文章编号：1003-0077(2009)03-0065-023

汉英词语对齐规范

赵红梅[1]，刘 群[1]，张瑞强[2]，吕雅娟[1]，隅田英一郎[2]，吴翠玲[2]

（1. 中国科学院 计算技术研究所 智能信息处理重点实验室，中国 北京 100190；
2. 日本情报通信研究机构—国际电气通信基础技术研究所，日本 京都 619-0288）

摘 要：该文介绍了一个新的汉英词语对齐规范。该规范以现有的 LDC 汉英词语对齐规范为基础，对其进行了较大的改进和扩展，特别是提出了一种全新的对齐标注方法 —— 将词语对齐区分为真对齐和伪对齐，真对齐又分为强对齐和弱对齐。这种细化的标注方法能够更好地刻画词语对齐的特点。该规范已经实际应用于大规模的人工词语对齐标注中。我们对对齐标注的一致性进行了评价。结果表明，在该规范的指导下，标注者内部和标注者间的对齐都取得了比较理想的一致性，两组强、弱、伪三种对齐的 Kappa 值分别为 0.99、0.98、0.93 和 0.96、0.83、0.68。最后，一个简单的实验初步证实了该规范在统计机器翻译中的有效性。

关键词：人工智能；机器翻译；汉英词语对齐规范；手工词语对齐；真对齐；伪对齐；强对齐；弱对齐；对齐和标注一致性

中图分类号：TP391 **文献标识码**：A

A Guideline for Chinese-English Word Alignment

ZHAO Hongmei[1], LIU Qun[1], ZHANG Ruiqiang[2], LV Yajuan[1], Eiichiro SUMITA[2], Chooi-Ling GOH[2]

(1. Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100190, China;
2. NICT-ATR, Kyoto, Japan 619-0288)

Abstract: This paper presents a new guideline for Chinese-English word alignment. Starting from the existing Guidelines for Chinese-English Word Alignment (Linguistic Data Consortium, 2006), we propose a completely different classification for word alignment annotation: genuine link (involving strong link and weak link) and pseudo link. This explicit distinction can represent the characteristic of cross-lingual word alignment. The proposedguideline has been successfully applied in a large-scale task for Chinese-English Word alignment, achieving good intra- and inter-annotator agreemenst at the Kappa coefficients of 0.99、0.98、0.93 and 0.96、0.83、0.68 for the strong link, weak link and pseudo link respectively. And a further experiment proves that such annotated word alignment is useful for SMT system.

Key words: artificial intelligence; machine translation; annotation guidelines for Chinese-English word alignment; manual word alignment; genuine link; pseudo link; strong link; weak link; alignment and annotation agreement

1 引言

1.1 词语对齐引起的机器翻译问题

在统计机器翻译的训练过程中，基本上都采用 GIZA++[1] 自动对齐工具进行不同语种间的词语对齐，这种做法导致翻译结果经常出现如下问题：如汉英翻译将“我国”统统译为“China”，这是一种指代混淆的问题，究其原因，主要是由于自动对齐只有一种对齐方式，没有将确定性的对齐和不确定性的对齐区分开。后来很多统计机器翻译研究开始转向词

收稿日期：2008-11-24 **定稿日期**：2009-01-20

作者简介：赵红梅(1968—)，女，学士，高级工程师，主要研究方向为机器翻译；刘群(1966—)，男，博士，研究员，主要研究方向为机器翻译和自然语言处理；张瑞强(1967—)，男，博士，高级研究员，主要研究方向为网络搜索技术、语音翻译和自然语言理解技术。

语的手工对齐，将手工对齐的语料视为黄金标准语料，也形成了一些比较成型的手工词语对齐规范。

1.2 几个比较知名的手工词语对齐规范

1) Blinker项目标注规范1.0.4版[2]：它是Dan Melamed于1998年2月为宾州大学Blinker项目建立的，以《圣经》的英文和法文版作为对齐语料，这个规范中几乎所有的对齐规则和方法都被后继的LDC的GALE汉英词语对齐规范所继承，包括我们后面要谈到的粘合的方法，规范中也出现了我们所采纳的全连线的对齐方式。

