

LCTL Translation Guidelines
Version 0.8
June 09, 2006
Linguistic Data Consortium
<http://projects ldc.upenn.edu/LCTL/>

1 Introduction

This document describes the specifications for translating text in a Less Commonly Taught Language (LCTL) into modern American English and vice versa.

2 The Translation Team

2.1 Members

A translation team must consist of at least two members: an LCTL-dominant bilingual and an English-dominant bilingual.

For LCTL-to-English translation, the LCTL-dominant bilingual performs the initial translation (first pass) and the English-dominant bilingual proofreads and edits the translator's output (second pass).

For English-to-LCTL translation, the roles are reversed.

2.2 Changes to Team

The translation team may not change during the translation of a file. That is, only one team may complete translation on a particular file.

A translation agency may have multiple teams working simultaneously on a set of files. Members may be shared among teams.

2.3 Assistance

The team may use the following as assistance:

- An automatic machine translation system; and
- A translation memory system.

2.4 Documentation

The team must be fully documented. Documentation includes:

- The name, native language, second languages, age, and years of translation experience of the translator(s);
- Tasking: who performed the first pass and who performed the second pass of each file;
- The name and version number of any automatic machine translation system or translation memory system used; and
- Anything else relevant to the quality of the translation.

If multiple teams are used to translate a set of files, the following documentation is also required:

- Which members are on which teams; and
- Which files each team translated.

3 Formatting

3.1 File Types

All files (source and translation) are text files (with a .txt extension) in Unicode (UTF-8) format.

3.2 Source Files

Each source file is formatted as such:

```
--Segment 1--  
Ηερεεσ τηε φηρστ σεγμαεντ ιν τηησ φηλε.  
  
--Segment 2--  
Ηερεεσ τηε σεχονδ σεγμαεντ ιν τηησ φηλε.  
  
--Segment 3--  
Ηερεεσ τηε τηηρδ σεγμαεντ ιν τηησ φηλε.  
  
...
```

In addition, there may be header or footer information in the source files, as in:

```
<DOC docid="FLOP_ENG_20050217.1237.014" lang="ENG">  
  
--Segment 1--  
Ηερεεσ τηε φηρστ σεγμαεντ ιν τηησ φηλε.  
  
--Segment 2--  
Ηερεεσ τηε σεχονδ σεγμαεντ ιν τηησ φηλε.  
  
--Segment 3--  
Ηερεεσ τηε τηηρδ σεγμαεντ ιν τηησ φηλε.  
  
...  
</DOC>
```

3.3 Translation Files

For each source file we send to the translation agency, we send a corresponding translation file for the agency to enter its translations into.

Each translation file is formatted as such:

```
--Segment 1--  
  
--Segment 2--  
  
--Segment 3--  
  
...
```

Translators should type the translation of a source segment after the --Segment x-- line with the same number, as such:

--Segment 1--

Here's the translation of the segment which appears after the "--Segment 1--" line in the source file.

--Segment 2--

Here's the translation of the segment which appears after the "--Segment 2--" line in the source file.

--Segment 3--

Here's the translation of the segment which appears after the "--Segment 3--" line in the source file.

...

In cases where a single source segment must be translated into multiple sentences, no blank lines should be inserted between the translation sentences. Only spaces may be inserted between translation sentences from a single source segment.

3.4 Alterations

Translators may not alter any part of the translation file other than the lines into which they enter the translations. In particular:

- No "--Segment x--" lines may be added or removed;
- Translation files may not be renamed; and
- Header and footer information in the translation files may not be altered in any way.

4 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.

4.1 Meaning and Style

The translation must be faithful to the source text in terms of meaning and style. If the source text is a news story, the translation should also be journalistic. The translation should mirror the original meaning as much as possible without sacrificing grammaticality, fluency, and naturalness.

Maintain the same 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.

4.2 Factual Translation

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.

4.3 Factual Errors

Factual errors in the source text should be translated as is; they should **not** be corrected.

4.4 Respecting the Cultural Matrix

The translation should also respect the cultural matrix of the source. For example, if the source text uses the phrase "Comrade Jiang Zemin", the translation should **not** be rendered as "Mr. Jiang Zemin".

4.5 Commenting

No bracketed words, phrases or other annotation should be added to the translation as an explanation or aid to understanding.

5 Translation Submission

Translations must be submitted electronically, as a zipped attachment to e-mail or via FTP.

6 Quality Control at LDC

6.1 Our Review

LDC has hired fluent bilinguals in the LCTL and English to review the quality of the translations. Every translation is subject to LDC's review. The translation agency will not be paid until the translation is to the LDC's satisfaction.

6.2 Sampling Translated Files for QC

For each translation package, we will score either the first five or the last five segments of a random subset of the files, where the total number of words graded is about 1,200. This sample of translations will then be scored using the system below.

6.3 Scoring of Errors

To ensure consistency from one review to another, the following system has been adopted for scoring translations:

<u>Error</u>	<u>Deduction</u>
syntactic	4 points
lexical	2 points
poor English usage	1 point
significant spelling or punctuation error	1/2 point (with a maximum of 10 points per file)

For each error found, the corresponding number of points will be deducted. For instance, if the source file reads "Bush will address the General Assembly of the United Nations tomorrow" and "tomorrow" is missing in the translation, 2 points will be deducted for the lexical error.

6.4 Unacceptable Translations

If more than 40 points are deducted from the 1,200-word sample, the translation will be considered unacceptable and the whole translation package will be sent back to the agency for improvement.

If a translation package is sent back to the agency, the corrected translation package must be received by the LDC within five business days.

7 Updates to These Guidelines

The LDC reserves the right to modify these guidelines. Agencies should always use the latest version.