Guidelines for Translating Arabic Text to English

Version 1.0

June 2, 2006

Linguistic Data Consortium

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

1 Introduction

Our goal is to create English translation of Arabic newswire, weblogs, newsgroup text, as well as transcripts of Arabic broadcast news, broadcast conversations, and telephone speeches to support Machine Translation research.

This document describes the format of the source text and its translation, and addresses specific issues when translating text from different genres.

2 The Translation Team

A translation team must consist of at least two members:

1) An Arabic dominant bilingual
2) An English dominant bilingual

One of them does the initial translation, the other one proofreads the translation. It's up to the translation agencies to decide who does the initial translation and who does the proofreading.

The team may use the following means as assistance:

1) An automatic machine translation system
2) A translation memory system.

The translation team must not change during translation, and the team must be fully documented. Documentation includes:

1) The name (or pseudonym), native language, second languages, age and years of translation experience of the translator(s)
2) The order of processing (i.e. the name of the person who performs the first pass, second pass, etc.)
3) The name and version number of any translation system or translation memory used
4) A description of any additional quality control procedures or other relevant parameters or factors that affect the translation

A translation service may have multiple teams working simultaneously. Proofreaders can be shared among teams, unless informed by the LDC not to. Once a team is setup, it should not be changed during the course of translation.

If multiple teams are used to complete the work, the following documentation should also be sent to the LDC along with the translations at the end of the work:

1) The team: names (or pseudonyms) of the translator and proofreader
2) The files, or segments of files the team translated

3 Arabic Source Text

The original text the LDC creates or acquires are in many different formats, which, besides speaker ID and transcripts, also include metadata such as section boundaries, turn boundaries and timestamps. The LDC reformats the source text before sending them to the translators to 1) make the source files easy to read; 2) to avoid translator's tampering of metadata; 3) to aid automatic processing after the translation is returned to LDC.

Each source file is formatted as such:

```
<ar=1> [speaker1] {Arabic sentence 1}
<en=1>
<ar=2> [speaker1] {Arabic sentence 2}
<en=2>
<ar=3> [speaker2] {Arabic sentence 3}
<en=3>
```

A source file contains multiple Arabic lines, each followed by an English line as the placeholder for the English translation of the Arabic sentence.

Each Arabic line consists of 3 parts:

1. "<ar=##>", where "##" is a unique identification number of the Arabic sentence;
2. "[speaker id]", which contains the identification of the speaker of the Arabic sentence; Speaker IDs apply only to transcripts of speech data, such as broadcast news and talk shows (broadcast conversation), other types of data (newswire, weblogs, newsgroup) do not have speaker IDs.
3. Arabic transcripts

English lines start with "<en=##>", where "##" is the id of the sentence to be translated.

4 English Translation File Format

The translated text is to be organized in exactly the same way as the source text. Translators should type the English translation after each "<en=##>" tag without altering any other part of the file.

Speaker IDs ([speaker1]) are provided to facilitate clear understanding of conversational speech. They do **NOT** need to be translated or copied over.

In cases where a single Arabic sentence is translated into multiple English sentences, **NO** blank lines should be inserted between the English sentences.

The English translation of each source text is to be rendered as plain ASCII text, as illustrated as following:

```
<ar=1> [speaker1] {Arabic sentence 1}
<en=1> {translation of Arabic sentence 1}
<ar=2> [speaker1] {Arabic sentence 2}
<en=2> {translation of Arabic sentence 2}
<ar=3> [speaker2] {Arabic sentence 3}
<en=3> {translation of Arabic sentence 3}
```

Electronic transmission of output translations (as zipped email attachments or ftp) must be used. Paper transmission is not acceptable. All the files should be in plain text file, we do not accept Microsoft Word documents.

5 Translation Quality

The goal of these translations is to take the Arabic source text - which was originally spoken, not written - and translate it, producing a result that sounds as if it was originally spoken in the target language.