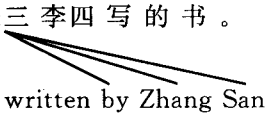
2) ARCADE词语对齐标记规范1.0版[3]：它是Jean Véronis于1998年4月为ARCADE项目词语系列(word track)的开发而建立的，由于对齐的性质特殊(仅限于对齐句子中一个指定的词或词组)，它在Blinker规范的基础上进行了不少修改，它要求标记者针对对齐本身(正常翻译、拼写错误、非平行的联合)、对齐的可信程度(不满意到满意四个维度)或未对齐的种类(指代表示、改写或解释性的翻译)使用不同的标签，还可以添加评论。在规范中也使用了粘合的方法。它的一个特点是很多部分都不要求对齐(包括各种限定词和不在固定搭配中的动词后的介词等)，值得注意的是，指代表示在“不能对齐但需要标记”之列。

3) LDC的GALE汉英词语对齐规范1.0版(以下简称LDC规范)[4]：是LDC于2006年9月参考了上面两个规范建立的，该规范内容比较全面，要求标注者在对齐的同时标记“被翻译”、“未被翻译”、“正确”和“不正确”标签，规范正式提出了粘合对齐的方式(glue approach)。2008年10月该规范有了3.0版[5]，较之第一版新版本变化不多，在变化的内容中突出了最小匹配原则，比如“今年(this year)”要求将“今”对齐“this”，而不是“今年”对齐“this year”。

以上几个规范普遍[2-4]存在如下问题，这些问题会产生一些不太合理的词组：

1) 都采用了粘合的对齐方式，但是没有区分粘合和被粘合的部分：

如：张三 李四 写 的 书 。

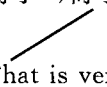


books written by Zhang San and Li Si

抽提出来的词组是：张三 by Zhang San

2) 除了ARCADE规范外[3]，其他规范都允许将指示代词对齐到被指代物：

如：我 买 了 张 椅子 ，椅子 很 贵 。



I bought a chair. That is very expensive.

抽提出来的词组是：椅子 that

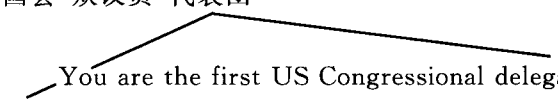
这些问题的产生都源于没有区分不同的对齐方式，而对不同的对齐加以区分正是我们提出的新规范的特点。

1.3 我们建立规范的思路

手工词语对齐结果语料要真正成为黄金标准语料，笔者认为有两点必须得到保证：

1) 有效性，即对齐结果真正能更好地为后继的机器翻译流程所用，能为更佳的翻译质量提供最大的帮助。比如，LDC规范[4]中有如下的对齐方式：

如：你们 是 我 今年 会见 的 第一 个 美国 国会 众议员 代表团



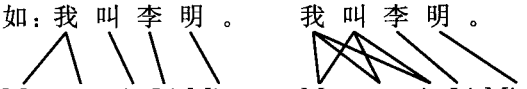
You are the first US Congressional delegation that I have met this year

这样势必抽出词组：代表团〈〉delegation that 。

这类词组和1.2中我们提到的那些词组都属于短语表中的“鸡肋”，有的或许能借助语言模型进行修改，但有的很难修改，因而直接影响了机器翻译的效果。

2)一致性，对齐的一致性决定了对齐结果语料本身的可靠性。对齐的不一致主要来自于标注者之间及标注者内部的差异，产生差异往往是由于对齐和标注的标准不一致和不明确造成的。

如：我 叫 李 明 。　　我 叫 李 明 。



My name is Li Ming.　　My name is Li Ming.

上例是对同一句话不同的对齐标注者的对齐结果，很明显，不同的对齐标注方式会抽出不同的词组，这样，一方面会造成对齐结果良莠不齐，如“我〈〉my name”这样的对齐方式就不太合理，而“我叫 My name is”这样的词组则比较理想；另一方面也削弱了翻译概率的可靠性。

为了增强对齐的有效性和一致性，必须有一部好的对齐规范作为指导，我们的规范做到了如下两点：

1) 针对统计机器翻译的实际情况，提出了一种新的对齐标注方式：真对齐和伪对齐。真对齐指确

定的对齐，伪对齐指不确定的、只在小范围内语义相通或语法相关的对齐。真对齐又分为强对齐和弱对齐，强对齐是词和词之间确定的一一对齐，弱对齐是真对齐的词组中语义相关但不完全对等的词和词之间的对齐。这种明确的对齐区分解决了以往由于词语对齐而引发的翻译问题，增强了对齐结果语料的有效性。

