amin, are children more susceptible to trauma than adults?

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:

صار زمان بتشتغل هونيك ولا لأ ؟
You've been working there for years or not?

وله علاج مش هيك ؟
It has a cure, doesn't it?

المشكلة العربية الإسرائيلية أتعقدت، مش ده رأيك برضه ؟
The Israeli Arab problem is highly complicated, isn't this your opinion also?

Rhetorical questions should also receive a Question SU label:

ألا يقولون ليس هناك سلام رديء ولا حرب جيدة ؟
Isn't it said that peace is always acceptable and there is no such thing as a good war?

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 SU. Incomplete SUs frequently occur when a speaker self-interrupts or is interrupted. When a speaker interrupts him/herself and then restructures the utterance and continues speaking on the same topic. 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:

انا قلت لل --
الموضوع ده انا مش موافق عليه ابدا .
I said to --
I am not in agreement with this subject at all.

الطريق الوحيد هو بضمان وحدة العراقيين بجميع الطوائف و و--
The only way is ensuring Iraqi unity with all its sects and and --

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:

speaker1:	يمكن أولادهم يمكن أولاد أولادهم يمكن اولاد اولاد--
speaker2:	دكتورة أمين دعني أستوقفك لنتكلم مع ضيفتنا من القاهرة .

speaker1:	Their children, their grandchildren their great-great-grand- --
speaker2:	Dr Amin, let me interrupt you to introduce our guest from Cairo.

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 when to insert an SU boundary 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.

Complex sentences are very common in spoken Arabic and can be tricky to segment into SUs. In general, annotators should lean toward creating a single SU for complex, multi-part sentences. This is particularly true when two parts (clauses) of the sentence depend on one another for the completion of an idea, for instance:

(1)

مش بس منلوث مياه البحر بس كمان منخلي المجاري تصب رأسا بالبحر.

Not only should we not pollute the water, but we should also not let the sewers empty straight into the sea.

(2)

على مستوى المشاعرجزء كبير من الاشخاص يصاب بقلق شديد لدرجة أنه لا يستطيع النوم ليلا ان لم يأخذ منوم أو مهدىء للأعصاب.

As far as emotional reactions go, some people are unable to sleep without a sleeping pill or a sedative.

(3)

انفجرت قنابل بنفق قرب الفندق ودمرت السيارات الواقفة بالجوار.

A bomb exploded in a tunnel near the hotel, causing damages to many cars in the area.

(4)

مصادر الشرطة ذكرت أن شخصا فجر سيارة كان يقودها قرب الحافلة العسكرية بمحيط مطار المدينة في ساعة مبكرة من صباح اليوم مما أدى إلى إصابة الجنود إضافة إلى مدنيين إثنين.
Police sources said that one person blew up the car that he was driving near the military bus circling the city airport in the early hours of the morning, leading to the injury of the soldiers in addition to two civilians.

(5)
نستضيف من كربلاء المقدسة من العراق الأستاذ فؤاد الدوركي الباحث الإسلامي ومعنا أيضا من مدينة قم المقدسة السيد حسين المحكيم أستاذ الحوزة العلمية.
We host from Holy Karbala, from Iraq, Professor Fuad Al Dourki, the Islamic researcher, and also with us from the Holy City of Qum, Mr. Hussein Al Hakim, Professor of the scientific Hawza.

(6)
ولهذا حينما نلاحظ الإمام الحسين عليه السلام ونأخذه هو بشكل خاص كنموذج أيضا نجد تغيرا في موقفه عليه السلام.
That is why, when we look at Imam Al-Hussein, peace be upon him, and take him in particular as a model, we also find a change in his attitude, peace be upon him.

أمابالنسبة إلى الإجابة على سؤالك تفضلت إنه ما هي الأدلة ما يحتاج إلى أدلة يا حبيبي.
As for your question, when you said what is the evidence, it’s that it does not need evidence, dear.

In Arabic we frequently see a subject introduced in the first clause of a narrative and then dropped repeatedly from subsequent clauses. In such cases, annotators treat each clause as a sentence, as the following examples show:

(1)
دعت الجمعية الطبية الأطباء لمؤتمر طبي.
The Medical society invited the physicians to a conference.
وناقشت موضوعات طبية ساخنة.
And discussed hot medical topics.
وفي النهاية دعت الأطباء الي العشاء.
And by the end invited the physicians for a dinner.

(2)
قام الرئيس جورج بوش بزيارة الى فرنسا.
President Bush went on a visit to France.
وتقابل مع الرئيس الفرنسى لبحث المشكلة العراقية.
And met the French president to discuss the Iraq situation.

(3)
أعلن إسماعيل هنية رئيس الوزراء الفلسطيني أنه سيرفع الغطاء عن أي شخص أو جهة تخرق اتفاق وقف إطلاق النار بين حركتي حماس وفتح
The Palestinian Prime Minister, Ismail Haniyeh, announced that he is going to expose any person or organization that violates the ceasefire agreement between Hamas and Fatah.
وأكد على أن هذا الإتفاق يشكل صفحة جديدة نحو المحافظة على الهدوء

```
And emphasized that this agreement will set up a new phase in
maintaining tranquility.
```

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, and 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.

