Tagging Guidelines

For Chinese-English Word Alignment

Version 1.0 – 4/26/2009

Linguistic Data Consortium
Created by: Xuansong Li, xuansong@ldc.upenn.edu
With contributions from: Niyu Ge, niyuge@us.ibm.com
Stephanie Strassel, strassel@ldc.upenn.edu

Table of contents

1 Introduction

The goal of the tagging guidelines is to provide richer information via tagging translated-correct links and unaligned words. If the alignment task aims to map symmetric deep-structural semantic equivalence, the tagging task will describe asymmetric surface-structure differences leading to such semantic equivalence.

In this guidelines, we tag all the links, unmatched words, and all cases of Chinese DE 的. Blue color is used for an alignment link, and yellow color for word tagging.

2 Types of links

For links, we distinguish semantic (content word) links from function links. Pure semantic or function links are those which cannot be further decomposed because no extra words/components will be needed to build semantic translation equivalence. This type of link can be regarded as context-free links, the interpretation of which involves less contextual clues. The link is symmetric in structure because the source and target resemble each other in surface structure construction.

To handle the idiosyncratic features of the problematic Chinese 的, we designed three Chinese 的 links.

Grouping or attaching function words to their head words can show word dependency relations and contextual expressive function equivalence. With extra words attached, links can still be of a semantic or function nature. The attached words can be functional or contextual, providing grammatical or contextual/semantic clues. The missing words on the other side of the link then can be inferred grammatically or contextually. Therefore, according to how a missing word is inferred, we have grammatically inferred link and contextually inferred link. Stripping off extra unmatched words, we get pure semantic or function links.

Therefore, we have the following types of links:

Semantic links
Function links
DE-clause links
DE-modifier links
DE-possessive links
Grammatically inferred semantic links
Grammatically inferred function links
Contextually inferred links

2.1 Semantic links

Revealing direct equivalence between content words/phrases of source and translation, semantic links refer to links between content words. For different language, there are different standards for categorizing contents words. For English, content words are nouns, verbs, adjectives and adverbs. For Chinese, there are more categories than those in English. Here, we adopt the English standard. If the words on both sides of the link are content words, the aligned link will be a semantic one, otherwise it will be a function link.

敬请收看走遍中国特别节目 .
We respectfully invite you to watch a special edition of Across China.

中国→ China
特别→ special

百团大战纪念碑巍然屹立在太行山上，
Standing tall on Taihang Mountain is the Monument to the Hundred Regiments Offensive.
屹立→Standing tall

Idioms, set (frozen) expressions or proverbs, which are translated semantically as a whole minimum translation unit, belong to semantic links even if we may find function words inside the links. As long as there is set/frozen expression on one side of the link, we would regard the link as a semantic one.

胡锦涛指出实行改革开放很重要。
Hu Jintao point out that it is important to have open-up policy.

我可以弹奏几种乐器，比如吹笛子，弹吉他。
I can play quite a few musical instruments, for instance, the flute and the guitar.

2.2 Function links

Function links involve two cases. One case is that function words appear on both sides of the link. The other case involves function words only on one side of the links, either in source or translation. In other words, as long as there is a word that is not within the categories of nouns, verbs, adjectives or adverbs appearing on source or translation of the link, the link will be a function type. The function links describe how a function word is translated into another language, revealing function equivalence between two languages. Punctuations are covered in this category.

它由主碑，副碑，一座大型圆雕和烽火台，长城等组成。
It is composed of a primary stele, secondary steles, a huge round sculpture and beacon tower, and the Great Wall, among other things.
和→and

它由主碑，副碑，一座大型圆雕和烽火台，长城等组成。
It is composed of a primary stele, secondary steles, a huge round sculpture and beacon tower, and the Great Wall, among other things.
等→among other things

百团大战是八路军在抗战期间发动的规模最大的一次战役 。
The Hundred Regiments Offensive was the campaign of the largest scale launched by the Eighth Route Army during the War of Resistance against Japan.
八→Eighth

Punctuations involve two cases: both sides of the links are punctuations; one side of links (source or translation) is punctuation.

这场战役打破了日军对敌后根据地的封锁，振奋全国人民抗日精神，影响了世界人民反法西斯战争的形势。
This campaign broke through the Japanese army's blockade to reach base areas behind enemy lines, stirring up anti - Japanese spirit throughout the nation and influencing the situation of the anti - fascist war of the people worldwide.
，→ and
。→ .