2）在 LDC 规范[4]的框架下，进行了大量的修改和补充，形成了一部比较完善的对齐规范（以下简称为 ICT-NICT 规范），并附加了对齐指导手册，规范包括明确的对齐标注定义及区分、较为完整的语言分类体系、丰富的对齐规则和实例，这个规范加上我们精细化的项目质量管理，全方位地保证了手工词语对齐的一致性。

在该规范的指导下，我们完成了 5.8 万句对的汉英词语对齐任务（包括 2.3 万句对的篇章语料和 3.5 万句对的口语语料），随着工程的不断进行，我们不断扩充和完善了各种详细的对齐规则，并配以精心挑选的实例，力图使规范能覆盖对齐中遇到的各种语言现象，使得标注者在对齐过程中处处有规可循。另外，我们还将标注者提出的各种具体问题以及根据规范中的定义所作的相关解答分类汇总成《汉英词语对齐指导手册》，作为规范的补充，每周更新发放给标注者作为参考，标注者在项目初期，经常根据更新的规范对原先的工作进行修改。这种精细化的管理方式大大降低了对齐中容易出现的标注者之间及标注者内部的差异，从而增强了对齐结果的一致性。

我们将在下面的第二部分介绍 ICT-NICT 规范的几种对齐标注形式，第三部分介绍 ICT-NICT 规范对齐规则的变化和扩展，第四部分介绍对齐和标注一致性评价，第五部分是本篇的结语。

2　ICT-NICT 规范的几种对齐标注形式

词语对齐的目的是为了抽提出语义对等的词或词组（通过真对齐），以及抽提出其他对翻译有用的信息（通过伪对齐）。在下面的对齐图形中，我们用粗线、细线、虚线分别代表强、弱、伪三种对齐。

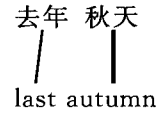
2.1　真对齐（强对齐和弱对齐）

真对齐是词或词组汉、英两部分在语义完全对等的基础上建立的对齐，这样的词或词组其汉、英两部分在一个较大的范围或某个专业领域内可以互译，互译具有较大的普遍性，可以构成一部内容非常详尽的双向翻译词典（包括专业词典）的词条内容，比如“大选〈〉general election”、“有 麻烦〈〉in trouble”，再比如商业领域的“完税〈〉duty paid”等都属于真对齐。

真对齐可以进一步分为强对齐和弱对齐，解释如下：

在对齐时，存在着词组语义完全对等、但组成它们的词找不到语义对等部分的情况。

如：去年　秋天



last autumn

很明显，“去年”和“last”语义并不对等，但是词组“去年秋天”和“last autumn”语义却是完全对等的，属于真对齐，为了抽提出这样的词组，我们将词组中语义相关但不对等的部分之间建立的对齐标注为弱对齐标记，而不管词还是词组，词和词之间只要语义对等、位置对应而建立的对齐都叫强对齐，可见，弱对齐是专门为词组的真对齐而设的，而强对齐是词和词之间的一一对齐，强对齐的词也可以构成词组的一部分，这样的区分可以帮助我们在翻译的后面步骤中过滤掉很多垃圾词对（如：去年〈〉last），抽提出更多有用的对齐（如：秋天〈〉autumn，去年秋天〈〉last autumn）。

2.2　伪对齐

请看以下 A、B、C、D 四种情况：

A. 如：美国　总统　谈　美国　与　东盟　关系

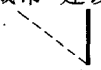
Clinton on U. S. -ASEAN Relations

例句中如果“美国 总统”和“Clinton”对齐，抽出的词组很可能使得其他句子中的“美国 总统”都翻译成“Clinton”，类似于前面提到的将“我国”统统译为“China”，我们把这一类情况都归于指代关系的对齐（还包括指示代词和被指代对象的对齐等）。

B. 上句中的“谈”和“on”在本例中语义对等、位置对应，按说也可以对齐，但是这种对齐关系并不具有普遍性，只局限于某些特定的语境，比如标题。

C. LDC 规范[4]指出：语言中还存在着省略和增加的现象，有一类省略和增加只是字面上的，原文和译文语义还是一致的（我们称为语义一致的省略）。

如：厦门 加强 城市 建设



Xiamen speeds up construction

其中，construction 可以理解为“城市 建设”。

D. 每种语言都有自己特定的语法要求，

如：他 买 了 一 本 书

He bought a book.

