Guidelines for

BOLT Chinese-English Word Alignment

Version 2.0 – April 10, 2014

Linguistic Data Consortium

Created by: Xuansong Li xuansong@ldc.upenn.edu
With contribution from: Niyu Ge niyuge@us.ibm.com
Stephanie Strassel strassel@ldc.upenn.edu

TABLE OF CONTENTS

1 Introduction

This version of word alignment guidelines used for the BOLT project was developed based on the guidelines for the GALE word alignment project.

The task of word alignment consists of finding correspondences between words, phrases or groups of words in a set of parallel texts. The resulted annotated data can be used as gold standard training data for machine translation. With references to Blinker project guidelines and ARCADE project guidelines, this guideline is especially designed to suit the task of Chinese-English word alignment, and a visualized tool is developed by LDC to facilitate the task.
In this guideline, the data used for word alignment is first presented in the beginning section. In the section followed, the tasks are specified and the conventions adopted in this guideline are explained for better understanding. In section 4, the general strategies of annotation are addressed to deal with universal language features in word alignment. Then more detailed specifications and rules are elaborated with examples in section 5. Section 6 handles unaligned words. Section 7 describes approaches toward distinctive features of Chinese language. Section 8 handles alternate translations. Last section discusses discussion forum features with some examples.

2 Data

The data type is discussion forum.

Tokenization is done automatically without human corrections.

Tokenization of English follows the same guidelines used in Penn English Treebank: split words by white spaces, separate punctuations from the preceding/following words, apostrophe S (‘s) is treated as separate tokens. Penn English Treebank treats most hyphens as separate tokens, but some as part of words.

We tokenize the Chinese text as follows: each Chinese character is treated as a separate token; English words in Chinese text are tokenized the same way as described in the previous paragraph; In addition, we separate punctuations from the preceding/following words/characters. In Chinese word alignment project, all hyphens are treated as separate tokens.

3 Tasks and Conventions

3.1 Tasks

a) Link words or phrases in a source language (Chinese) to those in a target language (English)
b) Make judgments on translated or not-translated parts in source and target language
c) Attach unmatched words to their related parts according to attachment rules
d) Reject the alignment using the “reject segment” button for blank sentences, unmatched sentences, half translated sentences and pure English sentence on both sides.

3.2 Conventions

For word alignment annotation and compilation of this guideline, the following terms and symbols are employed.

a) translated, not-translated, correct, incorrect (labels appeared in the tool for links and markups)
b) <> (this symbol is used to show equivalence between source language and target language, e.g. “牛奶” <> milk)
c) parenthesis ( ) is used to indicate emptiness or omission (e.g. ()<>the, which indicates “the” appears in translation but no counterpart in source language.)
d) colors and squares in examples (for illustration of proper correspondent links and markups, blue color and square representing corresponding links, and yellow color and bold representing markups)

4 Concepts and General Approach

For efficiency and accuracy, some points about good practice need to be emphasized and memorized before annotation. Of vital importance to the whole successful alignment process, they are bulleted below before the general strategies are discussed.

- Before the actual alignment, in order to get the main idea, an annotator needs to glance over both the source sentence and target sentence in the two windows on the right side of the tool interface.
- Based on and to begin with the source sentence, an annotator should first link all the content words before they move on to function word links.
- With all the equivalent links mapped, the left-over words or phrases can be either glued to other parts or marked up as not-translated.
- Most of the links are translated and correct links; “not-translated correct” links are chiefly for purely functional or grammatical word insertions due to language idiosyncrasy, or for contextual insertions for the sake of effective communication and discourse coherence.
- All words in both source and target languages should be linked or marked. No single piece could be left unattended. An annotator can reject to align

a sentence if he/she thinks that sentence is not a suitable one for alignment.

4.1 Translated versus Not-translated

4.1.1 Translated

In translating from one language to another, it is common there are several translation versions for a particular word or sentence in source language. All these versions may convey the same meaning, i.e., they are semantically equivalent, but only different as a result of choice of words or styles. Thus, each of these versions is the correct translation of the source words or phrases. Therefore, all these versions are considered to be “translated” and “correct”. For mapping, an obvious lexical item or items can be found in source and translation.

E.g. 他拒绝了他的提议. <> (version 1) He refused his suggestion.
(version 2) He declined his suggestion.
(version 3) He said no to his suggestion.

E.g. most-favored nation trade status
最惠国待遇

E.g. 高兴死了。
Extremely happy.

E.g. 如果我是他, …
If I were him, …

E.g. 如果我是他的话, …
If I were him, …

A pair is also treated as “translated” and “correct” even if there is a change in part-of-speech or sentence structure.

E.g. 关系的改善

Improve relations

There are two types of links in terms of “translated” type:

1) Translated and incorrect

If a word or phrase is translated in a wrong way -- either semantically wrong or grammatically wrong -- they are translated but incorrect. Typo or grammatical errors in target language can be treated as translated and incorrect.

E.g. 江泽民说你们是我今年会见的第一个美国国会众议员
Jiang Zemin said, "They are the first US Congressional delegation that I have met this year."

E.g. 香港人无论是从政从商学生老师…
The people in Hong Kong, whether they was teachers, students, or involved in politics or business…

2) Translated and correct

They are the normal links that both properly conveys the meaning and grammatically right. Most links in this task are of this type. The most easily detected link is the one where direct equivalence of both form and content could be spotted, such as:

E.g. 屋子很大。
The room is very big.

