BOLT Program: Chinese to English CTS Translation Guidelines

Version 1

December 19, 2013

Linguistic Data Consortium

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

These guidelines provide an overview of requirements for translation of Chinese text into English to support the BOLT Program evaluation of machine translation technology. These guidelines cover translation of transcribed telephone speech.

This document describes the format of the source text and its translation, and addresses

specific issues when translating text from speech 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: a Chinese 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 Chinese segment in the source file consists of 3 components.

- **<cn=##>** is a unique identifier for the Chinese segment;

- **[speaker id]** is a speaker identification 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.
- **Chinese 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 Chinese segment.

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

A sample source file appears below:

```
<cn=1>[speaker1] 来看几条日本的消息。
<en=1>
<cn=2>[speaker1] 最近在日本的政坛呢有这样的一个比较有震动性的消息呀。
<en=2>
```

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 Chinese 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 Chinese 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:

<cn=1>[speaker1] 来看几条日本的消息。
<en=1> Let’s take a look at some news items from Japan.
<cn=2>[speaker1] 最近在日本的政坛呢有这样的一个比较有震动性的消息呀。
<en=2> Recently, in Japan’s political arena, there is some rather sensational news.

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 the subject header:

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

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 BOLT Translation process is to take Chinese source text drawn from many different genres, both spoken and written, and translate it into fluent English while preserving all of the meaning present in the original Chinese 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 Chinese 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.
- 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 Chinese text uses the phrase *Comrade Jiang Zemin*, the translation should **not** be rendered as *Mr. Jiang Zemin* – 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 a Chinese proper name has an existing conventional translation into English, that standard translation should be used, for example, 孙中山 is translated into *Sun Yat-sen* instead of *Sun Zhongshan*.

For names without an existing conventional translation, Pinyin should be used in most cases. However, some Taiwanese, Hong Kong and overseas Chinese names do not use Pinyin by tradition, so those names should be rendered in their conventional transliteration. For example, the former Taiwanese president should be translated as *Lee Teng-hui*, not *Li Denghui*.

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 *Jiang Zemin*, this should **not** be changed to *Zemin Jiang* in the translation.

In summary, this is the order of preference in translating Chinese names, mainland and outside the mainland.

1. Common transliteration (English name or most commonly used translation of the name).
2. Regional standard for this name, with regional standard determined by

context.
3. Pinyin

When addressing people in Chinese, it may be awkward to directly translate their titles to English. For example, 张老师 can be left as "Teacher Zhang" even if this does not qualify as "appropriate" in English. This cultural consideration applies to all similar titles.

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

Occasionally the translator may observe that the name appearing within the speaker ID tag is spelled incorrectly. Translators should always use the correct English spelling/transliteration of a name within a file, even when the speaker ID for that name is incorrect.

Proper names from languages other than Chinese should be translated into English following the standard conventions for that 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 Uyghur names. Names that originally come from English should be translated using their most common English form. Some names, although they sound like Chinese names, are not Chinese names at all. For example, 彭定康 (former Hong Kong governor) should be translated as *Chris Patten*, not *Peng Dingkang*.

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 recurring names, 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. Lacking other information to the contrary, translators should simply proceed as if the name was a Chinese 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 and Units

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 years.

> 选举名单上有*40578*个投票人。
> There are *40,578* voters on the electoral rolls.

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

> 大陆有*8亿*农民，不能把台湾模式硬加在大陆上，这是行不通的。
> The mainland has *800 million* peasants, so rigidly applying the Taiwanese model to the mainland won't work.

> 中国现有十三亿人口，发达国家约有十亿人口，让中国全民生活质量达到发达国家水平，其任务超过发达国家总和。
> China currently has a population of *1.3 billion*. The population of the developed countries is approximately *one billion*. Improving the quality of life of China's population to the level of developed countries will be a task greater than the task of all the developed countries combined.

> 近*2万*个投票人参加了这次选举。
> Nearly *twenty thousand* voters took part in this election.