中文有量词“本”，英文没有相应的量词，

如：天气 真 棒 啊 ！

What a wonderful day!

中文有句尾语气词“啊”，英文没有相应的句尾语气词。

对比一下汉语和英语：

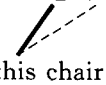
表1 汉英语法现象对比（★表示为某种语言特有）

	冠词	量词	句尾语气词	时态助词	从句先行词	不定式标记
汉语		★	★			
英语	★			★	★	★

在 LDC 规范中，或者将上面的 C 和 D 用粘合的办法（glue approach）来进行对齐，比如将 C 中的“城市”和“建设”粘合后再对齐到“construction”上，将 D 中的冠词粘合到中心名词上，再对齐到名词的对等部分，或者不进行对齐，只标注为“未翻译和正确的”，如 D 中汉语的句尾语气词。

ICT-NICT 规范借用了 LDC 规范的粘合对齐方法，与 LDC 规范不同的是：我们将粘合式的对齐区分出来，归为伪对齐一类，并且将以上四种类型的对齐都并入伪对齐。

如：这 张 椅子



this chair

依 LDC 规范的对齐结果为：“这 张〈〉this”，依 ICT-NICT 规范的对齐结果避免了本来译文比较确定的 this 产生众多翻译的可能，如“这位、这篇、这本、这辆……”，这样将不确定的、依赖上下文语境的伪对齐从确定的真对齐中分离出来，避免了“眉毛胡子一把抓”，也避免了 A 中提到的错误，使得对齐的层次比较分明；另外，伪对齐如“张〈〉this”也不同于对齐到空（不进行对齐），它不仅起到了一个“占位”的作用，为后继的翻译搜索提供了方便，而且这个对齐信息本身在下一步的翻译流程中还可以结合上下文的词类信息、语义信息和各种不同的翻译策略等加以利用，这种做法无疑增强了对齐结果的有效性。

伪对齐分两种情况：

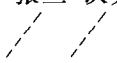
① 语义相通：仅局限于小范围（如：当前语句或特定的语境）内的、因语义相通而建立的词语对齐，比如指代（A）、小范围内的语义对等（B）、语义一致的省略（C）等；

② 语法相关：与特定语言的语法要求相关的对齐，比如 D。

ICT-NICT 在 LDC 规范的基础上重新定义了伪对齐中的粘合方法：当某些词（主要是虚词、实词中的量词和代词）在当前译文中找不到语义对等部分时，可以附着在跟其意义密切相关的主词上（如：助动词附着在中心动词上，限定词附着在中心名词上等），伪对齐到那个主词的对等部分上，这样的对齐方法叫做粘合。粘合的前提是该主词必须有可以对齐的词，否则就只能对齐到空。

伪对齐的部分完全不能用来作为翻译词典（包括专业词典）中的词条内容，但是可以为下一步的翻译搜索提供有效的帮助，是很有用的信息。

如：张三 认为



Zhang said

因为真对齐要求被对齐双方必须在较大范围内可以互译，上例中“张三”可以翻译成“Zhang”，但反过来的译文不具有普遍性，所以是伪对齐；同样，“认为”和“said”语义也不完全对等，只是在小范围内语义相通，所以也是伪对齐。

2.3 全连线

全连线是词组的一种对齐方式，主要用于这种情况：中文词组与对应的英文词组可以进行真对齐或伪对齐，但是词组内部有的词根本找不到可以独立对齐的部分。

如：华约 集团　　拉脱维亚 与 中国

The Warsaw Pact　The two countries

这样的对齐方式只能抽出词组的对齐，不能抽出单个词的对齐。

2.4 对齐到空

当一个词不能与对应译文中的任何词建立强、弱、伪对齐时，我们将之对齐到空。

如：酒店 有 理发师 吗 ？ Is there a hairdresser 's in the hotel?