2.3 DE-clause links

对于没有经历过战争的人来说，战争是生动的故事，对于经历过战争的人来说，战争是人生的历程，是生命的一部分。
To those who have never experienced wars, wars are vivid stories. To those who have experienced wars, war is the course of life, part of life.

可是让他们没有想到的是，自主创新的产品生产出来以后，市场并不像他们所想像的那么好。
But what they had not expected was that when the independently invented products were produced, the market did not turn out to be as good as they had imagined.

一九九四年，研究者们开发出了完全具有自主知识产权的华中一型高技能数控系统，
In 1994, researchers developed the Huazhong Model 1 high - functioning numerical control system whose intellectual copyright was completely and independently owned by them.

2.4 DE-modifier links

Chinese 的 is useful and frequent in a modifying relationship between words or constituents. This function word can be translated differently. When it does not assume the clause or possessive function during translation, it is then regarded as a DE-modifier. In translation practice, it is normally translated into prepositions in English.

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.

2.5 DE-possessive links

的 in DE-possessive links assume the possession function in a modifying relationship.

我们今天请到演播室的嘉宾是中国政法大学的蔡定剑教授。
The guest we invited to our studio today is Professor Cai Dingjian at China University of Political and Law.

日军这一阴谋引起八路军总部朱德，彭德怀的高度警惕。
This plot of the Japanese army drew great attention from Zhu De and Peng Dehuai of Eighth Route Army headquarters.

2.6 Grammatically inferred semantic links

In grammatically-inferred semantic links, stripping off extra words or divergent expressing words, we get terminal semantic links. In this type of link, the extra words will be further tagged according their grammatical/expressive function.

正因为这样，国外呢都把高级数控机床当做是战略物资，对我国长期封锁。
This is why advanced CNC machine tools are seen as strategic materials overseas, uh, and for a long time have been kept away from our country.
把…机床→ machine tools (This is a grammatically-inferred semantic link, and 把 will be further tagged inside this link)

随后华中高层决定将这项成果变成产品推向市场。
Then the Huazhong leadership decided to turn this success into a product to put on the market.
将…成果→ success This is a grammatically-inferred semantic link, and 将 will be further tagged inside this link)

2.7 Grammatically inferred function links

In grammatically-inferred function links, stripping off extra words or divergent expressing words, we get terminal function links. In this type of link, the extra words will be further tagged according their grammatical/expressive function.

嗯，卫生员是相当忙的。
Eh, medics were quite busy.
的 。→ . (This is a grammatically-inferred function link, and 的 will be further tagged inside this link)

就是以铁路为一个柱子，以公路为一个链子，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 railways as a pillar, roads as a chain, and strongholds as a lock, to carry out siege warfare
一个--> a (This is a grammatically-inferred function link, and 个 will be further tagged inside this link)

中国大地上的二战标志之太行永志。
WW II Landmarks on the Great Earth of China: Eternal Memories of Taihang Mountain
上→on (This is a grammatically-inferred function link, and 的 will be further tagged inside this link)

2.8 Contextually inferred link

In contextually inferred links, the extra words added to one side of the link are obligatory for the context. Without them, the grammatical structure might be acceptable, but it is not semantically sensible. Such extra words are triggered by its co-occurring (associative) word or collocation words, that is, they have lexical clues or co-occurring words/collocation partners on the surface structure at a locally phrase level or sentence level. The missed words on the other side of the link will be accordingly implied via word association or collocation. If a missed word is inferred pragmatically or contextually above the sentence level, they will be handled as independent unaligned words.

大家好！欢迎收看这频道。
Hello, everyone! Welcome to this channel.
欢迎收看 → Welcome (this is a contextually inferred link, and 收看 will be further tagged in this link)

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.
会见活动→ meeting (this is a contextually inferred link, and 活动 will be further tagged in this link)

他对会议的圆满举行感到高兴。
He was glad at the success of the meeting. (举行 will be further tagged)

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 will be further tagged)

新华社华盛顿 4 月 20 日电(记者应谦)
Xinhua News Agency, Washington, DC, April 20, by wire (reporter Qian Ying) (DC will be further tagged)

八路军　击队异常活跃给华北地区的日军制造了越来越多的麻烦。
The Eighth Route Army guerrillas were extraordinarily active, creating more and more trouble for the Japanese army in North China. (地区 will be further tagged)

Exceptions:

When a single character word is contextually attached to another single character word, even if they are of the contextual type, they show strong coherence between each other, and annotators feel very hard to tear them apart, as in the following cases, thus we treat two-character contextual word as a unit of real translation, that is, a semantic link, rather than a contextually inferred link, and no further tagging is needed for such two-character words.

