Qatar Computing Research Insitute

Arabic Language Technologies

Translation Guidelines

Rev. 0.1.3 Date: Sun Sep 21 17:09:01 AST 2014.

Added timestamps for corrections. Fixed some typos, and lowercased English examples. Fixed description for foreign tag.

Added numbers to sections

Fix typo on 9.4 لهجة

These are the general guidelines for translating video subtitles and audio transcription. See below for an example excerpt. Each of the segments is marked with a number for the sake of clarity:

Example text (5 segments)

```
1> [HES] so we are here to talk about functional design .
2> now hopefully we've all got a better idea, from than we did leaving the last meeting,
as to what it is we are up to now .
3> so here's an agenda .
4> right, forty minutes for this meeting,
5> so a bit more than the last one
```

1. General instructions for translation

Translating spoken text can be challenging. Unlike prose, spoken text requires more analysis and understanding. Below are a few instructions to facilitate the task.

1. Before translating a segment, read it until you understand it completely. Try to paraphrase (mentally) it if necessary.

   ```
    2> now hopefully we've all got a better idea, from than we did leaving the last
   meeting, as to what it is we are up to now .
   ```

   For example this sentence (although not FLUENT) says that:

> *Hopefully at this moment, we all have more knowledge of what we need to do, in comparison to when we were leaving last meeting.*

In case of confusion, always keep in touch with other translators

2. Divide the segment into independent clauses

```
2.A> now hopefully we've all got a better idea
2.B> from than we did leaving the last meeting
2.C> as to what it is we are up to now
```

3. Translate each clause respecting the context (e.g. respecting co-refences, discourse connectives, etc.). If some part of the original clause is ungrammatical or does not have a translation (e.g. idiom) you should use the tag [LIT/حرفي X] to denote a literal translation.

```
2.A> نـأمل الآن أن تـكون لـديـنـا فـكرة أفـضل
2.B> مـما كـنـا عليـه تـرك الاجتمـاع الأخير [حرفـي مـن]
2.C> على مـا نـحن بـصدده الآن
```

4. Put the clauses back together, respecting the connectives, and puctuation marks, but respect the original order of the clauses as much as possible.

```
2> نـأمل الآن أن تـكون لـديـنـا فـكرة أفـضل [حرفـي مـن] مـما كـنـا عليـه عند تـرك الاجتمـاع الأخير على
مـا نـحن بـصدده الآن
```

5. If a segment is incomplete (a clause spans over two segments), then read both, and translate them as if they were part of the same segment, but in different clauses. After translating it, re-segment it to align with the original content as much as possible.

   For example

```
4> right, forty minutes for this meeting,
5> so a bit more than the last one
```

   We make them a single segment

```
4'> right, forty minutes for this meeting, so a bit more than the last one.
```

   Then repeat steps 2 to 4.

```
4'> حسنا، أربعون دقيقة لهذا الاجتماع، إذا وقت أطول قليلا من الأخير.
```

Then resegment into corresponding segments respecting their original order as much as possible.

```
4> حسنا، أربعون دقيقة لهذا الاجتماع،
5> إذا وقت أطول قليلا من الأخير.
```

6. You might add additional words or connectives to improve readability of the text, but including this within a [ADD/إضافة x] tag

```
2.B> مما كنا عليه ترك الاجتماع الأخير [حرفي من]
```

Becomes

```
2.B> مما كنا عليه [إضافة قبل] ترك الاجتماع الأخير [حرفي من]
```

2. Basic rules

1. The translation must be faithful to the original text in terms of both meaning and style. The translation should mirror the original meaning as much as possible while preserving 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 politeness. If the source is rude, excited or angry, the translation should convey the same tone.
3. 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 former US President, the translation should not be rendered as **President Bush**, **George W. Bush**, etc.
4. Except for the allowed tagset, no bracketed words, phrases or other annotation must be added to the translation as an explanation or aid to understanding.
5. The translation should also respect the cultural assumptions of the original source. 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 source.
6. All translations should be spell checked and reviewed for typographical errors before submission.
7. After translating a segment, read it again to detect any errors of translation, grammar, structure, etc.
8. Be consistent in translation of identical phrases in similar contexts.
9. For Arabic, diacritics should be used only for disambiguation or adding readability.

```
لديّ ماء عذب يُمكنني من مواجهة الحر
هذه السلع تُخرج إلى السوق يوميا
```

3. Proper Names

1. All named entities should be tagged with the corresponding Named Entity [NE/علم] tag. All names should be translated/transliterated consistently across files.

```
[NE:PER Ahmed Salama]
[علم:شخص أحمد سلامة ]
```

2. Whenever an Arabic proper name has an existing conventional 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.
3. The order of last name, first name presentation for the name in the source file should be preserved.
4. For specific proper names such as names of agencies, programs, conferences, films, and other media, translators should follow the generally accepted or commonly used form. The form used should be consistent througout every translation.

```
[NE:ORG UN] programme for education.
برنامج [ علم:منظمة الأمم المتحدة] للتعليم
```

4. Numbers

1. As a general rule of thumb, numbers in the translation should appear according to how they appear in the source text (either spelled out in full (e.g. twenty-three), or digits (e.g. 23).
2. Exception: Always use figures for years (e.g. 2013, the 70's), statistics (e.g. 70%), and for ranges (e.g. 4-100).

5. Capitalization

1. In English, no capitalization should be used except within named entities.

6. Punctuation

1. 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.