上句的"'s"对齐到空。

3 对齐标注一致性评价

为了评价 ICT-NICT 规范的一致性，在项目快结束时，我们让一名对齐人员（标注者 1）重新对齐了 1 个多月前对齐过的 110 个口语句子，然后又对齐了另一名对齐人员（标注者 2）对齐过的 100 个口语句子；因为人工对齐是在自动对齐的基础上进行的，为了考察对齐人员是否对自动对齐结果有依赖性，我们在忽略对齐方式的情况下，分别计算了标注者 1 的 110 句对齐的第一次对齐结果和标注者 2 的 100 句对齐结果分别和 GIZA++[1] 自动对齐结果的一致性；另外我们进一步考察了篇章语料和口语语料的人工对齐结果和 GIZA++ 自动对齐结果的一致性及其差异，计算了 1.5 万句篇章语料（LDC 语料）和 3.5 万句口语语料（BTEC 语料）的人工对齐结果和 GIZA++ 自动对齐结果在忽略对齐方式的情况下的 Kappa 系数和 Dice 系数，以及以前者为标准答案的 GIZA++ 的对齐错误率（AER）[6] 和各种对齐方式的召回率。由于我们允许对齐人员修改原文中的错误（包括切分错误）和放弃对齐，为了保证被比较的文件原文词串完全相同，我们在统计时将修改过原文和放弃对齐的句子排除在外，实际参加统计的数据规模请参见表 1。

表 1　各组实际参加统计的数据规模

	篇章语料	口语语料	A11	A1g	A12	A2g
句对数	15 781	35 384	103	105	83	93
汉语词数	416 372	309 506	866	884	684	794
英语词数	479 260	334 082	969	991	737	849

我们采用 Kappa 系数[7-8] 和 Dice 系数来评价对齐的一致性，分别考察：

1）忽略对齐标注方式的一致性：即将强、弱、伪对齐都视为一种对齐；

2）区分对齐标注方式的一致性：即考察强、弱、伪三种对齐各自的对齐一致性。

在计算 Kappa 系数和 Dice 系数时，在以汉语词串和英语词串构成的矩阵中（如表 2），如果任意一对汉英词对（如"他〈〉He"）符合指定的对齐要求（如"强对齐"），我们就将它归为"T"类，如果不符合指定的对齐要求或者没有对齐，就归为"F"类。

Kappa 系数的计算公式为：

$$Kappa=(Po-Pe)/(1-Pe)$$

其中 Po、Pe 分别为观察值和期望值，以表 3 中的 A、B、C、D 来计算，A、B、C、D 分别是符合相应条件的总次数。

$Po=(A+D)/n$，$Pe=[(A+B)\times(A+C)+(D+B)\times(D+C)]/n^2$，其中 $n=A+B+C+D$。

表 2　Kappa 计算表 B

	他	去	北京	了
He	二			
has				---
gone		二		
to		—		
Beijing			二	

注：表 2 小格中上面的横线表示标注 1 的对齐，下面的横线表示标注 2 的对齐，粗、细、虚线分别代表强、弱、伪三种对齐方式。

表 3　Kappa 计算表 A

		标　注　1	
		T	T
标注 2	T	A	B
	F	C	D

如果我们要计算上例中强对齐的 Kappa 系数，那么 A=2，B=1，C=0，D=17，n=20，

Po=(2+17)/20=0.95, Pe=(3×2+17×18)/20²=0.78, Kappa=(0.95−0.78)/(1−0.78)=0.77

Dice系数的计算公式为:

$$Dice=2\times I/(A1+A2)$$

Dice系数只考虑非空对齐的情况,其中A1和A2分别是两份对齐的"T"类的数量,I是这两个对齐的"T"类的交集(相同对齐)的数量。

如果我们要计算上例中强对齐的Dice系数,那么A1=2, A2=3, I=2, $Dice=2\times2/(2+3)=0.8$

我们最终的实验结果如下:

结论:(Dice系数和Kappa系数存在着极高的一致性,我们以Kappa系数为例)

1) 依据本规范进行的词语对齐的标注者内部和标注者之间的对齐一致性都比较理想,两组强、弱、伪三种对齐的Kappa值分别为0.99、0.98、0.93和0.96、0.83、0.68,除标注者之间的伪对齐之外均超过了代表可靠程度好的0.8的阈值,而后者也达到了可以得出"是一致的"实验性结论的0.67的阈值[9];标注者内部和之间的强、弱、伪对齐的对齐一致性都呈递减趋势,符合我们的预期;

表4 标注者内部和标注者之间以及与GIZA++自动对齐的对齐一致性(口语语料)