A typo in the source side can still make a correct link to the translation when the translation is right.

E.g. 北进市
the city of Beijing

4.1.2 Not-translated

A word is not-translated in the sense that the word is both semantically and lexically missing, or in the sense that the semantic meaning is neither lost nor added, only extraneous words could be found. Thus two cases are recognized for "not-translated" markups.

1) not-translated and incorrect

This kind of markup is proper when both the word form and the information are lost.

E.g. 他买了牛奶**和大米**. (not-translated and incorrect)
He bought milk.

新华社二十五日电
Xinhua News Agency, Beijing, 25th.

For words like “较好，继续” are sometimes translated as “well, continue” to show partial concept entailment between source word and translated words. We treat these words as not-translated incorrect because partial meaning is lost anyway.

她干得较好。
She did well.

我们仍继续做这个课题。
We continue to work on this project.

Topic related extra words are treated as not-translated incorrect because topic clue may be found within or beyond a sentence, and such inference may simply rely on reader’s understanding of the discourse focus.

他去了一个纺织产品展览会， 深深地被纺织产品所吸引。
He went to a textile product exhibition. He was greatly attracted by the products there.

2) not-translated and correct

“Not-translated correct” refers to extra words inserted during translation. These words are either grammatically needed or contextually needed to make the translation grammatical or fluently understandable. For instance, 个, **把, 给, 将, 向…** in the following sentence are all extra words needed for Chinese to make the sentence grammatical and fluent.

E.g. 他是第一个提出此议案的人。
He is the first to propose such a motion.

E.g. 他把鸡蛋吃了。
He has eaten the eggs.

Extra words can be inserted at phrase level, sentence level or discourse level. For phrase (or local) level insertions, we use attachment approach to glue them to their related words (see section 5). For other insertion cases not mentioned in section 5, we mark all of them as not-translated correct (see section 6: unmatched/unattached words)

4.2 Minimum-match

4.2.1 Minimum-match in literal translations

In this guideline, the principle of a word-for-word linkage is strictly followed, that is, the smallest number of words will be preferred.

E.g. instead of 近年
recent years

the correct way: 近年
recent years

E.g. instead of : 今年
this year

the correct way: 今年
this year

When annotators have a difficult time to decide if they should break down a word or not, they may ask “how does a translator translate this word?” For instance,

E.g. 科技
science and technology

去年
last year

能源
energy source

The above cases show the translators are translating each word separately. These words are actually the abbreviated form of extended words. Therefore we separate them and do the sub-part link. Breaking down to sub-part alignment can reveal translation rules and general translation practice at character level.

4.2.2 Maximum-match in idioms and non-literal translations

Sometime as many words as necessary are to be selected for link to ensure a two-way semantic equivalence. In other words, select as many characters as we can to make it a semantically independent unit. Such cases include set or frozen expressions, idioms (or near idioms), proverbs and hyphenated words. Other cases include very sticky structures such as verb phrases (verb particles), structures with suffix or prefix.

E.g. 所谓

What is known as…

E.g. 三个臭皮匠，顶个诸葛亮。
Two heads are better than one.

E.g. 齐头并进

Keep abreast with

However, for proverbs, if a word-for-word translation (content words) is found, sub-part links would be preferred (especial for those directly borrowed from English)

条条大路通罗马。
All roads lead to Rome.

We follow the same approach for proper noun, for which we have a special section (section 5.5).

Non-literal translations other than idioms or proverbs are also very common. In such cases, breaking down to constituent parts would result in semantic-nonequivalent alignments. To avoid such nonequivalent alignments, we would prefer to treat them as a whole unit.

E.g. 带您进入这个话题。
Let's get to this topic.

E.g. 他此刻热血沸腾.
His heart is now racing.

Hyphenated words in English are treated as one unit if the subparts of hyphenated words are inseparable. In tokenization, these inseparable words are tokenized as one word. For those which are not tokenized as one word, subpart word can be aligned.

E.g. 合作
Co-operation

4.3 Attachment Approach

Phrase-level (or local level) extra words are usually attached to their dependency constituents when they are grammatically or contextually needed (see section 5 attachment rules).

When an extra word is glued to its generating head, a more complex situation is that its generating head may be conjunctive or other complex structures, where we can glue this extra word to its nearest head noun.

E.g. 最大的发展中国家和最大的发达国家
the largest developing country and largest developed country

5 Alignment and Attachment Rules

5.1 Anaphora (pronoun)

Anaphora is the co-reference of one expression with its antecedent. The links and markups can be done as follows.

Case 1: Pronoun <> Pronoun

E.g. 卢兹说他一贯支持 和遵循……
Lutz said, he has consistently supported and followed ……

Case 2: Pronoun <> Referent

E.g. 卢兹说他一贯支持中国。卢兹也遵循……
Lutz said, he has consistently supported China, and he followed ……

In Chinese, a subject is usually dropped after its first mention whereas English tends to keep it. We attach the pronoun or subject to its referents.

E.g. eBay 今日宣布 3.1 亿美元收购票务网站 StubHub
EBay Announces Today It will Acquire Ticket Website StubHub for 310 Million US Dollars

E.g. 我记得以前写过篇文章《深圳人均 GDP 之谜》
I remember I wrote an article "The Mystery of Shenzhen's GDP per capita."