For units of measurement that may differ between English and Chinese (for example "degrees" or "tons"), the translation should reflect the true number using units of measurement familiar and accepted in English. For example, 吨 should be translated to "metric ton", just as 度 should be translated to "degrees celsius".

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. Chinese source text uses standard Chinese punctuations and they need to be converted to their English equivalents. Transcripts typically have reasonably standard punctuation, which

should be preserved in the translation. When transcripts are missing punctuation, it is due to a transcription error, so translators should **add** punctuation in the translation following standard English punctuation conventions.

Often in transcripts of conversational Chinese, 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 5.13).

5.6 Idioms

Idioms, colloquial expressions (成语，习语 refer to http://baike.baidu.com/view/2990.htm on definition of 成语) and the like are particularly difficult to translate for human translators, let alone for machine translation engines. To help machine learning, we will provide both intended meaning and literal meaning of the idiomatic expressions. If a similar expression exists in English, you should use it as the intended meaning. The literal meaning should always be rendered in fluent English, rather than as a word-for-word translation. The translations should be surrounded by [ ] as in [text1 | text2] with | separating the two translations. Text1 should be the intended meaning and text2 should be the literal meaning. For example:

Format: [intended meaning | literal meaning]

NOTE: Intended meaning should always precede literal meaning.

有传闻，二胖*翘辫子*了。
There is a rumor that [Kim Jong Il | Fat guy No.2] [**has kicked the bucket | ‘s pony tail has stuck up**].

Intended meaning: has kicked the bucket
Literal meaning: ‘s pony tail has stuck up

反腐败要想迅速获得效果，要上下齐动员，对腐败分子形成一种*“过街老鼠，人人喊打”*的氛围，在此基础上建立一种清廉的社会风气，让人民觉得生活在充满正义的天地之中
For anti-corruption efforts to achieve rapid results, there must be an even mobilization at all levels, creating an atmosphere in which corrupt individuals are [**immediately called to attention like a rat crossing the street | treated like a rat crossing the street and everybody cries “kill it”]**. On this foundation, an uncorrupted society can be built that will make people feel that they are living in a heaven-on-earth full of justice.

Intended meaning: immediately called to attention like a rat crossing the street
Literal meaning: treated like a rat crossing the street and everybody cries “kill it”

法院、检察院离政府机关稍微远，有*天高皇帝远*的感觉，欲所欲为。
The court and the procuratorate are somewhat far from the government organizations,

giving a feeling of [**distance from central government control | the heaven being high and the emperor far away**] and action in accordance with one's own will.

Intended meaning: distance from central government control
Literal meaning: the heaven being high and the emperor far away

我没吃过猪肉，还没见过猪跑吗?
[**I'm not a complete neophyte, you know? | I haven't eaten pork, but haven't I seen pigs running?**]

Intended meaning: I'm not a complete neophyte, you know?
Literal meaning: I haven't eaten pork, but haven't I seen pigs running?

Not all translations of idiomatic expressions need alternatives. If the intended meaning and literal meaning of the idiomatic expression in a context are identical, there is no need to provide alternatives. For example:

黑油窑在武汉之多之久，且高大耸立，排放黑烟**遮天蔽日**
So many illegal oil kilns have been operating for so long in Wuhan, and they are huge and towering, emitting black smoke **obscuring the sky and the sun**.

武汉黑油窑还要**肆无忌惮**作孽到何时？
How long will illegal oil kilns in Wuhan continue their **unscrupulous** evil operations?

Be careful not to provide translation alternatives for non-idiomatic expressions. Words/phrases like 老百姓 are commonly used regular expressions, so there is no need to provide alternatives even though the etymology of this word comes from "hundred names."

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.

For example, if the source text were the following line:

<cn=1> 亚西尔·阿拉法特 (Yasser Arafat) 1 9 2 9年8月出生于耶路撒冷，是 一位逊尼派穆斯林。

It should be translated as so:

<en=1> Yasser Arafat (Yasser Arafat) was born....

