GALE Program: Arabic to English Translation Guidelines

Version 2.7

September 7, 2010

Linguistic Data Consortium

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

NOTE: If for any reason translators are uncomfortable working with any particular document included in their assignment, please contact LDC at translations@ldc.upenn.edu to request a replacement.

1 Introduction 3
2 Translation Teams 3
3 File Formats 4
  3.1 Source file 4
  3.2 Completed translation file 5
  3.3 File Naming Conventions 5
4 Delivery of Completed Translations 6
  4.1 Email Correspondence about Translation 6
5 Specific Rules for Translation 6
  5.1 General Principles 6
  5.2 Proper Names 7
  5.3 Numbers 8
  5.4 Capitalization 9
  5.5 Punctuation 9
  5.6 Idioms 10
  5.7 English or other language content 10
  5.8 Factual Errors in Source Text 11
  5.9 Typographical errors 11
  5.10 Difficult to Translate Source Text 11
  5.11 Headlines and Titles 12
  5.12 Special Issues for Translation of Speech Sources 12
    5.12.1 Disfluent Speech 12
    5.12.2 Speaker Noise 13
    5.12.3 Filled Pauses 13
    5.12.4 Translation of أه, إيه ,أم 13
    5.12.5 Repetition and Restarts 14
    5.12.6 Partial Words 14
    5.12.7 Semi-intelligible and Unintelligible Speech 15
    5.12.8 Program Names 15
    5.12.9 Transcription Mark-ups 16
  5.13 Special Issues for Translation of Weblogs and Newsgroups 16
    5.13.1 Emoticons (Emotion Icons) 16
    5.13.2 URLs 16
    5.13.3 Punctuation 16
6 Quality Control at LDC 16
7 Guidelines 19
Appendix: Transcription and translation mark-ups 20

1 Introduction

These guidelines provide an overview of requirements for translation of Arabic text into English to support the GALE Program evaluation of machine translation technology. Genres covered by these guidelines include the following:

- Broadcast News (BN) transcripts consisting of "talking head"-style news broadcasts from radio and/or television networks.
- Broadcast Conversation (BC) transcripts consisting of talk shows plus roundtable discussions and other interactive-style broadcasts from radio and/or television networks.
- Newswire (NW) consisting of newswire feeds.
- Web Newsgroups (NG) consisting of posts to electronic bulletin boards, Usenet newsgroups, discussion groups and similar forums
- Weblogs (WL) consisting of posts to informal web-based journals of varying topical content

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

2 Translation Teams

Initial translation is performed by translation services employing professional translators. Services must assign files to be translated to a ***team*** consisting of at least two members: an Arabic dominant bilingual and an English dominant bilingual. One team member does the initial translation while 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 makeup may not change during translation of a particular data set.

A translation service may have multiple teams working simultaneously. Proofreaders may be shared among teams, unless LDC provides instructions to the contrary for a particular project.

Translation teams may use an automatic machine translation system and/or a translation memory system to assist them during translation.

Translation teams must be documented as follows:

- Translator and proofreader profiles consisting of name or pseudonym, native language, second languages, age and years of translation experience. When multiple translation teams are used, also indicate team membership for each person.

- Work assignment information consisting of the team number or the name of the translator and proofreader for each file in the data set.
- The name and version number of any translation system or translation memory used.
- A description of any additional quality control procedures or other relevant parameters or factors that affect the translation.

This documentation should be submitted to LDC along with the completed translations.

3 File Formats

The source text for translation comes in many different data formats, and may include such metadata as speaker labels, timestamps, section and turn boundaries or other information. LDC converts all source text into a standard translation format before sending data out to translators in order to 1) make the source files easy for translators to read; 2) to avoid translator's tampering with the metadata; and 3) to aid automatic processing of the data after the translation is returned to LDC.

3.1 Source file

Each source file delivered from LDC is divided into segments that roughly correspond to sentences or sentence-like units. Each Arabic segment in the source file consists of 3 components.

- **<ar=##>** is a unique identifier for the Arabic segment;
- **[Speaker id]** is a speaker ID label that is added for segments taken from speech sources like broadcast news and talk shows. Not all segments and files have corresponding speaker IDs. Data drawn from non-speech genres (like newswire and web text) will not have speaker IDs.
- **Arabic text to be translated**.