E.g. 在 Dreamer(不要问我从哪里来)的大作中提到…
In the contribution by Dreamer (Don't ask me where I came from), it was mentioned…

E.g. 我买了张椅子,很贵。
I bought a chair. That is very expensive.

E.g. 卢兹说他一贯支持…, 并遵循 ……
Lutz said, he has consistently supported… and he followed ……

E.g. 江泽民指出台湾问题关系到中国的主权, 始终是中美关系中最重要最敏感的问题
Jiang Zemin pointed out that the Taiwan question is a matter of Chinese sovereignty. It is always the most important and sensitive issue in Sino-US relations. (“it” is aligned to head word “question”)

We adopt the same approach for reflexive pronouns (`myself," ``yourself," ``herself," ``himself," ``itself," ``ourselves," ``yourselves," and indefinite pronoun (“one”, “some”, “someone”).

5.2 Demonstrative words

Links involving demonstratives (this, that, these, those, here, there) can be illustrated by the following cases.

Case 1: Demonstratives <> Demonstratives

E.g. 我喜欢这椅子。
I like this chair.

E.g. 我喜欢这。
I like this.

Case 2: Demonstratives <> Referent (or vice-versa)

E.g. 我买了张椅子。椅子很贵。
I bought a chair. That was very expensive.

Special cases:

这一著作的诞生非比寻常。
The birth of the book is unusual.

For extra demonstratives, we attach them their head words.

5.3 Measure words

Measure words can have links built from measure words from both sides:

就这个问题，他们进行了三轮会谈。
They had three rounds of talks about the issue.

Extra measure words are common in Chinese, and they will be glued to their head numbers, ordinal numbers or demonstratives:

E.g. 他买了一本书
He bought a book.

E.g. 他买了本书
He bought a book.

E.g. 这个椅子
this chair

E.g. 他是第一个提出此议案的人。
He is the first to propose such a motion.

E.g. 这三张椅子 measure words used together with number to count)
these three chairs

With temporal nouns “Year”, and “Date’ expressing time as in the following examples, we attach them to numbers or ordinal numbers (and later to tag them as measure word maker) when there is no equivalent.

一九四九年是不寻常的一年。
1949 is an unusual year.

他六月三日来到上海。
He came to Shanghai on 3rd of June.

5.4 Copular BE (link verb)

Link verb “to be” (am, is, are, be, been, being) qualifies or informs about the properties of its argument. Many languages do not use the verb "to be" to separate an adjective from its noun, such as in Chinese, then in link mapping, the extra link verb can be treated as “not-translated” and “correct”. Compare the following examples:

E.g. 中方是愿同美方共同努力增加共识的。
China is willing to strive with the US to increase mutual understanding.

E.g. 中方愿同美方共同努力增加共识.

China **is** willing to strive with the US to increase mutual understanding.

E.g. 她又高又苗条。
She **is** tall and slender.

5.5 Proper Noun

A proper noun is used to name a specific item. It is usually a one-of-a-kind item and it always begins with a capital letter.

We follow the basic minimum-match approach in aligning proper nouns, especially when the proper noun is compositional (usually one or more specific name plus one or more common nouns).

E.g. 山东省
Shandong Province

清朝
Qing Dynasty

西湖
the West Lake

黄河
the Yellow River

山东鲁能集团
Shandong Luneng Group

北京大学
Beijing University

第一服装厂
First Clothing Factory

E.g. 中东
the Middle East

白宫
the White House

There are cases where the proper noun in source conveys the same or a derivative concept as that in translation, which are semantically and contextually equivalent, so we treat it as a whole unit. When it is hard to decide for this case, the annotator can try to test it context-free or in any context, as in the following example, “Shandong” is equal to “Shandong Province” in meaning in any context.

e.g. 山东
Shandong Province

For names of people, we align first name and last name separately.

e.g. 毛泽东
Mao Zedong

For a name of country in its abbreviated form, we treat it as a whole unit.

e.g. 越南
Vietnam

美国
the United States of America

When the name is not abbreviated and has good one-to-one match, we'll break down to sub-piece alignment.

e.g. 中华人民共和国
the People's Republic of China

Acronyms of translation of proper nouns are treated as a whole unit.

E.g. 世界卫生组织
WHO

E.g. 中央电视台
CCTV

Abbreviations of Chinese for proper nouns can be linked separately if there is one-to-one match.

E.g. 北大
Beijing University

E.g. 文革
the Cultural Revolution

E.g. 中美两国
The two countries of China and the US

However, if there is no perfect match for the case of abbreviations of Chinese for proper nouns, then we glue them as a whole unit.

E.g. 中纪委
the Central Discipline Inspection Commission

5.6 Determiners (articles in English)

Determiners, here particularly articles in this guidelines, are used before nouns to indicate whether you are referring to something specific or something of a particular type. The definite and indefinite articles "a/an/the" are all determiners.

Sentences involving articles can have the following cases.

Case 1: Determiner <> Determiner

E.g. 卡斯特罗称赞中国的成功是一个奇迹
Castro praised as a miracle China's success.

Case 2: ( ) <> Determiner (or vice-versa)

E.g. 新年伊始
the beginning of the New Year.