	忽略对齐方式		区分对齐方式					
			强对齐		弱对齐		伪对齐	
	Dice	Kappa	Dice	Kappa	Dice	Kappa	Dice	Kappa
A11	0.988 3	0.986 9	0.987 4	0.986 6	0.977 9	0.977 2	0.928 4	0.926 9
A1g	0.780 4	0.755 2						
A12	0.945 7	0.939 3	0.961 4	0.958 9	0.832 9	0.828 3	0.686 5	0.680 0
A2g	0.792 9	0.770 5						

注:A11是标注者1内部的对齐一致性,A1g是标注者1与GIZA++对齐的一致性,A12是标注者1和2的对齐一致性,A2g是标注者2与GIZA++对齐的一致性。

表5 篇章语料和口语语料的手工对齐结果与GIZA++自动对齐结果的一致性及以手工对齐为标准答案的GIZA++的对齐情况

	在忽略对齐方式下与GIZA++的对齐一致性		GIZA++强对齐的召回率	GIZA++弱对齐的召回率	GIZA++伪对齐的召回率	GIZA++的对齐错误率(AER)
	Dice	Kappa				
篇章语料	0.655 6	0.644 9	0.872 9	0.588 5	0.171 9	0.361 3
口语语料	0.784 1	0.760 4	0.952 4	0.623 9	0.339 8	0.217 7

2) 标注者内部的对齐一致性明显高于标注者之间的对齐一致性,尤其在伪对齐上更加明显(差值为0.25);

3) 由于在第一个实验中采用的都是口语语料,所以人工对齐结果与GIZA++对齐结果之间存在着较高的一致性(Kappa系数为0.76左右),在忽略对齐方式的基础上,标注者1与GIZA++之间的Kappa系数为0.75,明显低于标注者内部的Kappa系数(0.99),标注者2与GIZA++之间的Kappa系数为0.77,明显低于标注者之间的Kappa系数(0.94),证明标注者对自动对齐结果的依赖程度不高;

4) 第二个实验中,GIZA++自动对齐结果与口语语料的人工对齐结果的Kappa系数为0.76,明显高于与篇章语料的人工对齐结果的Kappa系数0.64,以人工对齐结果为参考答案,口语语料的对齐错误率仅为0.22,低于篇章语料的对齐错误率0.36,可见GIZA++自动对齐工具针对口语语料的对齐优于针对篇章语料的对齐;GIZA++对齐结果对强、弱、伪三种对齐的召回率呈逐级递减的趋势,符合我们的预期。

4 初步的翻译实验及结果

为了验证本规范的有效性,我们用基于短语的统计翻译模型做了一个初步的实验:以3.5万依本规范对齐好的BTEC口语语料作为训练语料(共35 384个句对,369 587条对齐连线,其中强对齐占

54.17%，弱对齐占 25.34%，伪对齐占 20.49%），采用 Moses[10] 作为训练工具包，解码器采用了 NICT-ATR 自行开发的基于 Pharaoh[11] 的解码器，我们将 Moses 训练过程中 GIZA++ 双向对齐取交集、向邻边扩展这两个步骤的结果直接替换为手工对齐的结果，实验中都采用 IWSLT 2005 测试集作为开发集进行最小错误率训练（MERT）[12]，用 IWSLT 2008，2007 和 2006 测试集作为测试集。实验结果如表 7，其中，采用全部对齐的模型为 BTEC(swp)，只采用了强、弱对齐没有采用伪对齐的模型为 BTEC(sw)。

表 7　采用不同种手工对齐的结果与采用 GIZA++ 对齐结果的翻译 BLEU 值对比

Model	2008	2007	2006	对齐连线	短语表规模
GIZA++	0.471 6	0.307 5	0.183 7	375 353	626 502
BTEC (swp)	0.489 0	0.333 2	0.203 6	369 587	661 104
BTEC (sw)	0.499 6	0.312 9	0.186 7	293 848	1 339 597

结果表明，与采用 GIZA++ 对齐结果相比，BTEC(swp)的总体 BLEU[13] 值提升了大约 2 个百分点，与 GIZA++ 相比，BTEC(swp)的对齐连线少，但短语表更大；通过 BTEC(sw)的结果可以看出伪对齐在降低短语表规模、提高解码速度上作用非常明显。由于这只是一个初步的实验，怎样充分利用三种不同方式对齐的结果来提升翻译的质量？很多细节和策略还有待进一步研究。