Translation agencies will use their best practice to produce translations. While we trust that each translation agency has its own mechanism of quality control, we have specific guidelines so that all translations share a common ground. These are:

1) The English translation must be faithful to the original Arabic text in terms of meaning and style. If the Arabic source text is a news story, the translation should also be journalistic. If the Arabic source text is transcript of a talk show, the translation should be conversational. The translation should mirror the original meaning as much as possible without sacrificing grammaticality, fluency, and naturalness.
2) Try to maintain the same speaking style (or register) as the source. For example, if the source is polite, the translation should maintain the same level of politeness. If the source is rude or angry, the translation should be rude or angry.
3) If the source text is an unedited transcription of spoken conversations, it may sometimes be hard to read, and may make more sense if you read it aloud. You will see that the source text sometimes reflects the kinds of "mistakes" people say when they're speaking aloud. For example, "Uh no I'm um I think he's uh um his home is over there." In this case, the speaker pauses ("uh", "um") and restarts the sentence three times, changing what he's planning to say ("I'm, I think he's, his home is over there"). Your translations will also have this "spoken-sounding" flavor, somewhat different from what you produce when you translate prose.
4) The translation should be as factual as possible. For example, if the original text uses "Bush" to refer to the US President, the translation should **not** be rendered as "President Bush", "George W. Bush," etc. No bracketed words, phrases or other annotation should be added to the translation as an explanation or aid to understanding.
5) The translation should also respect the cultural matrix of the original. For example, if the Arabic text uses the phrase "Comrade Jiang Zemin", the translation should **not** be rendered as "Mr. Jiang Zemin".

6 Proper Names

Proper names should be translated using common practice. This is summarized as follows:

1) Whenever an Arabic proper name has an existing conventional translation into English, that translation should be used. For example, "Gamal Abdel Nasir" the late former president of Egypt, should be translated as "Gamal Abdel Nasir", not "Jamal Abdel Nasir" as Modern Arabic would have suggested.
2) The order "first-name" always first in the source should be preserved. For example, "Osama Bin-Laden" should never become "Bin-Laden Osama", with the "last name" moving to the left of the "first name".
3) Speaker IDs in between brackets, such as "[host]" and "[Anwar_Majed_Ishqi]" in the following example, are provided to the translators to understand the conversation, and they should NOT be translated or copied over. Sometimes the spelling of a speaker ID could be wrong, in which case the translators are expected to correct them in the English translation:

<ar=34> [host] .دكتور أنور ماجد عِشقي مرحباً بك

<en=34> Welcome Dr. Anwar Majid Ishqi.

<ar=35> [Anwar_Majed_Ishqi].أهلاً بك

<en=35> Welcome to you.

4) Non-Arabic proper names should be translated as they would be translated into English directly from the original language. In case of an original English name appearing in the Arabic text, the normal English form should be used.
5) Lacking preexisting knowledge of how to translate a foreign proper name, the translator should use existing resources (such as information gleaned from the www) to decide on a best translation. Failing this, simply proceed as if the name was an Arabic name.
6) Names must be translated consistently across all of the documents.

7 Numbers

Translation agencies will use their best practice to follow standard American writing for numbers:

a) بإمكانكم التصويت إه على رقم الهاتف من داخل دولة قطر تسعة صفرين واحد ثلاثة أصفار، من جميع أنحاء العالم صفرين تسعة سبعة أربعة تسعة صفرين واحد تسعة صفرين.

You can vote uh from inside the country of Qatar at phone number 9001000, and from all parts of the world at 009749001900.

b) نحن لم نصنع التاريخ أه، لفترة طويلة يمكن تصل لألف وربعمية سنة.

We have not made history, uh, for a long period of time that goes back 1,400 years.

c) الرجل لا يزال يتمسَّك بشعار الثورة العربية الكبرى، إه ثمانية شباط، تسعة أيلول.

The man is still holding on to the slogan of the Great Arab Revolution, uh the Eighth of February, the Ninth of September.

8 English Sentences

Occasionally, there are English sentences in the source text. This happens often in newsgroups when internet users post messages in English. It also happens in broadcast news or broadcast conversation when a speaker speaks in English.

English sentences in source text should be copied over to the English translation. Grammatical errors, if there is any, should be corrected to make the translation fluent English.

9 Factual Errors in Source Text

Factual errors in the source text should be translated as is, they should **NOT** be corrected.

a) زار موسكو اليوم الرئيس الأمريكي بوتين.

American President Putin visited Moscow today.

b) ستستضيف سول الألعاب الأولمبية في عام ألفين وثمانية.

Seoul will host 2008 Olympics.

10 Translation of Speech Transcripts

This section addresses issues related to translation of transcripts of speech data, such as broadcast news and broadcast conversations (talk shows, call-in shows).

10.1 Disfluent Speech

Speakers may stumble over their words, repeat themselves, utter partial words, restart phrases or sentences, and use a lot of hesitation sounds. Filled pauses, repetitions, restarts, should be translated into English to the extent possible. Partial words don't need to be translated, but they should be marked in the English translation.