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.

*美国总统普京*今天访问了莫斯科。
American President Putin visited Moscow today.

汉城将举办2008年奥运会。
Seoul will host the 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,

抗议活动发生在天蓝门广场。
The protest happened in =Tiananmen Square.

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 =aid.

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 Special characteristics of Chinese source texts

In Chinese conversation, weblogs and newsgroups, sentences may be grammatically incorrect and elliptical in nature. Translators must be faithful to the original source and stylistically similar, even if this calls for elliptical or ungrammatical English translations.

整治行动呢， 效果呢不错， 但是整治过后容易出现一些问题的回潮， 反弹反复。
Disciplinary actions, eh, results are not bad, but some problems easily resurge once the discipline is finished, rebounds again and again.

其中有一副照片上是一个小孩儿，呃， 身上穿着很多包装物的那样一个泡沫的东西， 看上去非常穷， 肯定是无家可归。
One of them is a picture of a young child, er, wearing a lot of packing materials on his body, something that looks like Styrofoam, looks very poor, definitely has no home to go to.

However, subject omission is an issue occurring in the source which should not be mirrored in the translation. Since subject drop is common in Chinese source, yet not

grammatical acceptable in English, the subject should be restored in English during translation.

应该说这两国还有相当大的差距。
I should say that there is a relatively large difference between these two countries.

5.11 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:

我们知道，呃，在现在整个岛内，不管是蓝营，还是绿营，还是媒体，还是主流民意，都是一片尊重司法， 呃，遵守司法判决的这个结果的一片这样子名义的呼声之下。

We know, um, that currently, on the whole island, regardless of whether it is the blue camp, the green camp, the media, or mainstream opinion, across the board, they all respect the judiciary, um,((and voice the opinion that the result of the judiciary's judgment should be respected.))

我对这种文艺晚会已经厌倦了，特别是国内几年的春节晚会，生硬的拼凑、低俗的搞笑和灌输内容的搀杂 ，让我失去了对文艺晚会的兴趣，体内的文艺细胞随着年龄的增长似乎也逐渐的消亡了 。
I had become bored of this kind of cultural gala, especially the Spring Festival galas inside the country over the last few years. The awkward way they are slapped together, the lowbrow joke-making and the ((inferiority instilled into the content)) have caused me to lose interest in entertainment galas, while my body's literary and art cells seem to be gradually dying off as my age advances.

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

5.13 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).

5.13.1 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.

5.13.2 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}.

从二零零六年 {breath} 开始到现在，经过媒体报道的类似的事件已经不少于十起了。
Starting from 2006 {breath} and up until the present, no less than ten similar incidents

have been reported in the media.

5.13.3 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.

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

我们将继续和，%嗯，北朝鲜进行会谈。
We will continue our talks with, %uh, North Korea.

这个问题，%啊，很严重。
This problem is, %ah, very serious.

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.

NOTE: Filled pauses in Chinese transcripts will be marked with a % symbol before the word. This symbol should be carried over to the translation.

5.13.4 Translation of ”嗯”

In Chinese conversational speech, 嗯 can be used in many ways. Translators should differentiate the multiple uses of 嗯 and translate accordingly.

嗯 can be used as a filled pause. See section 6.10.2 for discussion of how to translate filled pauses.

嗯 can be used 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: 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

In such cases, 嗯 should be translated to its English counterpart, such as *uh-huh* or *yes*.

Finally, 嗯 can be used to answer questions as in:

A: 你能不能把那本书给我递过来?
B: 嗯

When 嗯is used to answer a question, it means *yes* and should be translated as such.

5.13.5 Repetition and Restarts

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

这一次呢，很具体的是有一个叫亚非商，亚非的商务峰会就是今天，在雅达，呃雅加达举行了。
This time, very specifically, a business summit called Asian-African Business, Asian-African Business Summit was held exactly today in %pw, er, Jakarta.

*Incorrect: This time, very specifically, a business summit called Asian-African Business Summit was held exactly today in Jakarta.