5　结语

从最初的规范搭建、语料库试对齐到规范的基本确立总共历时 3 个月，其间经过了大量的讨论和几次大的修改，后来的标注者在对齐过程中提出的各种细节问题使得规范日臻完善，但远非完美。我们准备通过进一步的实验来提高这套对齐方式和规范的有效性，也欢迎同行们多多指正。

致谢

本规范的建立得到了 NICT-ATR 汉英词语对齐项目的资助，在项目的进展过程中，得到了日本情报通信研究机构山本博史博士、中国科学院计算技术研究所刘洋、黄瓒、夏天以及所有参加这个项目的对齐标注人员的帮助，在一致性评价实验数据方面，得到了新加坡信息研究院（I^2R）张民老师的指正，我们在此深表感谢！

参考文献：

[1] F. J. Och and Hermann Ney. A systematic comparison of various statistical alignment models [J]. Computational Linguistics, 2003, March, 29(1):19-51.

[2] Melamed, D. Annotation style guide for the Blinker project, Version 1.0.4. [R]. IRCS Technical Report # 98-06: University of Pennsylvania, Philadelphia, 1998.

[3] Jean Véronis. ARCADE Tagging guidelines for word alignment, Version 1.0. [OL]. 1998. http://aune.lpl.univ-aix.fr/projects/arcade/2nd/word/guide/index.html.

[4] Linguistic Data Consortium. Guidelines for Chinese-English Word Alignment, Version 1.1. [OL]. 2006. http://projects.ldc.upenn.edu/gale/Alignment/specs/GALE_Chinese_alignment_guidelines_v1.1.pdf.

[5] Linguistic Data Consortium. Guidelines for Chinese-English Word Alignment, Version 3.0. [OL]. 2008. http://projects.ldc.upenn.edu/gale/Alignment/specs/GALE_Chinese_alignment_guidelines_v3.0.pdf

[6] F. J. Och and H. Ney. Improved statistical alignment models [C]//Proc. of the 38th Annual Meeting of the ACL. Hong Kong, China, 2000: pages 440-447.

[7] J. Cohen. A coefficient of agreement for nominal scales [OL]. 1960. http://www.garfield.library.upenn.edu/classics1986/A1986AXF2600001.pdf.

[8] J. Carletta. Assessing agreement on classification tasks: the Kappa statistics [OL]. 1996. http://acl.ldc.upenn.edu/J/J96/J96-2004.pdf.

[9] K. Krippendorff. Content Analysis: An introduction to its Methodology [M]. Beverly Hills: Sage Publications, 1980.

[10] Philip Koehn et al. Moses: Open source toolkit for statistical machine translation [C]//Proceedings of the ACL Demo and Poster Sessions. 2007: pages 177-180.

[11] Philipp Koehn, Franz Josef Och and Daniel Marcu. Statistical phrase-based translation [C]//Proceedings of HLT/NAACL. 2003: pages 81-88.

[12] Franz Josef Och. Minimum Error Rate Training in Statistical Machine Translation [C]//Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics. 2003: pages 160-167.

[13] Papineni, K., S. Roukos, T. Ward, and W. J. Zhu. BLEU: a Method for Automatic Evaluation of Machine Translation [C]//Proceedings of 40th Annual Meeting of the Association for Computational Linguistics (ACL). Philadelphia, PA: 2002: pages 311-318.

附：汉英词语对齐规范

1 导言

词语对齐的任务包括：在一个平行语料文本中查找词或词组的对应关系，并对词与词之间的对应关系（即本规范中的强对齐、弱对齐、伪对齐和对齐到空）进行判断和标注。

本规范是在LDC汉英词语对齐规范（1.1版）的框架上建立的，并且参考了Blinker项目规范和ARCADE项目规范。

2 本规范术语说明

1）强对齐：源语言词与相应位置的目标语言词之间在较大范围或某个专有领域内语义完全对等，二者可以互译，可以抽提出来构成双向普通字典或专业字典的词条义项，这样的对齐叫强对齐，它是真对齐的一种，本规范中用↔表示强对齐。

如：牛奶↔ milk

2）弱对齐：为了词组整体的对齐而设立的单个词的对齐方式，表示组成对齐词组的词与其对应的词既不是强对齐，又不符合伪对齐的条件，但是为了抽提出词组又必须进行的对齐，它也是真对齐的一种，本规范用〈—〉表示弱对齐。