10.1.1 Filled Pauses

Filled pauses are hesitation sounds that speakers employ to indicate uncertainty or to maintain control of a conversation while thinking of what to say next. Filled pauses do not add any new information to the conversation (other than to indicate the speaker's hesitation) and they do not alter the meaning of what is uttered.

Arabic filler pauses include أه, أم, إيه, أوو etc. They should be translated to their closest counterpart in English, such as "uh", "um", "eh" and "ooh".

a) الولايات المتحدة أه لها دورٌ هام ورئيسي في مسار النزاع العربي الإسرائيلي.

The United States, uh, has an important and principal role in the Arab-Israeli dispute.

b) أه أم أعتقد إنه، نعم، هناك أمل، وهناك عزم.

Uh, um, I think that, yes, there is hope, and there is determination.

10.1.2 Translation of "أه","أم","إيه"

In conversational speeches," أه","أم", "إيه" can be used in many ways, translators should differentiate the different uses and translate accordingly. " أه","أم","إيه " can mean one of the following in a conversational speech:

1) **filled pauses**, as described in section 10.1.1;
2) **Back-channeling**, this is the practice of listeners giving positive comments to the speaker to encourage further talk or to confirm that the listener is listening. In such cases," أه","أم", " إيه "should be translated to its English counterpart, such as "uh-huh" or "yeah". The following conversation between speaker A and B shows the use of " أه" , " أم", "إيه" as back-channeling:

A: أنا أريد في هذه النقطة قبل أن نذهب إلى النقطة الثانية

B: أه

A: الحكومة الأمريكية لها أهداف إستراتيجية عالمية

B: أم

A: لها يعني ، عشرات السنين

10.1.3 Repetition and Restarts

Repetitions and restarts should be translated into English.

a) هناك مسألتين أساسيتين يحددوا، فعلاً، أه ما سوف يؤول إليه الذي تعرض لهذا ال- لهذا الأمر الفظيع.

There are two main issues, defining actually, uh, what will happen to those who were subjected to this %pw, to this horrible thing.

b) يعنى، ليست ليست إدارة عندما، عندما يطبل يطبل لها بعض الليبراليين المقيمين في أمريكا.

I mean, it is not, it is not an administration when some liberals residing in America beat, beat the drum for it.

c) الليبرالية تأسّست في في أوروبا.

Liberalism was established in, in Europe.

10.1.4 Partial Words

A speaker may stop in the middle of pronouncing a word, which results in a partial word. We use a dash "-" to indicate a partial word in the source text and the point at which word was broken off. Partial words do NOT need to be translated, but their existence should be indicated by "%pw" in the English translation.

a) على مستوى المشاعر، جزء كبير بيصاب بخ- بقلق شديد.

At the emotional level, a large portion is inflicted with %pw severe worry.

b) يبدأ يحصل أه أحلام مزعجة جداً، توص-، تصل إلي حد الكوابيس.

They start to have, uh, very disturbing dreams, %pw reaching the point of nightmares.

c) نتائج ستدل على مستوى عالي من التفاهم الوطني وال- التوافق ال- ال- أه الوطني.

Results that will show a high degree of national mutual understanding and %pw, accord, %pw, %pw, uh, national accord.

d) هكذا أنا ك- كمستمع.

This is how I %pw, as a listener.

e) إذا سمحت، إذا س-

If I may, if %pw

10.2 Mispronounced Words and Typos

Occasionally, there are typos in the source text. The translators should translate the intended meaning.

a) جزء كبير بيصاب بخلق شديد، بالتوتّر الشديد.

Large portion is inflicted with severe worry, with severe tension.

b) هذه المسألة الديموجرافية هل أيضاً سببها إسرائيل وأمريك؟

This demographic question is it caused also by Israel and America?

10.3 Semi-intelligible and Unintelligible Speech

Sometimes an audio file will contain a section of speech that is impossible to understand. In these cases, transcribers were instructed to use empty double parenthesis (( )) to mark totally unintelligible speech. For example:

قالت الشرطة العراقية إن 37 (()) قد قتلوا.

If it is possible to guess the speaker's words, transcribers transcribe what they think they hear and surround the uncertain transcription/text with double parenthesis. For example:

ضرب زلزال قوي بلغت قوته 6,8 على مقياس ((ريختر)) نيروبي الساعة 12,19 بتوقيت غرينتش.