因为她，那个女的那句台词就是说， 不要不一不能完全好，尽量好就可以了。
It was because she, the girl's line was that everything did not have to be, not, could not be perfect, it was okay as long as it was as good as possible.

不过最近（（））有一个，凤凰有一个很很新闻说， 是在酝酿，国家正在酝酿这个新的政策出台。就是那个房地产税，
However, recently, (()) there is a, a piece of news reported by Phoenix about the deliberation, that the state is deliberating upon the publication of a new policy, which is the real estate tax.

是吧？我愿意把这些，我传授给你.
Right? I'm willing to pass these, I'll pass them all on to you.

5.13.6 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:

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.

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.

5.13.7 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:

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 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:

我们的目的是让所有((塞尔维亚人))联合起来。

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

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.

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)).

5.13.9 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.

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
- Overlooked file or section: translation missing from a big portion or a file

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 A -- 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

Appendix B – Commonly Made Mistakes

The following are real examples of common mistakes made by translation agencies in the course of translating a file. The format are as follows: the first line displays the original source text, the second line shows the original translation made by the agency, and the third line shows the correct made by LDC, which should be the standard for translated text.

A1. Missing Elements in Translation

A1.1 Missing Verb

会议**预定**十二月六日至八日在非洲马利举行.
The Conference will be held in Mali, Africa, from December 6 - 8.

The Conference **is scheduled** to be held in Mali, Africa, from December 6 - 8.

李长荣老师在给同学们讲解数学题， **发现**窗外突然出现了一名二十多岁的年轻男子。
Teacher Li Changrong was explaining a math problem to the pupils when a young male in his twenties suddenly appeared outside the window.

Teacher Li Changrong was explaining a math problem to the pupils when **she found that** a young male in his twenties suddenly appeared outside the window.

A1.2 Missing Modifiers

病毒一旦扩散到**防疫**准备不足的国家,
Once the virus spreads to regions that are not well prepared,

Once the virus spreads to countries that are not **preventively** well prepared,

因为网络成瘾的话，在**我们**中国就是一个比例非常的高 ，高达百分之十三。
Because in China, there is a very high rate of Internet addiction, as high as 13%.

Because in **our** China, there is a very high rate of Internet addiction, as high as 13%.

纵观**国内外**视频搜索引擎的现状，可以看出，其发展空间是非常大的，前景**也非常**看好。
Looking at the current situation of domestic video search engines, it is apparent that the scope for development is tremendous, and its future prospects optimistic.

Looking at the current situation of **domestic and international** video search engines, it is apparent that the scope for development is tremendous, and its future prospects **are also very** optimistic.

A1.3 Missing Nouns

那么一个呢是这个当时呢这个英国很快就召见了这个俄罗斯驻这个英国的大使，呃，要求他们这个尽快提供**相关的这个情况**，这个配合这个调查。

So, uh, at the time, Britain very quickly summoned the Russian ambassador to Great Britain, uh, to ask that they promptly provide, uh, cooperation in the investigation over this affair.

So, uh, one is that at the time, this, Britain very quickly summoned, this, the Russian , this, ambassador to Great Britain, uh, to ask that they promptly provide **related information** and, this, cooperation in the investigation over this affair.

A1.4 Missing Filled Pauses

然后，**呃**，立即在就近的原则下，指派相应的急救人员进行相应的救治。
Then, immediately, based on the principle of proximity, it assigns corresponding emergency personnel to carry out the relevant treatment.

Then, u**h**, immediately, based on the principle of proximity, it assigns corresponding emergency personnel to carry out the relevant treatment.

这个案子呢，当中有的刑警就提出来呢，就是前三起啊包括第四起，犯罪分子都把受害人，呃，捆扎起来.

Some of the criminal police officers on this case proposed, that is, in the three earlier cases, ah, including the fourth one, the criminal always tied up the victim.

Some of the criminal police officers，eh, on this case proposed, eh, that is, in the three earlier cases, ah, including the fourth one, the criminal always，er, tied up the victim.