It's common to have determiner on one side and omitted on the other side, then glue the extra determiner to its head noun followed even if this head noun is sometimes a kind of far away from its determiner.

Case 3: Determiner <> different translations (or vice-versa)

It happens occasionally a determiner appears on one side while non-determiner translation equivalence appears on the other side. When a determiner is translated into different versions (or vice versa), we make the mapping according to the meaning expressed.

E.g. 我给了他一本书。这本书很有趣。
I gave him a book. The book is interesting.

5.7 Auxiliary verbs

An auxiliary or helping verb is a verb giving further semantic information about the main or full verb following it. An auxiliary verb can have many forms. It is used to serve such functions as passive, progressive, perfect, modal, or dummy. Whenever the auxiliary verbs appear on both sides, simply link them. When missing on one side, the extra auxiliary verbs will glue to the main verb.

E.g. 他不知道。
He does not know.

E.g. 蜡烛吹灭了。
The candles were blown out.

E.g. 他不知道。
He doesn't know.

The last example is an exception since “n’t” is connected to “does”. Other such examples include “couldn’t”, “wouldn’t”, “didn’t”, and so on. (in future annotations, we can avoid such exceptions by tokenization)

The word "do” may be used for "emphasis", where it should not glue to any part.

5.8 Infinitive "to"

In Chinese-English alignment, infinitive “to” is glued to the verb following it. Occasionally, some words might be inserted in between, a case which requires extra attention to make a right link.

E.g 教授让他迅速有效地解决问题。
The professor asked him to quickly and efficiently solve the problem.

Occasionally, infinitive “to” may find its translation equivalent, where a direct link would be possible.

E.g. 为了解决问题，我们讨论了一下午。
To solve this problem, we discussed the whole afternoon.

5.9 Expletives

Here “expletive” chiefly refers to syntactic expletives, words that perform a syntactic role but contribute nothing to meaning, such as “it”, “there” “here”. For cases where an equivalent counterpart can be found, just link them. If without equivalence, the extra “it”, “there” or “here” can be marked as “not-translated” and “correct”.

E.g. 努力准备考试很重要。
It is important to work hard for the exam.

E.g. 汽车来了。
There comes the bus.

5.10 Conjunction

5.10.1 Conjunctions with “and”

Four cases of conjunctions are involved under this category:

a) and <> and
b) and <> , (or vice versa)
c) and <> ( ) (or vice versa)

For case a), simply link both sides.

E.g. 美国政府和美国国会
the US government and the US Congress

For case b), link comma to comma, the extra “and” is treated as “not-translated” and “correct”.

E.g. 中方希望美国政府和美国国会谨慎妥善处理台湾问题，不要让台湾问题干扰中美关系的稳定发展。
China hopes that the US government and the US Congress will carefully and appropriately handle the issue of Taiwan, and not allow the Taiwan issue to interfere with the stable development of Sino-US relations.

For case c), link comma to “and”

E.g. 他挥手再见，开车走了。
He waved good-bye and drove away.

For case d), the extra “and” is marked as “not-translated” and “correct”.

E.g. 他抬头对我说……
He looked up **and** said to me that……

5.10.2 Conjunctions without “and”

Conjunctions other than “and” structures could be illustrated with the following examples:

E.g. 他既不吃也不喝。
He neither eats nor drinks.

E.g. 或者是中国或者是美国站出来承担责任。
Either China or the US stands out to shoulder the responsibilities.

5.11 Prepositions

For prepositions, if the equivalent parts can be detected, simply link the correspondent parts.

For cases of two or more propositions in English while one proposition in source side, glue the two prepositions and link them to the preposition counterpart in the other side.

For cases of one preposition in one side while no counterpart in the other, the extra proposition will be glued to the object of this preposition even if the object is a kind of far away.

合作领域广阔
a broad sphere of cooperation.

For cases where the extra preposition appears after a verb, it is glued to its object in spite of an obvious co-occurrence of the verb and the preposition. This same rule holds true to the preposition after a noun.

E.g. 我同意你的意见。
I agree with you.

But:

E.g. 我同意你的意见。
I am in agreement with you.

When the object of an extra preposition is a present participle, attach the preposition to the participle.

E.g. 做实验可以得出结果。
Results can be obtained by doing experiments.

In some cases, the two preposition counterparts do not have an exact semantic match (i.e., divergent propositions), but we still link them.

5.12 Verb Particles

A verb particle, also called a phrasal verb or multi-word verb or compound verb, is a verb combined with a preposition, or an adverb, or an adverbial particle. They appear in the form of two or more words combined together. Since these particles are sticky and are inseparable from their verbs when forming a fixed meaning, we treat them as a whole unit for this task (i.e, glue the particle to the verb).

E.g. 指出
point out

5.13 Possessives

Possessives are used to indicate possession. English possessives can take two forms: "'s" and "of". If possessives appear on both sides, just link them.

E.g. 中国的领土
Territory of China

E.g. 中国的领土
China 's territory

For extra “of” in English, it can glue to its “owner”.

E.g. 中国领土
territory of China

If the English plural possessive is only an apostrophe, then it is the same as the apostrophe with “s” since they are semantically equivalent.

E.g. 学生的教科书
Students ' textbooks

For extra ‘s where no equivalent can be found in source, they are marked as “not-translated correct”.