我们欢迎外商投资。
We welcome foreign investment.

他会见了几位法务代表。(this is a semantic link)
He met with several judicial representatives.

But:

他会见了几位外务大臣。(this is a semantic link)
He met with several foreign affair ministers.

3 Types of tags

The tags used for labeling extra words are classified according to their expressive functions which help to build semantic equivalence. Tagging links and unmatched words lead to translation rules. The translation rules from the Chinese-English tagging task are derived on contrastive features between Chinese and English, that is, how a function in source is realized in target or vice versa. The rules are no-directional. Tagging unmatched words also reveal cohesion and coherence features at local or phrase (within a sentence) level.

3.1 Tagging unmatched words inside links

In this guideline, all unmatched words are tagged. If the unmatched words have local dependency words to rely on, then we attach them to those dependency words. These extra words can be either grammatically inferred or contextually inferred.

In grammatically inferred links, the format can be one of the following two:

attached word/words + content word/words
attached word/words + function word/words

Tagging focus is on the attached word/words before + (plus) sign. The extras can be attached to any side of the link, so the tagging will be done on both source and target. Classification of tags is based on expressive functions. Chinese 的 is specially treated here with one special classification to address its "modifying" function, while other functions of 的 are tagged in "tense/passive", "omni-function-preposition", and "possessive" categories.

3.1.1 Tense/passive-marker

对于没有经历过战争的人来说，战争是生动的故事，对于经历过战争的人来说，战争是人生的历程，是生命的一部分。
To those who have never experienced wars, wars are vivid stories. To those who have experienced wars, war is the course of life, part of life. (过and have are tagged as "tense/passive marker")

他正看书。
He is reading a book. (正 and "is" are tagged as "tense/passive marker")

她急着想要知道一个答案，自己为什么会被拒绝 。
She was anxious to find out why she was turned down. (被 and "was" are tagged as "tense/passive marker") (note: here the object of 被 is omitted, and we attach it to verbs to show passive voice. Otherwise it will be attached to its objects — see section 4.2)

Chinese-DE can also be served as tense-passive marker

他看了提交的报告。(word order change is tagged here with "的")
He went through the report submitted. (的 is tagged as "tense/passive marker")

这是目前国内第一台使用国产高端数控技术制造的数控机床。
At present, this is the first CNC machine tool manufactured using domestically produced top - end numerical control technology. (的 is tagged as "tense/passive marker")

百团大战是八路军在抗战期间发动的规模最大的一次战役。
The Hundred Regiments Offensive was the campaign of the largest scale launched by the Eighth Route Army during the War of Resistance against Japan. (tag "的" as "tense/passive" marker)

3.1.2 Omni-function-preposition marker

Preposition can trigger various kinds of expressive function structures, such as location, time, purpose, possessive, etc. Due to this multiple-function feature, we use an omni-function-preposition marker for all preposition-related expressive functions with the only exception of "possessive" case, which has a special "possessive" marker. Some frequent Chinese prepositions are also treated here, such as (把, 将, …etc). Word order change is closely related to this type of tags.

他们分为 5 个小组,共约见了约 150 位议员或议员助手以及政府官员,用他们在中国投资办企业的经验, 阐述延长给中国贸易最惠国待遇的必要性。
They split into five groups and altogether visited about 150 members of congress, congressional, and government officials, using their experience of investing in managing enterprises in China to explain the necessity of extending giving China most - favored - nation trading status. (tag "in" as "omni-function-preposition")

中共中央总书记国家主席胡锦涛和越共中央总书记农德孟，共同会见中越两国青年代表.,
General Secretary of the CPC Central Committee and Chinese President Hu Jintao and General Secretary of the Communist Party of Vietnam (CPV) Nong Duc Manh together met with youth representatives from the two countries. (tag "with" as "omni-function-preposition")

正因为这样，国外呢都把高级数控机床当做是战略物资，对我国长期封锁。
This is why advanced CNC machine tools are seen as strategic materials overseas, uh, and for a long time have been kept away from our country. (tag 把 as "omni-function-preposition", relating to order change)

随后华中高层决定将这项成果变成产品推向市场。
Then the Huazhong leadership decided to turn this success into a product to put on the market. (tag "将" as "omni-function-preposition")

不到一分钟就可以整个把这个曲面把它加工出来。
In less than a minute we should be able to machine the entire curved surface. (tag "把" as "omni-function-preposition", relating to word order change, tag "the" as "determiner/demonstrative")

四零年在在百船大战前呢，中国国内这个妥协啊，啊投降啊这个倾向危机呢是空前严重的，就是在正，正面的这个((境况))下。
Before the Hundred Regiments Offensive in 1940, an inclination to compromise, ah, surrender, was an extremely serious crisis in the frontline ((situation)) in China.