A1.5 Missing Sub-clauses

两位专家指出,除了**一月份在北京举行的**第三次防治禽流感会议所决定的捐款十九亿美元之外
The two experts pointed out that apart from the 1.9 billion U.S. dollars decided upon during the Third Global Bird Flu Prevention and Control Conference,

The two experts pointed out that apart from the 1.9 billion U.S. dollars fund decided upon during the Third Global Bird Flu Prevention and Control Conference **that was held in January in Beijing,**

关于青少年网络成瘾这个事件，我们说的不是一天两天了。
In terms of the addiction of young people to the Internet, it's not just one or two days.

In terms of the addiction of young people to the Internet, **we are not talking about** one or two days.

A2. Inaccurate Elements in Translation

A2.1 Inaccurate Nouns

病毒一旦扩散到防疫准备不足的**国家**,如非洲、中东国家.
Once the virus spreads to regions that are not well prepared, such as Africa and the Middle

East,

Once the virus spreads to **countries** that are not preventively well prepared, such as Africa and the Middle East countries,

在阿芳的耐心**劝导**下， 蒋政国的情绪似乎有些平静。
Under Afang’s patient induction, Jiang Zhengguo’s mood seemed to calm down somewhat.

Under Afang’s patient **persuasion**, Jiang Zhengguo’s mood seemed to calm down somewhat.

扩大捐款与**经费需求**等问题将讨论解决。
that issues such as increasing donations and operating funds would be discussed and resolved.

that issues such as increasing donations and **the need of funds** would be discussed and resolved.

A2.2 Inaccurately interpreted phrases

国家要加大对于这种网吧这种经营场所的管理力度，更重要的是**帮助这些孩子们戒除网瘾**，首先是孩子身边的成年人。
The state should enhance its management of businesses such as Internet cafes, and the most important thing is to rid these children from Internet addiction, and this should start from the adults around these children.

The state should enhance its management of businesses such as Internet cafes, and the most important thing is to **help these children to get rid of Internet addiction**, and this should start from the adults around these children.

A2.3 Inaccurate Numbers

那么据说呢它这个比氰化物的这个毒性要毒**二点儿五亿倍**，是吧？
So reportedly, it is 2.5 times more toxic than cyanide, right?

So reportedly, it, this, is **250 million times** more toxic than cyanide, right?

A2.4 Inaccurate Parentheses

在这个九五年以前，**这个，这个**，那时候，不存在这个问题。
This issue did not exist at that time, um, um, before 1995.

This issue did not exist at that time, **uh, uh,** before, this, 1995.

何先生，我们知道**那个，那个，**两个法律，一个是这个反垄断法，还有一个是物权保护法，那么
Mr. He, we know the two laws: one is an anti-monopoly law, and the other is a property rights protection law, so
Mr. He, we know, **uh, uh**, the two laws: one is an anti-monopoly law, and the other is a property rights protection law, so

A2.5 Inaccurate interjections

哎， 后来我们一看，{breath} **哎呀**， 这个不得了， 这是一个时装设计师啊。
Ah, then we took a look, {breath}, ai-ya, my goodness, this is a fashion designer!

Ah, then we took a look, {breath}, **oh my goodness**, this is a fashion designer!

A3. Extra Elements in Translation

...解决了一个大问题，
...solved a **very** big problem.

...solved a big problem.

此外外汇管理局提高了境内机构经常项目外汇帐户保留外汇的限额。
In addition, the administration of Foreign Exchange **also** raised the ceiling of foreign exchange retained by domestic organizations in the foreign exchange accounts under current account.

In addition, the administration of Foreign Exchange raised the ceiling of foreign exchange retained by domestic organizations in the foreign exchange accounts under current account.

对， 高。
Yes, **very** high.

Yes, high.

经常引发问题的网路未必是最显着的。
The **types of** websites that frequently lead to problems may not be the most obvious.

The websites that frequently lead to problems may not be the most obvious.