E.g. 中国领土
China's territory

E.g. 他带来了他的书。
He brought his books.

E.g. 他带来了书。
He brought his books.

Here in the first example, an exact lexical and semantic match can be detected in the link, which is an easier case to find and deal with. However, in the second example, only one lexical item “books – 书 ” can be found while “his” is omitted in source. A close examination of such omission could lead us to the conclusion: “his” is lexically omitted from the source language but is semantically conveyed in the source. In other words, only the word form is dropped. Such drop-offs happen locally or in word-level.

5.14 Passive Sentences

If passive voice appears in both source and target languages, it might be easy to link each piece correspondingly. Whereas for the case of passive voice in source and active voice in the translation or vice versa, the word order would be very different and it's not easy to tear apart the corresponding parts. Great care is needed in correctly mapping the links.

E.g. 这是中央台报道的。
This is reported by CCTV.

E.g. 能否妥善处理台湾问题
Whether or not the issue of Taiwan can be appropriately dealt with

E.g. 桥冲垮了。
The bridge was destroyed.

E.g. 参议员做出的努力
efforts made by Senator

E.g. 由参议员做出的努力
efforts made by Senator

In Chinese, the passive voice can be easily recognized by the use of word like “被”, “给”, etc. These preposition words introduce the doer and change the normal word order of an active voice sentence. In alignment, these words can sometimes found its equivalents in translation, like in the following:

E.g. 飞机残骸是由两位青年农民发现的
The remains of the plane were discovered by two rural youths.

E.g. 桥被洪水冲垮了。
The bridge was destroyed by the flood.

However, these words have no equivalents in translation in many other cases, where we attach these extra words to main verbs to show the passive feature of the sentence.

E.g. 桥被冲垮了。
The bridge was destroyed.

5.15 Subordinate Clauses

When the clauses are placed in a hierarchical structure, the more important ones are main clauses and the less important are called subordinate clauses since they cannot stand on its own. A subordinate clause is introduced by a subordinate conjunction or a relative pronoun and contains both a subject and a verb. Subordinate conjunctions include *after, although, as, because, before, even if, in order that, once, provided that, rather than, since, so that, than, that, though, unless, until, when, whenever, where, whereas, wherever, whether, while, why,* etc. (see the following examples). In Chinese-English mapping, these conjunctions can normally find their lexical counterparts. However, for relative pronouns, one cannot always find a match in target or in source. Relative

pronouns are words like *that, which, whichever, who, whoever, whom, whose, whosever, whomever*. For Chinese-English task, the relative pronouns are usually omitted and the relative pronouns could be presented in other forms, such as by the structure "…的".

According to functions, subordinate clauses can work as subject, object, complement and adverbial.

a) clause as subject
The extra clause markers are treated "not-translated correct".

E.g 他是学生是个事实。
That he is a student is a fact.

b) clause as object

In Chinese-English task, the extra markers are attached to head verbs. Or it can be aligned to comma in source.

E.g. 江泽民指出台湾问题关系到中国的主权.
Jiang Zemin said **that** the Taiwan question is a matter of Chinese sovereignty.

E.g. 江泽民指出,台湾问题关系到中国的主权.
Jiang Zemin said **that** the Taiwan question is a matter of Chinese sovereignty.

a) clause as adverbial

Adverbial clause markers can usually find their counterparts, such as:

E.g. 因为缺货，我们不能卖给你所要的书。
We cannot sell you the book you need because we are out of stock.

d) Adjectival clauses

With no equivalent counterparts in source language, the relative markers are glued to their antecedents since no counterparts can be found, as in these examples:

E.g. 两位瑶族青年农民上山采药
two rural youths belonging to the Yao ethnic minority **who** had climbed the mountain to collect medicinal herbs.

E.g. 我在菲律宾马尼拉同克林顿总统再次举行了会晤

I was in Manila, the Philippines, **where** I met again with President Clinton.

d) clause as apposition

For clauses used as apposition, the relative markers are glued to their antecedents.

E.g. 我们完全同意中方的立场世界上只有一个中国
We fully agree with the Chinese position **that** there is only one China in the world.

5.16 Punctuations

All the equivalent punctuations can be linked, such as comma-to-comma, period-to-period, and so on. Extra punctuations can be marked “not-translated” and “correct”. Special cases in Chinese are:

Case a) : <> , (for Chinese-English alignment, map the link as “correct”)

E.g. 他说: “…..”.
He said, “…..”.

Case b) , <> . (comma in one side, period in the other)
For this case, link comma to period.

Case c) “and” <> “,” (or vice versa) (see section 5.3.5.1)

Case d) ( ) <> various punctuations (or vice versa)

For cases where there is a punctuation mark in one side while it disappears in the other side, the extra punctuation mark is treated as “not-translated” and “correct”.

E.g. 新华社23日电。(记者:王玮)
Xinhua News Agency, 23rd, by reporter Wang Wei.

5.17 Contextually attached words

During translation, translators may add some contextual words for better understanding. Without these extra words, the grammatical structure may be acceptable, but the semantic meaning is not sensible. These extra words are added based on word association or collocation clues. We attach these extra words to the words associated or related, like in the following examples.