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

[5] The XTrans toolkit requires annotators to provide speaker ID for each SU annotation.
[6] Note that the LRS (Listen Random Segment) and LAS (Listen All Segments) functions in XTrans are helpful for verifying speakerID assignment.

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

Speakers using other varieties/dialects of these languages, or speaking these languages with a heavy foreign language/dialect accent (for instance, French-accented Arabic, or British English) should be marked as non-native.

In the case of Arabic, all speakers will be native speakers of some regional variety of Arabic (e.g., Egyptian Arabic or Gulf Arabic) rather than native speakers of MSA. A native speaker of any Arabic dialect who is talking in MSA should be considered "native" for purposes of speakerID labeling. Do not mark native Arabic speakers as "non-native" when they are speaking MSA simply because you can detect a regional accent. Only speakers who are clearly **not** native speakers of Arabic, or who speak Arabic with a discernable **foreign language** accent, should be considered non-native.

See **Section 7.4** for additional discussion of Arabic 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 MSA orthography, word segmentation and word spelling. All files must be checked for typos and misspellings after transcription is complete. When in doubt about the spelling of a word or name, annotators should consult a standard reference, like an online or paper dictionary, world atlas or news website[7].

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	,

[7] The latest version XTrans also includes an Arabic spell-checker.

Transcripts should **not** contain quotation marks (?), exclamation marks (!), colons (:), semicolons (;), single (stand-alone) dashes (-), or ellipses (...). 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"> or <non-MSA> tag as needed; see **Section 7.4** for more details on foreign or dialectal speech).

Transcribed as ...	Language variety	Pronounced as ...	Numeral
إحدى عشرة	MSA (feminine)	<iHdaY Ea$rap(a)	11
أحد عشر	MSA (masculine)	>aHada EaSar(a)	11
<non-MSA> إحدعش </non-MSA>	Baghdad (Gulf)	<iHdaEa$	11
<non-MSA> إهدعش </non-MSA>	Baghdad (Gulf)	<ihdaEi$	11
<non-MSA> دعش </non-MSA>	Baghdad (Gulf), Levantine)	daEa$	11
<non-MSA> إدّعش </non-MSA>	Baghdad (Gulf)	<id~aEa$	11
<non-MSA> إيدعش </non-MSA>	Mosul (Gulf)	<iydaEi$	11
<non-MSA> إحداشر </non-MSA>	Egyptian (Nile)	<iHdA$ar	11
<foreign="English"> one </foreign>	English	one	1

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 "Osama bin Laden" the transcript should also use "Osama" when that name is spoken, not "Usama" or some other form.

7.1.5 Contractions

Contractions are extremely rare in Arabic. Annotators should limit their use to cases where they are actually produced by the speaker. In those rare cases, annotators must take care to transcribe exactly what the speaker says and what they hear using standard orthography.

Contracted form	Expanded form	English gloss
شقتله ؟	اش قلت له؟	What did you say to him?
نصّ بتّ	نصف بنت	half-daughter
خلّروح	خلّي نروح	Let's go

7.1.6 Acronyms

For acronyms pronounced as a single word, write them as they are pronounced:

Acronym	Transcribed as...
NASA	ناسا

AIDS	إيدز
UNESCO	يونيسكو
UNICEF	يونيسف

7.1.7 Spoken Letters

Abbreviations that are normally written as a single word, but are pronounced as a sequence of individual letters, should be written in Arabic as they are pronounced, with a space between the letters.

Note that the Arabic letters for English letters 'j' and 'n' should **not** be written as ج and ن but as full words, جيم and نون.

Abbreviation	Pronounced as...	Transcribed as...
IBM	ay by am	ای بی ام
UN	u an	یو إن
CIA	cy ay ayh	سی ای ایه

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 the 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 Arabic, filled pauses are limited to the following:

Transcribed as...	Pronounced as...	English gloss
أه	h	ah
إيه	<yh	eh
أم	>m	um
أوو	>ww	ooh
مم	mm	hm

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 understood. A single dash (-) **attached to the partial word** is used to indicate at which point the word was broken off.

ما زالت مس- مستمرة

```
It is continu- continuing.
```

ولا بد من أن تتّحد الأم- الأمة العربية
```
The Arab natio- nation must unite.
```

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.

