Guidelines for Translating Chinese 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 Chinese newswire, weblogs, newsgroup text, as well as transcripts of Chinese 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) A Chinese 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 Chinese 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:

```
<cn=1> [speaker1] {Chinese sentence 1}
<en=1>
<cn=2> [speaker1] {Chinese sentence 2}
<en=2>
<cn=3> [speaker2] {Chinese sentence 3}
<en=3>
```

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

Each Chinese line consists of 3 parts:

1. "<cn=##>", where "##" is a unique identification number of the Chinese sentence;
2. "[speaker id]", which contains the identification of the speaker of the Chinese 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. Chinese 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 Chinese 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:

```
<cn=1> [speaker1] {Chinese sentence 1}
<en=1> {translation of Chinese sentence 1}
<cn=2> [speaker1] {Chinese sentence 2}
<en=2> {translation of Chinese sentence 2}
<cn=3> [speaker2] {Chinese sentence 3}
<en=3> {translation of Chinese 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 Chinese 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 Chinese text in terms of meaning and style. If the Chinese source text is a news story, the translation should also be journalistic. If the Chinese 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) Because 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 Chinese 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 a Chinese proper name has an existing conventional translation into English, that translation should be used. For names without an existing translation, Pinyin should be used in most cases. However, some Taiwanese, Hong Kong and overseas Chinese names do not use Pinyin by tradition. For example, the former Taiwanese president should be translated as "Lee Teng-hui," not "Li Denghui."
2) The order of "last-name first and first-name last" in the source should be preserved. For example, the Chinese president should be "Jiang Zemin", not "Zemin Jiang."
3) Speaker IDs in between brackets, such as "[host]" and "[pu_jing]" 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 is wrong, in which case the translators are expected to correct them in the English translation:

<cn=33> [pu_jing] 你好。

<en=33> Hello.

<cn=34> [host] 欢迎您，普京总统。

<en=34> Welcome, President Putin.

4) Non-Chinese proper names should be translated as they would be translated into English directly from the original language. This is particularly important for translating Japanese, Korean, and Vietnamese names, and also for non-Han Chinese names such as Tibetan, (Inner) Mongolian, and Uigur names. Any names which were originally English should be translation into English using its normal English form. Some names, although they sound like Chinese names, were not Chinese names at all. For example, "彭定康" (former Hong Kong governor) should be translated as "Chris Patten", not "Peng Dingkang".
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 a Chinese 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:

The following guidelines apply to the majority of CU writing except for scientific, statistical, technical, and mathematical writing.

1) Spell out one to nine. Use numerals for 10 and above.

   Among those killed in the bombing, five were children.

   When she turned 21, she realized that she'd rather be a flight instructor.

2) Spelling out large round numbers is preferred.

   She gave the museum more than a hundred thousand artifacts.

3) Use a combination of numerals and words with numbers in the millions and larger.

   The population increased by 2.3 million.

4) Use a comma for numbers with more than three digits unless they are years.

   The book has 1,229 pages.

5) Spell out numbers at the beginning of a sentence or rephrase the sentence to avoid beginning with a number.

   Forty-nine students received the new degree at the May commencement.

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) 美国总统普京今天访问了莫斯科。

   American President Putin visited Moscow today.

b) 汉城将举办 2008 年奥运会。

   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 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.

Chinese filler pauses include 呃, 嗯, 啊, 这个 etc. They should be translated to their closest counterpart in English, such as "er", "um", "ah" and "uh".

a) 我们将继续和，嗯，北朝鲜进行会谈。

   we will continue our talks with, uh, North Korea.

b) 这个问题，啊，很严重。

   This problem is, ah, very serious.

10.1.2 Translation of ”嗯”

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

1) filled pauses, as described in section 5.2.1;
2) back-channeling, which 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, 嗯 should be translated to its English counterpart, such as "uh-huh" or "yes". The following conversation between speaker A and B shows the use of 嗯 as back-channeling:

   A: 我们的期中考试快完了，语文我考得还可以。

   B: 嗯

   A: 我的数学不怎么样。

   B: 嗯

3) answering questions. When 嗯 is used to answer a question, it means “yes”:

   A: 你能不能把那本书给我递过来？

   B: 嗯

10.1.3 Repetition and Restarts

Repetitions and restarts should be translated into English.

a) 我们，我们当然反对台独了。

   We, we certainly oppose the independence of Taiwan.

b) 整个电影的重心是他父母他父母对这个世界的态度。

   `The focus of the whole film is his parents, his parents' attitude to the world.`

10.1.4 Partial Words

A speaker may stop in the middle of pronouncing a word, which results in a partial word. Sometime, 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) 整个电影的 zh- 重心是他父母他父母对这个世界的态度。

   `The %pw focus of the whole film is his parents, his parents' attitude to the world.`

b) 欧盟官员现在最担心的是即将从非，非洲飞回的候鸟。

   EU officials are now worried about the return of migrating birds from %pw Africa.

10.2 Mispronounced Words and Typos

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

a) 抗议活动发生在天蓝门广场。

   The protest happened in Tiananmen Square.

b) 连和国需要 47 亿美元用于人道主义原助。

   The UN needs $4.7 billion for humanitarian aids.

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:

中国十四个边境(())城市一九九五年经济建设取得可喜成果。

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:

我们的目的是让所有((塞尔维亚人))联合起来。

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) 中国十四个边境(())城市一九九五年经济建设取得可喜成果。

   Exciting accomplishment has been achieved in 1995 in the economic development of China's 14 border (()) cities.

b) 我们的目的是让所有((塞尔维亚人))联合起来。

   `Our goal is to unite all the ((Serbs)).`

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:

今日焦点	Today's Focus
鲁豫有约	A Date with Lu Yu
锵锵三人行	Behind the Headlines
新闻骇客	News Hacker
社会能见度	Social Watch
时事开讲	Newsline

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:

伊拉克释放数百名囚犯

Iraq Frees Hundreds of Prisoners

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 Chinese 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.