大家好！欢迎收看这频道。
Hello, everyone! Welcome to this channel. (收看 is attached to 欢迎)

参加第六届中越青年友好会见活动的青年朋友们，正相聚在这里，等待着中越领导人的到来。
The young people participating in the sixth friendship meeting of the Sino - Vietnamese youth were gathering here and waiting for the arrival of the leaders of China and Vietnam. (活动 is attached to 会见)

他对会议的圆满举行感到高兴。
He was glad at the success of the meeting. (举行 is attached to 圆满)

Sometimes words are added to show the change of part of speech of words. These added words are also treated as contextually attached words.

他对她进行威胁。
He threatened her. (进行 is attached to 威胁)

Exceptions:

When a single character word is contextually attached to another single character word, they show strong coherence between each other, therefore we treat them as a unit of real translation rather than a contextually attached word.

我们外商投资。
We welcome foreign investment.

Added words relating to “discourse topic” are treated as not-translated incorrect.

他们去了一个纺织产品展览会，深深被这些纺织产品所吸引。
They went to a textile product exhibition. They were greatly attracted by those products. (the discourse topic here is “textile product”)

5.18 Rhetorically attached words

For structures where repetition is involved, the extra ones are treated as not translated correct.

E.g. 内地的专家和台湾的专家 (share the same noun)
the experts from the mainland and Taiwan

她画了个粉色的车和房子。
She drew a pink car and a pink house.

6 Unmatched/unattached words

Helping words can be used to smooth the sentence, in other words, to make it more Chinese-like, such as 将, 而, 都, 就, 给, etc . These words sometimes carry meaning, but in most cases they are meaningless when they are used to adjust the word order of Chinese or to demonstrate special Chinese structures.

For instance, they carry meaning in the following sentences.

E.g. 就在此时，上海昂立投资公司对秋子的态度却发生了一百八十度的转变.
Precisely at this time, Shanghai ONLY Investment Company completely changed its attitude toward Qiu Zi.

E.g. 信中他表示就自己的言语过失向秋子道歉.
In his letter, he apologized for to Qiu Zi for his verbal offense.

E.g. 吴萍索赔金额将是六百万。
The compensation amount demanded by Wu Ping will be 6 million yuan.

E.g. 他给了她一本书。
He offered her a book.

However, they are not-translated correct in the following examples.

E.g. 他把鸡蛋给吃了。
He has eaten the eggs.

E.g. 国家每年都拨专款用于开展残疾人体育活动。
Every year the nation allocates special funds to be used in developing handicapped sports activities.

E.g. 这个案子当中有的刑警就提出来，就是前三起啊包括第四起，犯罪分子都把受害人捆扎起来.(for smoothing and connection)
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.

Interjection words like “呀”, “啊”, “哇”, etc., in the middle of the sentence show hesitation or change of mind, they should have their counterparts in target. Otherwise they are treated as not-translated and incorrect.

这一工作，啊，实在是，呃，困难而又花时间。
This work, ah, is really, er, difficult and time-consuming.

Interjection words at the end of the sentence such as “吗”, “么”, “啊”, etc. are used together with punctuations to form various types of sentences, in which

case, these interjection words have no equivalents in translation, and we'll glue them to punctuations.

E.g. 天气真棒**啊**！
What a wonderful day!

Insertions or omissions resulting from personal stylistic choice of words, genre-specific insertions and other pragmatic or discourse peculiar features can be treated as not-translated correct. For instance, in broadcast conversations, there are purely colloquial insertions to show hesitations, changes of topic, or personal habits of diction in speaking. They are not semantically important, thus are treated as "not-translated correct".

E.g. 那么敌后战场**呢**，日本为了要夺取，巩固他**这个**占领区，他开始**那个**新的战略。
Well, on the battlefield behind enemy lines, in order to take over and consolidate the area under its occupation, Japan began a new strategy.

E.g. 就是以这个据点**呢**为一把锁，**这样呢**进行一个囚笼子作战。
That was to use strong holes as a lock, to carry out siege warfare.

Apart from purely functional insertions or omissions, discourse semantic insertions or omissions are also treated as "not-translated correct" because, in such cases, the semantic connection is so distant and discourse context-depend that we almost have the tendency to treat them as "not-translate incorrect". For instance, at the beginning of an article, where we talk about "Chinese basketball team", and in subsequent sentences we omit the words "Chinese" from "Chinese basketball team" since it is obvious that "basketball team" and "Chinese basketball team" are identical from a discourse perspective. However, if examined from a word-level or locally, we can never say these two terms are semantically equivalent. They are equivalent only by adding discourse clues.

Extra conjunctions are also not-translated correct.

E.g. 他们的基础知识差，出现了一些错别字，无法纠正，结果落选。
They have poor basic knowledge, they wrote some wrong characters, and **since** there was no remedy, they finally lost in the selection.

A special case under this category is that sometimes there are English words inserted into the source language, which serve as an explanation of noise from audio file. They can be marked as "not translated" and "correct". (e.g. {laugh}, {breath}).

For other minor cases which are treated as "not-translated correct" will be touched separately in later sections.

7 Special Features in Chinese

7.1 Tense in Chinese

Because of the lack of inflectional morphology, Chinese verbs do not change forms, and “tense” and “time” are chiefly expressed in three ways: a) using nouns which indicate time, such as “昨天，三年前”; b) using 着, 了, 过 to express “present” and “past” behavior or state; c) using adverbials, such as已, 已经, 曾, 曾经, 正, 正在, 没, 没有; d) semantic implication, which is knowledge-base, like we all know “Paris Commune” is a past existence, and thus 巴黎公社规定 is translated as “Paris Commune stipulated the rules that …”, where past tense is used in English.