四 零 年→ in 1940 (tag "in" as "omni-function-preposition")

怀揣着梦想，梦想却被别人粉碎 。
She had dream in her heart but someone shattered her dream. (tag "被" as "omni-function-preposition")

3.1.3 DE-modifier marker

的 is used to modify an adjective or noun, 地 is used to modify verbs or adverbs, 得 is used to modify verbs while introducing complimentary result. When 的 is used to modify nouns, word order change may be involved.

八路军击队异常活跃给华北地区的日军制造了越来越多的麻烦。
The Eighth Route Army guerrillas were extraordinarily active, creating more and more trouble for the Japanese army in North China.
(的 is tagged as "DE-modifier")

而且依托这个囚笼呢啊来进一步地加强对根据地的进攻。
In addition, it relied on this cage , ah , to further strengthen its assaults against the base areas . (地 is tagged as "DE-modifier")

而且依托这个囚笼呢啊来进一步地加强对根据地的进攻。
In addition, it relied on this cage, ah, to further strengthen its assaults against the base areas. ("的"is tagged as "DE-modifier")

他干得好。
He did well. (得 is tagged as "DE-modifier")

3.1.4 Possessive marker

Both source and target can have possessive markers. Possessive markers may trigger word order change during translation.

美国商会广东分会会长康永华律师说,克林顿政府已经表示要延长中国的贸易最惠国待遇,因此,这次说的重点是那些较保守的议员。
The head of the Guangdong branch of the American Chamber of Commerce, the attorney Yonghua Kang, said the Clinton administration had already demonstrated that it wanted to extend China's most - favored - nation trading status and because of this, the focus of this round of lobbying was those members of congress who are more conservative. ("of" is tagged as "possessive" order-change)

而且依托这个囚笼呢啊来进一步地加强对根据地的进攻。
In addition, it relied on this cage, ah, to further strengthen its assaults against the base areas.
进 攻→ its assaults (its is tagged as "possessive")

3.1.5 To-infinitive marker

To-infinitive can be used to express various functions in English. In alignment, it may be in a terminal function link, like:

为了完成项目，他连夜工作。
To finish the project, he worked through the night.

When an infinitive has no word to match, it will be attached to verbs and tagged as “to-infinitive” marker.

四零年在在百船大战前呢，中国国内这个妥协啊，啊投降啊这个倾向危机呢是空前严重的，就是在正，正面的这个((境况))下。
Before the Hundred Regiments Offensive in 1940, an inclination to compromise, ah, surrender, was an extremely serious crisis in the frontline ((situation)) in China
妥 协 → to compromise (to is tagged as “infinitive”)

第二次世界大战二十万中国远征军出国作战，
During World War II, the two hundred thousand strong Chinese Expeditionary Force went abroad to fight.
作 战→ to fight (to is tagged as “infinitive”)

3.1.6 Sentence marker

是吧，路也给弄路断了，他汽车也不能来来来回了。
Yeah, the roads had been cut off, with its vehicles unable to drive back and forth.
了 。→ . (了 is tagged as “sentence marker”)

这样一个作战,这个他的目的，目的目标是什么呢？
What was the purpose, purpose and goal of this campaign?
呢 ？ -- ? (呢is tagged as “sentence marker”)

嗯，卫生员是相当忙的。
Eh, medics were quite busy.
的 。→ . (的is tagged as “sentence marker)

就像这种五轴技术的话，就说的国外很多是不卖给咱们的。
For example, it was said that a lot of those overseas won't sell this kind of five - axis technology to us. (的is tagged as “sentence marker)

我只需要编一条程序，它就自动地就可以走出这个弧线来了。
I just need to write a program, then it can automatically produce this curve. (了 is tagged as “sentence marker)

高技术含量的，高精度的，他们一般都不卖给咱们的，或者是落后的东西卖给你，可能是。

High - tech, high - precision, generally none of them will sell it to us, or probably sell you some obsolete stuff. (的is tagged as "sentence marker)

There are cases no alignment can be find when these words are used as clause markers in the middle of a sentence, thus no attachment could be done, we only tag them as "sentence marker".