Each corresponding English segment in the source file consists of one component:

- **<en=##>** is a unique identifier indicating where the English translation should be added for a given Arabic segment.

The ## for a given English segment is identical to the corresponding Arabic segment to be translated.

A sample source file appears below:

```
<ar=9> [speaker1] دكتور عمرو أه موسى يعني كيف تعلقون على مثل هذا الحادث ب- أه لبنان؟
<en=9>
<ar=10> [Amr_Moussa] أه أنا علقت وأصدرت بيان رسمي ينعى السيد رفيق الحريري.
<en=10>
```

3.2 Completed translation file

The completed translation should be formatted exactly the same as the source file; the only difference between the source file and the completed translation file is that an English translation will be added after each English segment identifier.

Translators should type the English translation after each "<en=##>" tag without altering any other part of the file. Altering anything in the Arabic segment, or adding unnecessary line breaks, carriage returns or other stray marks in the files makes it difficult for LDC to automatically post-process the translation files. Similarly, in cases where a single Arabic sentence is translated into multiple English sentences, no blank lines should be inserted between the English sentences.

Speaker IDs in square brackets, e.g. [speaker1], are provided to facilitate clear understanding of conversational speech. They should not be translated or copied over into the translation. Occasionally speaker names do not appear inside square brackets but instead appear at the start of the source text. In such cases, the names should be translated as usual, using the standard guidelines for translating proper names (see Section 5.2).

To summarize: the content inside square brackets should not be translated or copied over into the translation. Other content should be translated

English translations should be rendered in plain ASCII text using UTF-8 encoding.

A sample of a completed translation in the correct format follows:

<ar=10> [Amr_Moussa] أه أنا علقت وأصدرت بيان رسمي ينعى السيد رفيق الحريري.

<en=10> Uh, I have commented and issued an official statement about the death of Mr. Rafik Hariri.

<ar=11> [Amr_Moussa] أنا شخصيا في غاية الحزن والأسف على فقدان هذا الرجل العظيم اللي كان يعتبر وجها عربيا وليس فقط لبنانيا.

<en=11> I am personally very sad and sorry about the loss of this great man, who was regarded as a prominent Arab personality and not only a Lebanese one.

3.3 File Naming Conventions

It is very important that completed translation files use the exact same name as the initial file provided by LDC. Please do not add anything to the file name (like ENG or the agency name) or change anything in the file name, including the file extension.

4 Delivery of Completed Translations

Completed translation files should be submitted electronically, as zipped email attachments, via FTP or by web upload as specified by LDC. Paper transmission is not acceptable. All completed translation files must be in plain text, not in some proprietary software format (like Microsoft Word).

Translations must be delivered on time according to the schedule negotiated with LDC at the start of the project. Late submissions may result in payment penalties. If it looks like a translation delivery may need to be later than the agreed-upon submission date, please contact LDC as soon as possible to discuss the situation. In some cases, a partial delivery may be required before the final delivery date.

4.1 Email Correspondence about Translation

When sending correspondence about a translation project, please include the following information in the subject header:

- Agency Name
- Translation Package Name and Part Number (where applicable)
- A descriptive phrase like "Translation Delivery", "Invoice", "Question" etc.

For instance, the following subject headers are helpful:
ABCAgency GALETrain01.Arabic Part 1 - Translation Delivery
AgencyX GALEPhase2Eval.Arabic Part 2 – Invoice
FGH MT07.Arabic – Schedule Question

Using a descriptive header helps LDC direct your message or delivery to the right place, ensuring timely response.

5 Specific Rules for Translation

5.1 General Principles

The goal of the GALE Translation process is to take Arabic source text drawn from many different genres, both spoken and written, and translates it into fluent English while preserving all of the meaning present in the original Arabic text. Translation agencies will use their own best practices to produce high quality translations. While we trust that each agency has its own mechanism of quality control, we provide the following specific guidelines so that all translations are guided by some common principles.

