Guidelines for the Linguistic Data Consortium's Language Translation Evaluation Project for Translation of Arabic to English

Goal

Our goal is to support the development of automatic means of evaluating translation quality. To this end it is necessary that we have a number of different translations of the same source material.

The Translation Team

A single translation "team" **must** be used to translate all of the source language data. This team may be:

1) A single bilingual translator
2) An Arabic dominant bilingual who does initial translation and an English dominant bilingual who proofreads and edits the output of the first translator
3) A MT system that does initial translation and a translator who proofreads and edits the output of the translation system
4) Either 1), 2), or 3) above, assisted by a translation memory system.
5) Some other "team" that we have not anticipated but that we might be willing to entertain

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

In some cases, LDC may allow a translation agency to perform more than one translation. To be considered, the agency must propose completely different teams for each of the translations. In this case, the teams must be **completely independent**, with absolutely no communication or resource sharing between teams.

Arabic Source Text

Each story has SGML tags added at the beginning to aid automatic processing, as follows:

```
<DOC docid=XXXXXXXXXX>
--Segment 1--
{Arabic text to be translated}


--Segment 2--
{Arabic text to be translated}


--Segment 3--
{Arabic text to be translated}
```

Each story is divided at sentence (originally marked with a Arabic period except for the headline) boundaries. A story is organized into records of Arabic text separated by blank lines. Each sentence is preceded and followed by a blank line. Each segment has a number associated with it.

English Translation File Format

The English translation of each source story is to be rendered as plain ASCII text, with enclosing SGML tags that preserve the attributes of the original story, as illustrated as following:

```
<DOC docid=XXXXXXXXXX>
--Segment 1--
{ English translation }


--Segment 2--
{ English translation }


--Segment 3--
{ English translation }
```

The translated data is to be organized in exactly the same way as the source data. In cases where a single Arabic sentence is translated into multiple English sentences, **NO** blank lines should be inserted between the English sentences.

The headlines (the first segment in each story) should be capitalized. For example, instead of "Nets and pistons advanced to second round of NBA eastern conference", please use "Nets and Pistons Advanced to Second Round of NBA Eastern Conference".

Electronic transmission of output translations (as zipped email attachments or ftp) must be used. Paper transmission is not acceptable.

Translation Quality

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. The Arabic source text is usually a news story, thus the translation should also be journalistic. The translation should mirror the original meaning as much as possible without sacrificing grammaticality, fluency, and naturalness.
2) 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.
3) The translation should also respect the cultural matrix of the original. For example, if the Arabic text uses the phrase “Prince Nasir”, the translation should **not** be rendered as “Crown Prince Nasir”.

Translation of 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) Non-Arabic proper names should be translated as they would be translated into English directly from the original language. In the case of an original English name appearing in the Arabic text, the normal English form should be used.
4) 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.
5) Names must be translated consistently across all of the documents.

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 10 points)

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.

4) 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.

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

Guidelines

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

.

.

.