高技术含量的，高精度的，他们一般都不卖给咱们的，落后的东西卖给你，可能是。
High - tech, high - precision, generally none of them will sell it to us and sell you some obsolete stuff. (的is tagged as "sentence marker)

3.1.7 Measure-word marker

这个采用差别化氨伦丝生产技术改造的项目,总投资七千万元,累计年产氨纶丝一千五百吨。
This project, which has remodeled by adopting a differentiated urethane elastic fiber production technology, at a total investment of 70 million yuan, has an annual accumulated total urethane elastic fiber output of 1500 tons. (tag "个" as "measure-word")

就是以铁路为一个柱子，以公路为一个链子，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 railways as a pillar, roads as a chain, and strongholds as a lock, to carry out siege warfare
一个--> a (个 is tagged as "measure words")

啊过去人工打磨啊，比方像五叶就五片叶子的那个螺旋桨，船用螺旋桨，五个工人，就是一个就是咱们按八个小时计算的，要打大概十五天到二十天才能做一个浆，
Ah , in the past , ah , finishing them by hand -- for example , a five - blade -- this five - blade propeller , a boat propeller , five workers , that is , one -- we take it as eight hours , needs to work for about 15 - 20 days to do just one propeller . (个 is tagged as "measure words")

With temporal nouns "Year", and "Date' expressing time like 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. (this is a contextually inferred link, and 年 will be further tagged in this link)

他六月三日来到上海。
He came to Shanghai on 3rd of June. (this is a contextually inferred link, and 日 will be further tagged in this link)

他五点三十分到达北京。
He arrive in Beijing at 5:30.

Note: exception case. Units that are present on both sides are aligned to each other and treated as terminal links. E.g.

这 个 采 用 差 别 化 氨 伦 丝 生 产 技 术 改 造 的 项 目, 总 投 资 七 千 万 元, 累 计 年 产 氨 纶 丝 一 千 五 百 吨 。
This project, which has remodeled by adopting a differentiated urethane elastic fiber production technology, at a total investment of 70 millian yuan, has an annual accumulated total urethane elastic fiber output of 1500 tons.

3.1.8 Determiner/demonstrative marker

In both source and target, if there is no direct match for function words of "the, a, an, this, that", they will be tagged as determiner/demonstrative marker. They are all used to restrict or modify a noun.

破击正太铁路是这场战役的主要目标。
the main purpose of the campaign was to sabotage the Zhengtai Railway .
目 标→ the purpose (the is tagged as "determiner/demonstrative")

在这条铁路线上，有天险娘子关和日军在华北的重要燃料基地阳泉，井陉煤矿。
Along this railway line was the Niangziguan Pass, a natural barrier, and the Jingjing coalmine, a key fuel base in Yangquan for the Japanese military in North China.
基 地→a base (a is tagged as "determiner/demonstrative")

E.g. 他是第一个提出此议案的人。
He is the first to propose such a motion. (the is tagged as "determiner/demonstrative")

3.1.9 Clause marker

除太原与石家庄外，日军以第四，第八，第九三个混成旅团，分布在铁路沿线的五十个据点里。
Outside of Taiyuan and Shijiazhuang, the Japanese military deployed the 4th, 8th, and 93rd mixed brigades and regiments, which were scattered among 50 strongholds along the railway line
团→ regiments which (which is tagged as "clause marker")

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.

3.1.10 Anaphoric-reference marker

When a pronoun is dropped from one side of the link, the extra pronoun in the other side of the link will be tagged as anaphoric-reference marker. For such markers, they usually have lexical clues within the sentence level. If the clue is beyond the sentence level, the extra pronoun will be handled in empty links.

美国商会广东分会会长康永华律师说, 克林顿政府已经表示要延长中国的贸易最惠国待遇, 因此, 这次 说的重点是那些较保守的议员。
The head of the Guangdong branch of the American Chamber of Commerce , the attorney Yonghua Kang , said the Clinton administration had already demonstrated that it wanted to extend China 's most - favored - nation trading status and because of this, the focus of this round of lobbying was those members of congress who are more conservative. (tag "the" as determiner/demonstrative maker and "it" as anaphoric marker)

数控机床相信很多人啊见都没有见过，可是我们日常生活里却几乎离不了它。
I'm sure many people, ah, haven't even seen CNC machine tools, but in our daily lives we can't get away from them. ("we" is tagged as anaphoric marker)

3.1.11 Rhetorical marker

For rhetorical purpose, some words are omitted on one side of the alignment, while the omitted are restored on the other side. This difference in rhetorical expression is obvious when the same noun, verb, adjective, etc. are shared. We tag the extra words for this purpose as "rhetorical marker".