- The English translation must be faithful to the original Arabic text in terms of both 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 while preserving grammaticality, fluency, and naturalness.
- 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, excited or angry, the translation should convey the same tone.
- In the case of speech sources like broadcast news and talk shows, the source text is an unedited transcription of spoken conversations. In some cases this means the transcript is 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 make when they're speaking aloud, like hesitation sounds (um, uh), restarted sentences and partial words. Your translations should retain the flavor of this "spontaneous speech" style, which will be quite different from what you produce when you translate prose.
- The translation should contain the exact meaning conveyed in the source text, and should neither add nor delete information. For instance, if the original text uses *Bush* to refer to the current 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.
- The translation should also respect the cultural assumptions of the original. For example, if the Arabic text uses the phrase *Comrade Jalal Talabani*, the translation should **not** be rendered as *Mr. Jalal Talabani* – instead, it should keep the term used in the original source.
- All files should be spell checked and reviewed for typographical or formatting errors before submission.

5.2 Proper Names

Proper names should be translated using conventional practices.

Whenever an Arabic proper name has an existing conventional translation in English, that translation should be used. For example, *Gamal Abdel Nasser* (جمال عبد الناصر), the late former president of Egypt, should be translated as *Gamal Abdel Nasser*, not *Jamal Abdel Nasser* as Modern Arabic would have suggested.

The order of last name, first name presentation for the name in the source file should be preserved. For instance, if the source file reads *Osama Bin Laden*, this should **not** be changed to *Bin Laden Osama* in the translation.

Recall that Speaker IDs that appear inside square brackets are provided as an aid to comprehension and should **not** be translated or copied over into the translation.

In some cases there may be mistakes in the name used within the speaker ID. If this happens, translators must be careful to use the proper name within the translation file. They should not repeat the mistake shown in the transcript. For instance:

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

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

<ar=35> [Anwar_Majed_Ishqi] أهلاً بك.
```

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.

Lacking preexisting knowledge of how to translate a proper name, the translator should consult a standard resource or do a quick web search. This is especially important for names that re-occur, for instance names that are part of a news story or names of political leaders. For names of "regular people" that occur only once *and* are highly unlikely to be found on the web (for instance, the names of the author of a newsgroup post), translators should use their best judgment about how to translate the name, following standard conventions. Lacking information to the contrary, translators should simply proceed as if the name was an Arabic name.

For specific proper names such as names of agencies, programs, conferences, books, films, and other media, translators should follow the generally accepted or most commonly used form. If no common form exists or if there are multiple forms in frequent use, translators should provide the translation that follows linguistic rules instead of a translation that is word-for-word but incorrect or awkward.

Names should be translated consistently within and across files.

5.3 Numbers

As a general rule of thumb, numbers in the translation should appear either spelled out in full, or written as ASCII numbers, according to how they appear in the source text. However, there are some general conventions for writing numbers in American English that should be followed.

In American English, commas are generally used for numbers with more than three digits unless they are the names of years.

> 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.`

> 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.`

Use a combination of numerals and words for very large numbers.

> 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.`

> سعر هذا البيت مليون وثلاثمائة وخمسة وأربعون ألف دولار.
> `This house costs 1,345,000 dollars.`

5.4 Capitalization

Translators should follow standard written English rules for capitalization unless there is strong evidence in the source text that suggests a different treatment. Proper names should be capitalized, including personal names and names of organizations and geo-political entities. The first word of each sentence should also be capitalized.

5.5 Punctuation

Written standards for punctuation vary across languages. As a general rule of thumb, punctuation in the translation should match the flavor of the punctuation in the source data, while following Standard English punctuation conventions. Punctuation in the source text primarily serves to enhance readability, so translators should not spend too much time worrying about the exact placement of commas and internal punctuation in the English translation.

Different genres will vary widely in their use of punctuation, and the translation of each genre should respect the flavor of the source text when it comes to punctuation. Newswire files typically use standard written punctuation conventions, and these should be preserved in the translation. Broadcast transcripts typically have reasonably standard punctuation, which should be preserved in the translation. When broadcast transcripts are missing punctuation, it is due to a transcription error, so translators should **add** punctuation in the translation following Standard English punctuation conventions. Web text (newsgroups and weblogs) is the most challenging genre in terms of punctuation. For web genres, translators should **not** standardize punctuation in the translation when it is missing from the source text. Non-standard punctuation (dashes, ellipses, missing end-of-sentence punctuation and the like) should also be preserved in the translation for web text.

Often in transcripts of conversational Arabic, speakers tend to change the subject and restart a sentence in the middle of an unfinished one. When this occurs in the source text, the translation should mimic the source text's punctuation in interpreting the type of pause, restart, or change of subject occurring in speech (see Section 5.12).

5.6 Idioms

Idioms, colloquial expressions and the like are particularly difficult to translate. If a similar expression exists in English, you should use it. When there is no direct translation in English, you should preserve the meaning of the Arabic expression but render it in fluent English rather than providing a literal word-for-word translation.

إش جاب لجاب

There's no comparison.

Literal but incorrect Translation: *what did he brings to bring*

تعبك راحة

No problem. *(*A courteous way to say)

Literal but incorrect Translation: *To get tired for you is a comfort*

5.7 English or other language content

Occasionally English or another language may appear 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 exactly as they appear in the source text. Do not make any changes or corrections to the English, even if the English contains grammatical or other errors.