7. Idioms and hard to translate source text

1. If a similar expression exists in English, you should use it. When there is no direct translation into English, you should preserve the meaning of the Arabic expression but render it into fluent English rather than providing a literal word-for-word translation.

```
إش جاب لجاب
Correct: there is no comparison
Literal but incorrect: what did he brings to bring
```

2. Idioms are sometimes hard to translate. If a similar expression does not exist in the target language, produce a fluent translation with an aim to preserve the meaning of the expression. If this is not possible, provide a literal translation embedded in the [LIT/حرفي X] Tag . Do not provide literal word-for-word translation outside this tag.

```
  To fit like a glove
  [حرفي  تناسب مثل القفاز]
```

3. In rare cases the source text may be so difficult to understand that translation is very difficult to perform. In such cases, translators should make their best guess about the appropriate translation, but add the tag [GUESS].

```
  شقدفونا وشلوا البقرة من البيت، ومالقوه أخذوه
  [GUESS They expelled us] and [GUESS took] the cow from home, they took everything
they found.
```

8. Errors in Source Text

In case you find errors in the source text, please take the following precautions:

1. **Factual errors**: in the source text should be translated as is. They should not be corrected

```
 الرئيس الأمريكي بوتين زار موسكو اليوم
[NE:MISC American President] [NE:PER Putin] visited [NE:LOC Moscow] today.
```

   Putin is not the President of America, yet, you need to provide the translation of the original text.

2. **Punctuation error**: Do not add a punctuation that does not exist on source language side. If a comma, colon, or other punctuation mark is missing from the original text, you might add it to the translation ONLY with the tag [ADD/إضافة].

```
they bought flowers chocolates and gifts
اشتروا  أزهارا [إضافة ،] شوكلاتة وهدايا
```

3. **Spelling/Typographical error**: If you find a spelling mistake, perform the translation as if the mistake

was not there. The intended meaning should be translated but should add the flag [FIX] before the translated word to indicate that it is a correction of a typo. The mistakes must be reported in the comment section with their corresponding timestamp per error. If several errors of the same type are found in the same timestamp, they should be separated by comma.

```
   (BAD)         [علم:مكان طاربلس] كما كنت أعانى حتى آخر مرة زرت فيها
   (INTENDED)    [علم:مكان طرابلس] كما كنت أعاني حتى آخر مرة زرت فيها
   (TRANSLATION) as i was [FIX suffering] until last time i visited [NE:LOC [FIX
Tripoli]]

   ----In comment section----------
   Spelling Errors:
   [10:05:07]: أعانى، طاربلس
```

4. **Other erros**: Other types of errors such as unbalanced tags, wrong tag, etc. must be noted in the comment section and the corrected version should be translated.

```
  (BAD)         .[علم:آخر جامعة مني سوتا] في أمريكا أنا من
  (INTENDED)    .[علم:مكان أمريكا] في [علم:منظمة جامعة منيسوتا] أنا من
  (TRANSLATION) i am from [NE:ORG University of [FIX Minnesota]] in [NE:LOC
America].

  ----In comment section----------

   Spelling Errors:
  [4:35:15] جامعة، مني سوتا
  Missing tags:
  [4:35:15][علم:مكان أمريكا]
```

9. Speech Translation

1. There are several disfluencies already marked in the text. Do not translate any word or phrases that occur inside of a tag like “[REP he said that]”.
2. Do not copy other speech-related tags like; FALSE, REP, CORR, INTERP, HES, INTERJ, BREATH, APPLAUES, NOISE, MUSIC

Exceptions only copy:

1. The named entity tags in the transcription like

```
[NE:PER Henry Ford] lives in [NE:LOC Qatar]
[   [علم:مكان قطر] يسكن في [علم:شخص هنري فورد]
```

2. The foreign language tag [FOR:]. Translations of foreign tags should be transliterated into the target language script, and use the corresponding translated tag.

```
[FOR:fr La vie en rose ] was the best film
[أجنبي:فرنسي لا في ان روز]  كان آفضل فلم
```

3. When the [FOR] is the target language, the expression should be transliterated into the target language with no tags.

```
سأقدم [أجنبي:إنجليزي بريزنتايشن] عن الخلايا الجذعية.
i'll present a presentation about stem cells.
```

4. When translating from Arabic into English, Dialectal Arabic should be translated using the tag [ORI: X]

```
فقلت له [لهجة:مصري معلش]، سأحاول مرة أخرى.
and i told him [ORI:EGY sorry], i'll try another time.
```

5. the [UNK/مبهم] tag should be passed as is.

```
 1> [UNK] [HES ] so we are here [REP to] to talk about [INTERP func] functional
design .
 1>[مبهم] لذلك نحن هنا للحديث عن التصميم الوظيفي.
```

10. TAGSET

Below is a summary of the tagset that can be used for translation:

1. [ADD/إضافة: X] When adding some word in the target language to improve readability
2. [GUESS/خمن: X] When the meaning is hard to convey, the translator provides the best guess.
3. [LIT/حرفي: X] When adding a literal translation (use sparringly)
4. [NE/علم:TYPE X] When translating a named entity
5. [FIX/عدل: X] When the translation corresponds to a corrected typographical error on the source side.
6. [UNK/مبهم] When unrecognized voice is reporeted in the transcript.
7. [ORI: X] When dialect word is translated

References:

LDC guide line for translating from Arabic to English
https://catalog.ldc.upenn.edu/docs/LDC2012T17/GALE_Arabic_Translation_Guidelines_V2_7.pdf