E.g. 内地的专家和台湾的专家 (share the same noun. Here the 2nd occurrence is tagged as 'rhetorical')
the experts from the mainland and Taiwan

3.1.12 Local context marker

We use "local context marker" for unmatched words in contextually inferred links. These extra words are triggered by word collocation or association in practical use. Contextual marker is only for lexical/phrase/sentence level inferred words. Words inferred above the sentence level will be attached to any word, they are directly tagged as context obligatory maker or context non-obligatory marker.

大家好！欢迎收看这频道。
Hello, everyone! Welcome to this channel. (tag 收看 as contextual marker)

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. (tag "活动" as contextual marker)

当地时间十五点三十分左右，胡锦涛和农德孟一起来到会见大厅，看望参加活动的二百位代表 。
Around 15:30 local time, Hu Jintao and Nong Duc Manh came to the meeting hall together to visit the 200 delegates participating in the activities. (tag "参 加" as contextual marker)

借此机会，我要向第六届中越青年友好会见活动的圆满举行表示热烈的祝贺 。

I wish to take this opportunity to warmly congratulate the success of the sixth friendship meeting of the Sino - Vietnam youth. (tag "举 行" as contextual marker)

中越两国的未来和希望，寄托在青年身上，中越友好事业的未来和希望，也寄托在青年身上，中国党和政府一贯重视和支持两国青年交流。.
The future and hope of China and Vietnam are in the hands of the young people; the future and hope of the Sino - Vietnamese friendship are also in the hands of the young people. The Chinese party and government have always attached importance to and support the exchanges between the young people of the two countries. (tag "事 业" as contextual marker)

3.2 Tagging independent unaligned words

There are two types of markers for indepent unaligned words. One type is context obligatory markers, which are grammatically needed and semantically inferred words. The other type is context non-obligatory markers, which are used for both grammatically and semantically unnecessary extra words.

3.2.1 Context obligatory marker

The context obligatory marker is used for the following cases:

Case A: Grammatically required words having no convenient dependency constituents to attach, such as copular BE, it-sentence, "that" objective clause marker, etc.

就是说在这样一个国际，国内严峻的形势下，这个中国共产党领导这个八路军啊，就开始那么就啊战略性的进攻战役，叫百团大战。
So, it was amidst such a grave international and domestic situation that the Eighth Route Army led by the Chinese Communist Party, ah, launched, ah, a strategic offensive called the Hundred Regiments Offensive.

那天黑夜里，人们都睡了，村里头人欢马叫，呀人们都起来，是不是有情况(())了?
On that dark night, everyone was sleeping when human voices and neighing horses were heard within the village. People all got up. Did something (()) happen?

设备从安装到投产只用了三个月时间,开发了企业自己的专利技术,为公司下一步对外输出氨伦丝生产技术奠定了基础。
It took only a three - month period from equipment installation to production startup. The company has developed its own patented technology, laying the foundation for its next move into the exportation of urethane elastic fiber production technology.

这个采用差别化氨伦丝生产技术改造的项目,总投资七千万元,累计年产氨纶丝一千五百吨。
This project, which has remodeled by adopting a differentiated urethane elastic fiber production technology, at a total investment of 70

million yuan, has an annual accumulated total urethane elastic fiber output of 1500 tons.

你们不要讲怕，
Don't say you are scared.

Case B: Anaphora (pronoun) reference beyond sentence level

Case C: Subject-drop

数控机床相信很多人啊见都没有见过，可是我们日常生活里却几乎离不了它。
I ' m sure many people , ah , haven ' t even seen CNC machine tools , but in our daily lives we can ' t get away from them .
I (no lexical or grammatical clue)

怎么又停下来了？
Why have you stopped again?

这个呢就是刚才说的这个程序。
This, uh, is the program I was just talking about.
I --pronoun-drop context drop

像制造这样的大型螺旋桨机件，只能依靠人工打磨等土办法。
For instance, manufacturing these large propeller parts, we have to use backward techniques like finishing them by hand.
we -- infer (knowledge)

他将用一台普通车床和数控车床分别加工一个球体来进行现场比较。
He would use a normal machine tool and a CNC machine tool and make a sphere with each that we could make a comparison on the site.

不好掌握，凭经验。
It's not easy to control, you do it by experience.