He have a good handwriting.

5.8 Factual Errors in Source Text

Factual errors in the source text should be translated as is. They should **not** be corrected. This also applies to grammatical errors or other speaker "mistakes" in the source text.

زار موسكو اليوم الرئيس الأمريكي بوتين.
American President Putin visited Moscow today.

ستستضيف سول الألعاب الأولمبية في عام ألفين وثمانية.
Seoul will host 2008 Olympics.

5.9 Typographical errors

Translators will occasionally notice obvious typographical errors or obvious incorrect use of homophones in the source text. In such cases, translators should translate the intended meaning but should add the flag = before the translated word to indicate that it is a correction of a typo. For example,

المعنة التى تمنحها الصين الى افريقيا
the =aid that China provides to Africa

Be careful to distinguish obvious typographical errors, which should be corrected, from factual errors in the source text, which should **not** be corrected. If it is not clear whether the item is a typographical error or a factual mistake, translators should **not** correct the item.

5.10 Difficult to Translate Source Text

In rare cases the source text may be so difficult to understand that translation is very difficult. In such cases, translators should make their best guess about the appropriate translation, but should surround the translated text with [ ] to indicate that this is a guess based on confusing source text, for instance:

الإننتخابات 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.

شقدفونا وشلوا البقرة من البيت، وما لقوه أخذوه.
[They expelled us] and [took] the cow from home. They took everything they found.

As always translators should use available resources including the internet to find the most appropriate translation for unfamiliar terms or phrases.

5.11 Headlines and Titles

Newswire sources frequently contain a headline or title as the first sentence of the document. Headlines are typically written in English a little differently than other kinds of news text, and so headline translations should observe these special guidelines.

Content words in the English translation of headlines/titles should be capitalized. Function words like *the, and, of, is* should not be capitalized. For example:

رئيس بلدية بلجيكي يمنع عرض عمل فني يصور صدام حسين

Belgian Mayor Bans Display of Artwork Depicting Saddam Hussein

When formatting titles, translators should follow style of the original source text: if two titles or headings exist, the translation should reflect this; if no titles exist in the source, the translation should not have titles either.

English headlines follow some special stylistic guidelines, summarized as follows:

- State or imply a complete sentence in the present tense.
- Avoid using passive voice.
- Limit use of punctuation within headlines. End-of-sentence punctuation is not required.
- Omit most "helping" and "to be" verbs

  Road Improvements Planned for Belvedere Avenue Southwest

  instead of

  Road Improvements are Planned for Belvedere Avenue Southwest
- 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
- Use infinitive instead of future tense

  City Council to Consider Budget Recommendation

  instead of

  The City Council will Consider the Budget Recommendation

5.12 Special Issues for Translation of Speech Sources

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

5.12.1 Disfluent Speech

Speakers may stumble over their words, repeat themselves, utter partial words, restart phrases or sentences, and use a lot of hesitation sounds. The sections below describe how to deal with these phenomena in translation.

5.12.2 Speaker Noise

Transcripts may sometimes include markup for speaker-produced noise like coughing, sneezing and laughter. These markers should be copied over into the translation using their original formatting, e.g. {cough}, {laugh}. Please refer to the Appendix for detailed description and examples.

5.12.3 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, but they do provide structural information and are an important part of spoken language so they should be translated.

Arabic filled pauses include أه, أم, إيه, أوو etc. They should be translated to their closest counterpart in English, such as *uh*, *um*, *eh* or *ah*.

> 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.`
>
> أه أم أعتقد إنه، نعم، هناك أمل، وهناك عزم.
>
> `Uh, um, I think that, yes, there is hope, and there is determination.`

It may sometimes be difficult to decide where to place filled pause in the translated text, since languages vary in where filled pauses occur in speech. The translator is free to shift the location of the filled pause in the translation to make the English more natural, but the filled pause should **not** be deleted or ignored.

5.12.4 Translation of أه, إيه ,أم