Translators should transfer the double parenthesis to the English translation, with the words (if there is any) in between the parenthesis translated into English.

a) قالت الشرطة العراقية إن 37 (()) قد قتلوا.

Iraqi police said that 37 (()) were killed.

b) ضرب زلزال قوي بلغت قوته 6,8 على مقياس ((ريختر)) نيروبي الساعة 12,19بتوقيت غرينتش.

A strong earthquake, measuring 6.8 on the ((Richter)) scale, hit Nairobi at 12:19 Greenwich Time.

IMPORTANT: translators should NOT introduce NEW "(())" in the English translation. In case the source text is very difficult to understand, they can use "[ ]". The translators should always use available resources, such as the Internet, to find correct or most appropriate translation of terms that translators are not familiar with.

الإنتخابات العربية لا تجري على مرشحٍ واحدٍ يفوز بمائة بالمائة من الأصوات كما فاز القائد الضرورة قبل سقوطه

Arab elections are not held with one candidate who gets a hundred per cent of the vote, just as the [bloody] leader did before his fall.

10.4 Program Names

There is many ways to translate a program name of TV/radio station to English. Translators should use the standard translation by which these programs are known in American English.

The following table provides translation of some of the programs we are currently recording:

من واشنطن	From Washington
حوار مفتوح	Open Dialogue
الاتجاه المعاكس	Opposite Direction
منبر الجزيرة	Al Jazeera Platform 1
بلا حدود	Without Bounds
أكثر من رأي	More Than One Opinion
نهاركم سعيد	Naharkum Saiid

11 Translating Newswire, Weblogs, and Newsgroups

11.1 Headlines and Titles

Capitalization: most news, weblogs and newsgroup contain a headline or title, which is usually the first sentence of a news story or article. Content words of the English translation of headlines/titles should be capitalized, function words – such as "the", "and", "of", "is" – do NOT need to be capitalized. For example:

رئيس بلدية بلجيكي يمنع عرض عمل فني يصور صدام حسين
Belgian Mayor Bans Display of Artwork Depicting Saddam Hussein

Style: translation of headlines should use common practice. This is summarized as follows:

a) State or imply a complete sentence in the present tense.
b) Avoid using passive voice.
c) Omit most "helping" and "to be" verbs: *Road Improvements Planned for Belvidere Avenue Southwest* instead of *Road Improvements are Planned for Belvidere Avenue Southwest*.
d) Cut articles (*a, an, the*): *School District Schedules Open House on Proposed Curriculum Changes* instead of *School District has Scheduled an Open House on the Proposed Curriculum Changes*.
e) Infinitive is preferred to future tense: *City Council to Consider Budget Recommendation* instead of *The City Council will Consider the Budget Recommendation*.

11.2 Emoticons (Emotion Icons)

An Emoticon is an ASCII glyph used to indicate an emotional state in email, news or online posting. Emoticons should be copied over to English translation.

The following is an incomplete list of popular emoticons you may see in weblogs and newgroup text:

:-) Standard Smiley (you are joking; satisfied)
:) Standard Smiley for lazy people
,-) Winking Smiley. You don't mean it, even if you are joking
;-) Winking Smiley. See above
:-> Follows a really sarcastic remark

12 Quality Control at LDC

To assure the quality of the translations, LDC will enforce the following policies:

1) LDC has hired fluent bilinguals in Arabic and English to control the translation quality. Every delivery is subject to the reviewers' review. The translation teams are not paid until the translation is to our

satisfaction.

2) For each delivery, we will randomly select a subset of the documents, and choose either the top or the bottom 5 segments, until the total number of words add up to about 1,200. The selected sample translation will then be graded using the system described below.
3) To ensure consistency from one review to another, the following scoring system has been adopted for grading translations:

Error	Deduction
Syntactic	4 points
Lexical	2 points
Poor English usage	1 point
Significant spelling or punctuation error	½ point (to a maximum of 5 points)

4) For each error found, the corresponding number of points will be deducted. For instance, if the original text says “Bush will address the General Assembly of the United Nations tomorrow”, and “tomorrow” is missing in the translation, 2 points would be deducted.
5) If more than 40 points are deducted from the 1200-word sample, the translation will be considered unacceptable and the whole delivery will be sent back to the translation team for improvement.
6) If a delivery is sent back to the translation team for further proofreading, the improved version must be completed within 5 business days.

13 Guidelines

In case these guidelines prove to be unclear, LDC reserves the right to modify them. Agencies will always use the latest version.