Case D: Conjunction (phrase/clause/discourse) drop

这次新投产的生产线,由该公司自行设计,自行开发,自行调试。
The production line newly put into operation this time was self - designed, self - developed, and self - debugged by the company.

Case E: obligatory punctuations and hyphens

Professional use of punctuations and hyhens is grammatically required, such as comma before “and”, “etc”, and hyphen used to join words which are used as a single grammatical unit, thus they are obligatory. By default, all punctuations and hyphens are obligatory except they appear as a typo.

这次新投产的生产线，由该公司自行设计，自行开发，自行调试。
The production line newly put into operation this time was self - designed, self - developed, and self - debugged by the company.

美国商会广东分会会长康永华律师说,克林顿政府已经表示要延长中国的贸易最惠国待遇,因此, 这次 说的重点是那些较保守的议员。
The head of the Guangdong branch of the American Chamber of Commerce, the attorney Yonghua Kang, said the Clinton administration had already demonstrated that it wanted to extend China's most - favored - nation trading status and because of this, the focus of this round of lobbying was those members of congress who are more conservative. (Hyphens here can be tagged non-obligatory because the words around it can standalone)

他们分为 5 个小组,共约见了约 150 位议员或议员助手以及政府官员,用他们在中国投资办企业的经验,阐述延长给中国贸易最惠国待遇的必要性。
They split into five groups and altogether visited about 150 members of congress, congressional aides, and government officials, using their experience of investing in and managing enterprises in China to explain the necessity of extending giving China most - favored - nation trading status.
(",", before "and" here is tagged as "context obligatory") (priority: word link first)

马英九
Ma Ying - jeou ("-" is here tagged as "context obligatory")

而且依托这个囚笼呢啊来进一步地加强对根据地的进攻。
In addition, it relied on this cage, ah, to further strengthen its assaults against the base areas.

国务院副总理邹家华参加了今天举行的投产剪彩仪式。
The Vice Premier of the State Council, Jiahua Zou, attended the ribbon cutting ceremony held today for the start of operations.

3.2.2 Context non-obligatory marker

Some extras are not needed grammatically/semantically or they simply carry style or modal features. For those extras sharing the same head word, they are omitted or added simply for rhetorical use, and thus they also belong to the non-obligatory category.

从小小的手机到刚刚上路的高速列车，没有数控机床这些东西呢就只能停留在设计图上。
From the small cell phone to the high - speed train just put on the rails; without CNC machine tools, these things, uh, could only remain designs on paper.

至于航空航天用的那些高精尖设备就更得靠数控机床来加工。
And the high - precision top - notch equipment used in aviation and space flight is even more reliant on CNC machine tools for their manufacture.

在就业的路上，她又该何去何从。
What course should she follow in her career.

呃在求职，都已经和总公司签了合同了，
Er, she is looking for a job, and she had already signed an employment contract with the head office.

因为之前我们也没有，真的没有想去伤害她。
Because we did not, truly did not want to hurt her before.

四零年在在百船大战前呢，中国国内这个妥协啊，啊投降啊这个倾向危机呢是空前严重的，就是在正，正面的这个((境况))下。
Before the Hundred Regiments Offensive in 1940, an inclination to compromise, ah, surrender, was an extremely serious crisis in the frontline ((situation)) in China

那么敌后战场呢，日本为了要夺取，巩固它这个占领区，它开始那个新的战略。
Well, on the battlefield behind enemy lines, in order to take over, consolidate the area under its occupation, Japan began a new strategy.

Chinese "所" is an auxillary, and is treated as context non-obligatory since it can be omitted in most cases.

The person I know is him.
我所认识的是他。(context non-obligatory)

What I understand is to go.
我所理解的是要去。(context non-obligatory)

All I understand is that I should go.
我所理解的是要去。(non-obligatory)

4 Conflicting rules

Whenever there is a rule conflict, we'll prioritize the alignment or tagging for consistency purpose.

a. Terminal link and attachment rules

Link rules always have the priority over attachment rules.

b. Conjunction rule and punctuation rule

Word link has the priority over punctuations.

他们分为 5 个小组,共约见了约 150 位议员或议员助手以及政府官员,用他们在中国投资办企业的经验,阐述延长给中国贸易最惠国待遇的必要性。
They split into five groups and altogether visited about 150 members of congress, congressional aides, and government officials, using their experience of investing in and managing enterprises in China to explain

the necessity of extending giving China most - favored - nation trading status.
以及 <--> and (",” before “and” here is tagged as “context obligatory”)
(priority: word link first)

c. Passive voice conflicting rules

First priority: link to link

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

Second priority: BEI is attached to its object

怀揣着梦想，梦想却被别人粉碎 。
She had dream in her heart but someone shattered her dream. (tag “被” “omni-function-preposition”)

Least priority: BEI is attached to verb (as tense/passive marker)

E.g. 桥被冲垮了。
The bridge was destroyed.

d. 的 sentence-marker and other 的-related tags

First priority: when 的 is at the end of a sentence or a clause.