In conversational speech, أه, أم, إيه can be used in many ways. Translators should differentiate the multiple uses of these words and translate accordingly.

أه, أم,إيه can function as a filled pause. See section 6.12.3 for discussion of how to translate filled pauses.

أه, أم,إيه can also function as a backchannel, in order to provide positive feedback to the speaker to encourage further talk or to confirm that the listener is listening, as in the following example between speaker A and speaker B:

> 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

In backchannel cases, the appropriate English counterpart is something *uh-huh* or *yes*.

5.12.5 Repetition and Restarts

When a speaker repeats him/herself or restarts a sentence halfway through, the repeated words should be translated into English:

هناك مسألتين أساسيتين يحددوا، فعلاً، أه ما سوف يؤول إليه الّذي 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.

. يعنى، ليست ليست إدارة عندما، عندما يطبل عندما 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.

الليبرالية تأسّست في في أوروبا.
Liberalism was established in in Europe.

5.12.6 Partial Words

A speaker may stop in the middle of pronouncing a word, which results in a partial word. Transcribers will sometimes use a dash “-“ to indicate a partial word in the source text and the point at which word was broken off. Translators should **not** attempt to translate the partial word; their existence should be indicated by using the symbol *%pw* in the English translation. For instance:

على مستوى المشاعر، جزء كبير بيصاب بق- بقلق شديد.
At the emotional level, a large portion is inflicted with %pw severe worry.

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.

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.

هكذا أنا ك- كمستمع.
This is how I %pw, as a listener.

إذا سمحت، إذا س-

If I may, if %pw

5.12.7 Semi-intelligible and Unintelligible Speech

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

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

If it is possible for the transcriber to guess the speaker's words, they 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 12,19بتوقيت غرينتش.

Translators should transfer the double parenthesis over to the English translation, along with the translated words (if there are words to translate). For instance:

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

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

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.

5.12.8 Program Names

There are many ways to translate names of Arabic TV and radio programs into English. Translators should use the translation that is in common use for these programs. The following table provides common translations for some of the programs in the LDC source data:

من واشنطن	From Washington
حوار مفتوح	Open Dialogue
الاتجاه المعاكس	Opposite Direction
بلا حدود	Without Bounds
أكثر من رأي	More Than One Opinion
نهاركم سعيد	Naharkum Saiid
العراق اليوم	Iraq Today
من العراق	From Iraq

أحداث الساعة	Affairs of the Hour
صناع القرار	Decision Makers
العالم اليوم	The World Today

5.12.9 Transcription Mark-ups

Some mark-ups are used in transcripts to mark special speech phenomenon, such as mispronounced words, dialect, foreign language etc.. Some signs need to be copied over into the translation. Please refer to the appendix on how to handle the transcription mark-ups.

5.13 Special Issues for Translation of Weblogs and Newsgroups

Weblogs and Newsgroups frequently contain emoticons and URLs. Please follow the instructions below when translators come across these cases.

5.13.1 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 newsgroup 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

5.13.2 URLs

URLs that appear in the source text should be copied directly into the translation.

5.13.3 Punctuation

Translators should **not** standardize punctuation in the translation when it is missing from the source text. Non-standard punctuation (dashes, ellipses, missing end-of-sentence punctuation and the like) should also be preserved in the translation for web text.

6 Quality Control at LDC

The quality of a translation is not determined by its style of prose or elegance in use of English, but mostly in its accuracy. Our definition of “quality” first and foremost requires the translation delivery to be faithful to the source, preserving its original meaning and style. This should be accomplished with the consideration that the translation should also be comprehensible and in fluent English.

Each translation delivery received by LDC is reviewed for completeness, accuracy and overall quality. Payment for completed translation is contingent upon successful completion of the quality review.

Fluent bilinguals working at LDC select a portion of each delivery and grade it according to several criteria. The amount of data selected for review varies depending on the delivery size, but at minimum constitutes 1,200 words drawn from multiple documents.

The grading system used by all translation reviewers is outlined below:

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)
Overlooked file or section	40 point

Below are the error categories:

- Syntactic: syntactic error
- Missed translation: lexical error, when certain word is missing from translation
- Added translation: lexical error, when certain word is inserted in translation
- Wrong translation: lexical error, when translation is wrong
- Poor English Usage: lexical error, when English is not awkward
- Punctuation: punctuation error, when punctuation is changed or missing in translation
- Spelling error
- Format problems: spelling/punctuation error, when translation mark-up is missing or incorrect
- Overlooked file or section: translation missing from a big portion or a file is missing its translation

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.

If more than 40 points are deducted from a 1200-word sample, the translation will be considered unacceptable and the whole delivery will be sent back to the translation team for improvement.

If a delivery is sent back to the translation team for further proofreading, the improved version must be completed within 5 business days.

Upon completion of the QC review, the LDC translation team will receive a summary report that includes the following components:

Part 1: Data Profile
Information about the data under review (volume, genre, etc.) and an overall rating for this delivery as *excellent, very good, good, fair, poor.*

Excellent	- 2 or less points deducted
v. good	- 3-10 points
good	- 11-20 points
fair	- 21-40 points
poor	- more than 40 points

Part 2: QC Report Summary
Number of words checked:
Error tally: __ points deducted overall

- Syntactic: __ points deducted
- Missed translation: __ points deducted
- Added translation: __ points deducted
- Wrong translation: __ points deducted
- Poor English Usage: __points deducted
- Punctuation: __ points deducted
- Spelling error: __ points deducted
- Format problems: __ points deducted
- Overlooked file or section: __ points deducted
- Other: __ points deducted

Part 3: QC Report Details
For each significant deduction above, at least one example is provided, along with the following information:

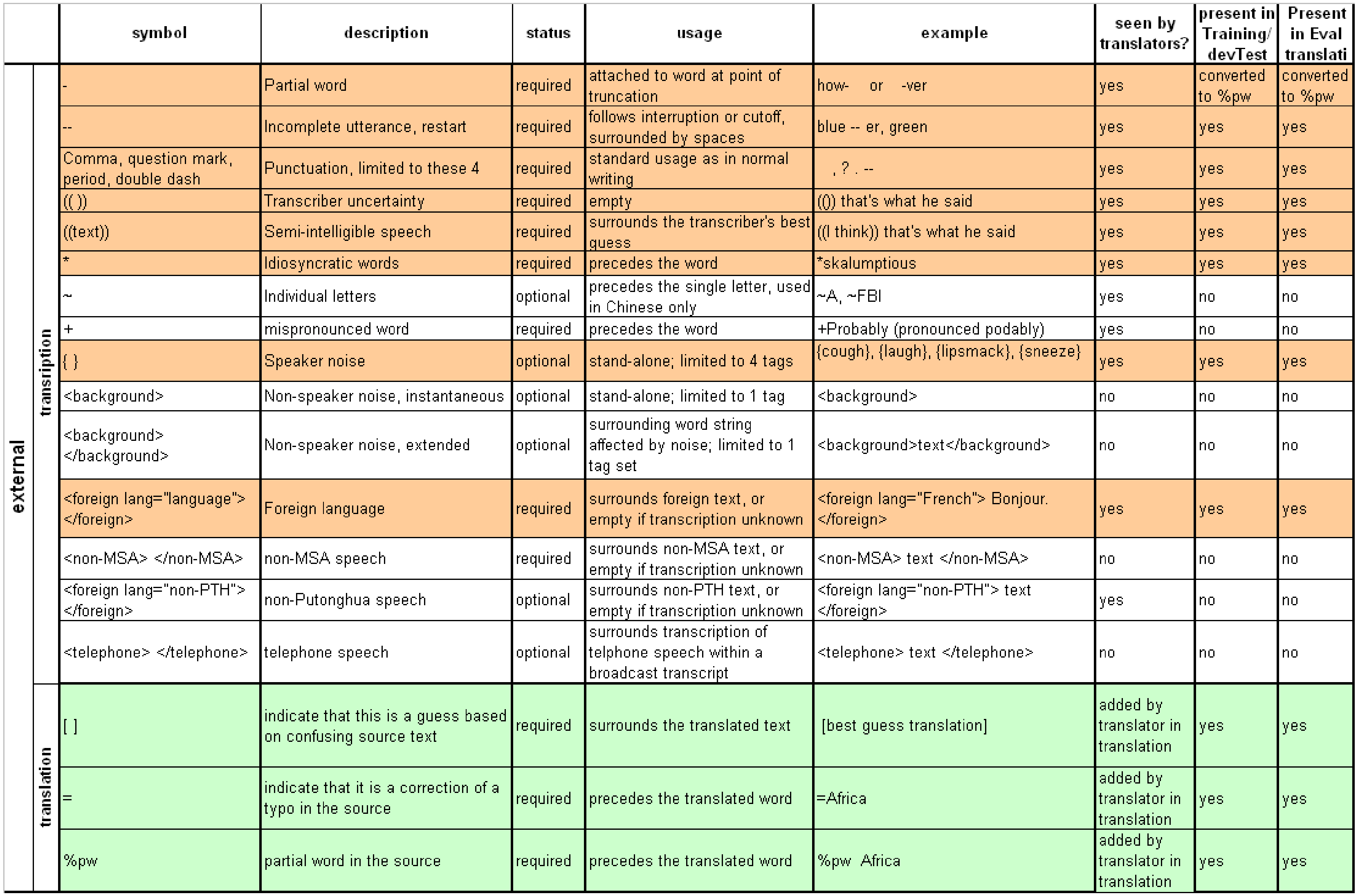
- FileID: e.g. google.com._edfdkfjd_1223.txt
- Your translation
- Suggested translation

- Discussion: a description of what should be changed and why

7 Guidelines

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

Appendix: Transcription and translation mark-ups



		symbol	description	status	usage	example	seen by translators?	present in Training/ devTest	Present in Eval translati
external	transription	-	Partial word	required	attached to word at point of truncation	how- or -ver	yes	converted to %pw	converted to %pw
		--	Incomplete utterance, restart	required	follows interruption or cutoff, surrounded by spaces	blue -- er, green	yes	yes	yes
		Comma, question mark, period, double dash	Punctuation, limited to these 4	required	standard usage as in normal writing	, ? . --	yes	yes	yes
		(( ))	Transcriber uncertainty	required	empty	(()) that's what he said	yes	yes	yes
		((text))	Semi-intelligible speech	required	surrounds the transcriber's best guess	((I think)) that's what he said	yes	yes	yes
		*	Idiosyncratic words	required	precedes the word	*skalumptious	yes	yes	yes
		~	Individual letters	optional	precedes the single letter, used in Chinese only	~A, ~FBI	yes	no	no
		+	mispronounced word	required	precedes the word	+Probably (pronounced podably)	yes	no	no
		{ }	Speaker noise	optional	stand-alone; limited to 4 tags	{cough}, {laugh}, {lipsmack}, {sneeze}	yes	yes	yes
		<background>	Non-speaker noise, instantaneous	optional	stand-alone; limited to 1 tag	<background>	no	no	no
		<background> </background>	Non-speaker noise, extended	optional	surrounding word string affected by noise; limited to 1 tag set	<background>text</background>	no	no	no
		<foreign lang="language"> </foreign>	Foreign language	required	surrounds foreign text, or empty if transcription unknown	<foreign lang="French"> Bonjour. </foreign>	yes	yes	yes
		<non-MSA> </non-MSA>	non-MSA speech	required	surrounds non-MSA text, or empty if transcription unknown	<non-MSA> text </non-MSA>	no	no	no
		<foreign lang="non-PTH"> </foreign>	non-Putonghua speech	optional	surrounds non-PTH text, or empty if transcription unknown	<foreign lang="non-PTH"> text </foreign>	yes	no	no
		<telephone> </telephone>	telephone speech	optional	surrounds transcription of telphone speech within a broadcast transcript	<telephone> text </telephone>	no	no	no
	translation	[ ]	indicate that this is a guess based on confusing source text	required	surrounds the translated text	[best guess translation]	added by translator in translation	yes	yes
		=	indicate that it is a correction of a typo in the source	required	precedes the translated word	=Africa	added by translator in translation	yes	yes
		%pw	partial word in the source	required	precedes the translated word	%pw Africa	added by translator in translation	yes	yes