For “tense” annotation, we attach 着, 了, 过 to verbs. For adverbials like 已, 已经, 曾, 曾经, 正, 正在, 没, 没有 which express “tense” in Chinese, we align them to their adverbial counterparts in English, otherwise, we attach them to verbs. There are also cases where 着, 了, 过 may not function to express “tense”, then we handle them case by case (for example, they may be not-translated correct or real translation).

E.g. 他听着音乐。
He was listening to music.

E.g. 我不喜欢他整天开着玩笑。
I don’t like him joking around all day. (着 is not translated correct because of no verb auxiliary can be found)

E.g. 我正等着他来。
I am waiting for him to come.

E.g. 我正在等他到来。
I am waiting for him to come.

E.g. 他正访问北京。
He is visiting Beijing.

E.g. 他去北京了。
He has gone to Beijing.

他去过北京。
He has visited Beijing.

E.g. 图书馆已经建成。
The library has already been completed.

E.g. 图书馆已建成。
The library has been completed.

7.2 Duplication

Duplication or repetition in Chinese is common, and exists in various forms. Generally, the duplicated characters are glued together in alignment if no repetition is found in English.

7.2.1 Noun duplication

E.g. 人人都到了。
Everyone arrived.

7.2.2 Verb duplication

E.g. 让他试一试。
Let him have a try.

E.g. 让他试试。
Let him have a try.

E.g. 让他解释解释。
Ask him to explain。

E.g. 他解释了又解释。
He explained and explained.

E.g. 他去散散步。
He went to take a walk.

E.g. 他做作业做得好。
He did his homework well.

7.2.3 Adjective duplication

E.g. 太阳红彤彤的。
The sun is red.

E.g. 屋子干干净净。
The room is clean.

E.g. 他傻不拉几的。

He is stupid.

7.2.4 Measure word duplication

E.g. 花一朵朵地凋凌。
The flowers withered gradually.

7.3 Separated verbs

The separated parts of a verb in Chinese should be glued together before a link is mapped.

E.g. 他深深地鞠了一躬。(glue and then link)
He made a deep bow.

E.g. 睡觉前洗个热水澡。(glue and then link)
Take a hot bath before going to bed.

7.4 “的” “地” “得”

“地” “得” can be attached to the words they modified.

E.g. 他干得好。
He did well.

E.g. 他深深地鞠了一躬。
He made a deep bow.

The usage of “的” is flexible. It assumes various grammatical functions. We can align it to different types of words.

a. align to clause marker (what, who, which, whom, whose, where, ….)

E.g. 他描述了他所看到的。
He described what he saw.

b. align to possessive words (preposition, ‘s, ‘)

E.g. 中国的领土
The territory of China

c. align to preposition (all kinds of prepositions)

螺旋桨是推动船舶前进的关键部位，直接影响船舶的性能，噪声等技术指标，

Propellers are key parts for driving boats forward and have a direct impact on a boat's technical indices such as performance and noise.

数控技术的重要性可见一斑。
This gives you an idea of the importance of numerical control technology.

When there is no equivalent alignment, “的” can be attached to the word with which it has a relationship.

a. attach to verb

他看了提交的报告。
He went through the report submitted.

b. attach to noun

E.g. 分享价值25万美元的20根金条
Sharing 20 gold bars worth 250,000 US dollars

E.g. 他是个二十岁的青年。
He is a 20 year old youth.

c. attach to adjective

E.g. 雪白的雪。
White snow

d. attach to preposition (attach to noun when preposition is missing)

她挥舞着手中的旗子。
She was waving the flag in her hand.

她挥舞着在手中的旗子。
She was waving the flag in her hand.

这是来自河南的张先生。
This is Mr. Zhang from Henan.

“的” can also be used at an end of a sentence to show a statement, where we attach it the punctuation followed.

E.g. 中方是愿同美方共同努力增加共识的。
China is willing to strive with the US to increase mutual understanding.

There are cases “的” can be used in idioms or set expressions or a part of a word, where it then becomes a constituent part of a translated correct link.

E.g. 所以我想不幸的是在过去一个多月的时间里，我们没有看到来自美国国内的一些主流菁英。
So, I think unfortunately, over the past one month or so, we haven' t seen a kind of deep reflection, like that against the Vietnam War in those years, among some mainstream elites in the United States ,

E.g. 所以因此的话，在那个动迁进来一个月，就是九月五号到十月四号那一个月，我天天都到那条街上去.
Therefore, as a result, during the month of the relocation period, which was the month from September 5 to October 4, I went to that street every day.

7.5 Prefix and Suffix

7.5.1 Prefix

Prefixes in Chinese “本, 该, 此” are usually linked directly to the counterparts without gluing since they carry actual meaning. However, “本人” is an exception since it has a new meaning when combined. Again apply here the rules about word segmentation. If a two-character word is highly sticky and co-appeared, and no character can be inserted in between, then treat it as whole unit.

E.g. 本人姓李。
My family name is Li.

E.g. 我遇到过此人。
I have met this person before.