Transcribed as...	Pronounced as...	English gloss
+شكولاتة	شكلاطة	
+سلاطة	سلاتة	
+جغرافية	طغرافية	

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 *. For instance,

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

إنت لية بُلُمْ

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, e.g.

```
(1)
```
اش قد أسعار الكلب الحراسة ?
```
How much the guarding dogs?
```

```
(2)
```
حيث انتصبت أقفاصا للكلاب و القطط والطيور ?
```
 Where cages for dogs, cats, and birds installed?
```

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, if the language is known by the transcriber.[8] For example,

```
<foreign lang="English"> Hello. </foreign>
<foreign lang="French"> Bon soir. </foreign>
```

If the annotator does not know the name of the language or what is being said, he or she inserts `"unknown"` into the foreign language tag:

```
<foreign lang="unknown"> </foreign> اش قلت
```

Note that borrowings from other languages are not marked as foreign language, but should be transcribed in Arabic. Usually these words have Arabic morphological markers. For instance:

Borrowed word *(language)*	**Transcribed as...**
computer *(English)*	اشتريت كمبيوتر.
TV *(English)*	شاف تلفزيون .
coiffeur *(French)*	ذهبت عند الكوافير.

7.4.2 Dialectal Arabic

Annotators will frequently encounter non-MSA dialect especially in the broadcast conversation programs. Non-MSA dialects include the following:

- Gulf Arabic: Saudi, UAE, Qatar, Oman, Bahrain, Kuwaiti, Iraqi
- Levantine: Syrian, Jordanian, Lebanese, Palestinian
- Maghrebi: Moroccan, Tunisian, Algerian
- Nile: Egyptian, Sudanese

It can be very difficult to distinguish when someone is speaking MSA and when they are speaking in a colloquial dialect, and speakers may move back and forth rapidly within a single statement. Nevertheless, because the target language for this transcription task is MSA, it is helpful to indicate when a speaker is obviously speaking in a colloquial dialect. Therefore, annotators should do their best to identify portions of speech when someone is obviously speaking in an Arabic dialect rather than MSA.

[8] 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.

Regions of non-MSA speech should be identified using a special marker:

```
<non-MSA> text </non-MSA>
```

The words should be transcribed using standard Arabic orthographic conventions. If the conversation switches back and forth between MSA and non-MSA dialect, mark just the non-MSA portions using the convention described above, and leave the MSA portions unmarked. Note also that SU segmentation is unaffected by the presence of non-MSA speech. A single SU segment may contain all MSA, all non-MSA, or a mix of both.

The examples show several SUs in Iraqi dialect:

<non-MSA> يعني أكو بها مبالغة يعني بالاحداث ياللي بتصير بالداخل. </non-MSA>

<non-MSA>يعني بدليل انه الوزير بنفسه ديصرح ديقول ل- أكو يعني أجانب جايين وطلع الجوازات يعني قدامك عالشاشة. </non-MSA>

<non-MSA> يـ عـ ني أكو قتلى موجودين يعني حتى جثث موجودة (()) يعني هل من المعقول أكو مساجد يكون داخلها جثث.</non-MSA>

This example shows a segment with a mixture of MSA and Non-MSA:

انا لست متفقا مع الدكتور <non-MSA>لاني شايف انو</non-MSA> على خطأ

Annotators may also encounter **MSA** spoken with a regional accent. This should be transcribed using standard Arabic orthography, without any special markup. Accented words should **not** be labeled as mispronounced words. Annotators should not transcribe any accent features, for example, *g* for *j* or *g* for *q*, but must use standard orthography. For example:

Pronounced as...	**Regional accent**	**Transcribed as...**
dagAg	Egyptian (Nile)	دجاج
ygul	Iraqi (Gulf)	يقول

It is most important in transcription that annotators only transcribe what they hear, instead of what they think is correct. Annotators should not attempt to normalize dialectal features. For example,

Speaker says الي even in a MSA context; transcriber should **not** turn it into الذي.
Speaker says انو ; transcriber should **not** turn it into أنه

Another thing transcribers keep in mind is not to allow their own regional background to influence their transcription. Transcribers write what they hear, not what they expect to hear.

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 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:

يعني هل من المعقول أكو مساجد يكون (()) يعني أكو قتلى موجودين يعني حتى جثث موجودة يـ ـعـ ني <non-MSA> داخلها جثث.</non-MSA>

If an annotator is truly mystified and can't at all make out what the speaker is saying, he or she uses empty double parentheses to surround the untranscribed region. For example:

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

Do not skip the region.

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	خمسة	Written out in full; use foreign language tags as needed
	Punctuation	Comma, question mark, period, double dash	, ? . --	Limited to these four symbols.
	Individual letters	No markup	ای بی ام	Written in Arabic as they are pronounced, with a space between letters
Disfluent speech	**Filled pauses**	No markup	أه, إيه, أم, أوو, مم	Limited to these 5 words; use standardized spellings
	Partial words	-	مس- مستمر	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	+	+Probably (sounds like Podably)	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	n/a	***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. **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". DO NOT leave the "Language" definition blank. **NOTE**: Do not use this convention for foreign borrowings that are common in the target language, e.g. *apropos*, *computer*.

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[9]	
12. save your work frequently	`Alt+F Alt+S`
13. repeat steps 4-12	
14. save and exit	

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