E.g. 太阳红彤彤的。
The sun is red.

5 Summary

5.1 A summary of link types and word tagging types

The following is summary of types and links with their abbreviations used for tagging task.

Types of links (8 types):

Semantic links
Function links
DE-clause links
DE-modifier links
DE-possessive links
Grammatically inferred semantic links
Grammatically inferred function links
Contextually inferred links

Tagging types (14 types)

Tense/passive marker
Omni-funciton-preposition marker
DE-modifier marker
Possessive marker
To-infinitive marker
Sentence marker
Measure-word marker
Determiner/demonstrative marker
Clause marker
Anaphoric reference marker
Local context marker
Rhetorical marker
Context obligatory marker
Context non-obligatory marker

5.2 How 的 is handled

In link tagging, there are three kinds of 的 links: DE-modifier, DE-clause, DE-possessive. For unmatched DE, we tag it using tense/passive marker, DE-modifier markers (when modifying nouns, adjectives, adverbs), possessive marker, and sentence marker.

5.3 How inference is handled

The missed words or meaning can be inferred from context. For missed function words, they can be inferred either locally (within the sentence level) or at a discourse level. The missed function words are required because without them, the source or translation will be grammatically unacceptable. The missed function words from the local level can be inferred from the lexical clues within a sentence, either from their dependency constituents or related lexical clues within sentences. This type of inference is grammatical inference. In the guideline, the unmatched words inside grammatically inferred links are of this type. The other type of words is contextually missed words due to word association/collocation (or pragmatic feature). The missed meaning or words can be inferred via collocation/association context. In this guideline, the unmatched words inside local context links are of this kind. If a missed function or semantic word is beyond the sentence level (that is, if there is no lexical clue to attach within the sentence level), it will be inferred at a discourse level. In the guidelines, the context obligatory unaligned words are of this type.

5.4 How to distinguish possessive-preposition and modifying-preposition

Prepositions can assume various expressive functions, including possession and modifying functions.

Possessive means:

Person or organization(country) owns(belong to) concrete things
Person or organization(country) owns (belong to) persons
Person or organization(country) owns (belong to) organizations
Person or organization(county) owns (belong to) abstract concepts (his hope, he owns the hope)
Person or organization(country) owns (having the behavior) (his doing)

而且依托这个囚笼呢啊来进一步地加强对根据地的进攻。
In addition, it relied on this cage, ah, to further strengthen its assaults against the base areas.
进 攻→ its assaults (its is tagged as "possessive")
这是我的书。
This is my book. (的 is tagged as "possessive marker")

啊这个呃断绝他们的互相联系，
ah, so as, er, to cut off their communication with one another. (的 is tagged as "possessive marker")

你们不要讲怕，我们的首长就要来了，跟你们说你们回到黑，安心睡觉。
Don't say you are scared. Our commander is coming. I am telling you to go back to % pw sleep peacefully. (的 is tagged as "possessive marker")

All of the following cases involve possession:

China's population 中国的人口
Its population 它的人口
Population of China 中国的产品
China's products 中国的产品
Products from China 中国的产品
Products of China 中国的产品
A professor at Upenn 宾大的教授
A professor from Upenn 宾大的教授， 来自宾大的教授
A professor of Upenn 宾大的教授

The following cases are modifying cases:

这场战役打破了日军对敌后根据地的封锁，振奋全国人民抗日精神，影响了世界人民反法西斯战争的形势。
This campaign broke through the Japanese army's blockade to reach base areas behind enemy lines, stirring up anti - Japanese spirit throughout the nation and influencing the situation of the anti - fascist war of the people worldwide.

借此机会， 我要向第六届中越青年友好会见活动的圆满举行表示热烈的祝贺。向中越两国青年 致以诚挚的问候.

I wish to take this opportunity to warmly congratulate the success of the sixth friendship meeting of the Sino - Vietnam youth.

6 Acknowledgement

Thanks to Rich Schwrtz, Daniel Marcu, Dekai Wu and other members of word alignment discussion group for immediate feedbacks on the guidelines during its compilation.