E.g.该书描述了他的一生。
The book describes his whole life.

E.g.该组织有五人。
There are five people in the organization.

E.g. 她是他的前妻。
She is his ex-wife

But:

E.g. 她是他的前妻。
She is his former wife.

E.g. 前总统克林顿

Former president Clinton

7.5.2 Word Suffix

Suffixes are normally glued to words before them since they usually do not carry meaning.

7.5.2.1 suffix after noun

Noun suffixes like “者，儿，员 …. ” glue to nouns.

E.g. 汤姆是个小青年儿。
Tom is a young man.

5.5.2.2 suffix after verb

Verb suffixes like “来，去，…. ” glue to verbs.

E.g. 他继续写下去。
He continued to write.

5.5.2.3 suffix after preposition

Suffixes after prepositions are of many forms and very flexible in usage. How to glue them is shown below.

E.g. 对美国来说
To the US

E.g. 在近年来
In recent years

E.g. 在桌子上面
On the table

7.5.3 Sentence suffix

Some characters such as “的, 吗, 么, 呀, …” appear at the end of a sentence, which serve to indicate a sentence type. They are attached to punctuations followed.

E.g. 你去过北京吗？

Have you ever been to Beijing?

E.g. 真漂亮呀！
How beautiful!

7.6 Word Segmentation

For the Chinese-English word alignment, word segmentation task is combined into the annotation process.

7.6.1 General Principles

In alignment it is usually hard to make a decision if the two-character words or three-character words are one-words. If the annotator can find perfect equivalence even with two characters, they might as well break down into sub-piece links.

E.g.经济强国

Economically-strong country (literal translation)

E.g. 经济强国
Economic power

7.7 Conflicting rules

7.7.1 Possessive versus co-reference

他带来了书。
He brought his book.

Higher priority: possessive rule (i.e., attach “his” to “book”)

8 Alternate Translations

In cases where two versions of translation are provided for the same source word or phases (common for idioms, proverbs or some special web languages), the alignment should be performed on the version which is more literally translated. The other version will be tagged as an alternate translation (see tagging guidelines Section 3.3).

E.g.

问一下出过书的楼主，哪国的奴隶的身份人有投票权？
I want to ask the [original poster | floor host] who has published books a question , in which country do people who are slaves have the right to vote?

9 Informal Language Features and Alignment Examples

The alignment annotation is particularly difficulty with informal language, such as weblogs, twitters, emails, informal discussions. Typical features of such languages includes: 1) inappropriate content and language, such as obscene, sexual and threatening language, or words or phrases associated with fraudulent schemes, chain letters and other common types of unsolicited email or messages; 2) non-standard use of punctuation or signs such as multiple exclamation marks “!!!”, question marks “???” or special “%&#$@*” characters, or use of all capital letters; 3) frequent use of emoticons; 4) informal abbreviations, slang, acronyms, etc. in text messages, such as “lol” and “IMO”; 5) numbers representing words such as numbers 520 in Chinese for "I love you" (我爱你) and 748 for (去死吧) "go to hell"; 6) old words taking new meaning such as “杯具”, originally only indicating “cups”, now having the new meaning “tragedy”; 7）new coined words like “宅男” (referring to “youth staying at home addicted to internet, computer games, etc.”). Here are some alignment examples with heavy language features of the discussion genre.

Examples:

Careless typos:

视我中华！真乃，是可忍孰不可忍！
=Defying us Chinese! This is really, [an extreme annoyance ! | intolerable, if this can be tolerated, then what cannot be tolerated?]

Frequent use of abbreviations and acronyms:

对于现在的ZF我还能说什么？？？
What can I say about the current [government | ZF] ???

然后24号刚刚在嘉定开盘的安亭新镇，听说看盘当天价格是从9500起，火爆了，均价也只在1w2左右。
Then Anting New Town, just opened in Jiading on the 24th, it was said that the prices right on the opening day were 9,500 RMB and upwards, it was hot, the average price was just around 12,000 RMB.

New words or randomly coined words:

但因温度过高，起不了止“汗 瀑”的作用。
But because the temperature is so high, they are no use in stopping the "sweat waterfall".

哎，这就是武汉，我都吃了好多年的地沟 油了
Sigh, this is Wuhan exactly, I have eaten [illegally recycled waste cooking oil | ditch oil] for many years

邓们的心思不可谓不毒 ！！
You can't say that the ideas of the leaders like Deng aren't poison !!

这个性价比是很高的，像这种花园叠墅，在哪里也没有这么便宜啊！
This is good value for the money -- like these kinds of garden – stacked villas, there is nothing this cheap anywhere else!

Colloquial style:

在中国人吃人的地方多的去了
In China, there are quite a lot of places where people cannibalize each other.

看来房价真的跌了哇！
It seems that housing prices really went down!

Typo caused by improper “Pinyin” input:

等青柴大人到啦，会给你升冤的。
Wait for the great =imperial commissioner to arrive, he'll =redress the injustice for you.

Frequent use of borrowed words:

首先本国大路上跑的75%都是大众，宝马，苯次之类的车。
First of all, 75% of the vehicles running on the highways of this country are Volkswagen, BMW, =Mercedes - Benz, and the like.

Non-standard use of punctuations:

如题，房价是真的跌了吗？？？
As in the title, did housing prices really go down???