如：香港 中央 图书馆 **开 馆**

Hong Kong Central Library **Opens**

开↔ Opens　　馆〈—〉Opens（弱对齐）

伪对齐：伪对齐分两种情况，一种是仅局限于小范围（当前语句）因语义相通而建立的词语对齐，比如指代、语义一致的省略等；一种是与特定语言的语法要求相关的对齐。比如找不到对等部分的量词的粘合等。本规范用〈---〉表示伪对齐。

如：啊〈---〉！

4）对齐到空：当一个词不能与对应译文中的任何词建立以上三种对齐时，我们将之对齐到空，本规范用"〈～〉()"来表示对齐到空。

如：他 买 了 牛奶 **和** **大米** 。

He bought milk.

和〈～〉()　　大米〈～〉()

5）粘合：某些词（除量词、代词之外，主要是虚词）在当前译文中找不到语义对等部分，这时可以附着在跟其意义密切相关的主词上（如：助动词附着在中心动词上，限定词附着在中心名词上等），伪对齐到那个主词的对等部分上，这样的对齐方式叫做粘合。粘合主要用于语法相关的伪对齐。粘合的前提是该主词必须有可以对齐的词，否则就只能对齐到空了。关于什么样的词可以粘合，以及伪对齐到哪个部分，请参照本规范相关章节的具体说明。

6）实词：有具体意义的词。包括名词、处所词、方位词、时间词、区别词、数词、量词、代词、动词、形容词等。

7）虚词：虚词有的只起语法作用，本身没有什么具体的意义，如"的、把、被、所、呢、吗"，有的表示某种逻辑概念，如"因为、而且、和、或"等等。包括副词、介词、连词、助词、语气词、拟声词、感叹词。

3 对齐的原则

3.1 语义对等原则

在将一种语言翻译为另一种语言时，源语言的一个特定的单词或句子往往有多种翻译译文，所有这些译文都表达了同样的意思，即他们在语义上是对等的，只是选词和风格不同而已。这样，每个这样的译文都是源语言的单词或短语的正确译文，都可以与源语言的单词或短语进行对齐。

如：他 **拒绝** 了 他 的 提议 。

（version 1）He **refused** his offer.

（version 2）He **declined** his offer.

（version 3）He **said no to** his offer .

拒绝↔ refused

拒绝↔ declined

拒绝〈—〉said　拒绝〈—〉no　拒绝〈—〉to

如：难受 **死 了**

Extremely sad.

死〈—〉Extremely　　了〈—〉Extremely

注意：即使在词性和句子结构方面发生了变化也可以将它们连线对齐：

如：关系 的 **改善**

Improve relations

改善↔ Improve

如：德国 领土 上 存在 占领军

The occupying armies existing **in** German territory

上↔ in

如：为 双方 在 **领事** ，商业 和 其他 正式 代表 机构 的 活动 提供 便利

to offer convenience to the activities of each other's **consular**, commercial and other official representative institutes

领事↔ consular

3.2 位置对应原则

除了满足语义对等原则之外，要对齐的译文还必须出现在与原文相应的上下文语境的合适位置。

比如：某词在原文中不同地方出现了两次，但只有其中一个有对应译文，另一个没有出现相应的译文，这时应遵循位置对应原则，将没有出现对应译文的词对齐到空（请与后

面提到的真对齐中的“并列同指”加以区分）。

如：这个 问题 是 应 由 中国 人 自己 解决 的 问题

This issue is up to the Chinese themselves to resolve

问题(第二个)〈～〉()

如：**卢兹** 说 他 一贯 支持 … ，并 遵循……

Lutz said , he has consistently supported… and **he** followed ……

卢兹⟷ Lutz

()〈～〉 he(第二个)

3.3 最小匹配原则

在本规范中，严格遵循词与词优先对齐的原则：先考虑词一级的对齐，然后才考虑词组一级的对齐，除了专有名词外，真对齐尽量不采用全连线的对齐方式。

如：**最** 大 的 便利

as much convenience **as possible**

最〈—〉as(第一个)　　最〈—〉as(第二个)

最〈—〉possible

大〈—〉much　　的〈—〉much

便利⟷ convenience

如：去年 下半年

late last year

去年〈—〉last　　去年〈—〉year

下半年〈—〉late　　下半年〈—〉year

3.4 最大匹配原则

有时，为了保证双向语义对等，在进行一次连接时，被对齐的词需要多少就选取多少。

换言之，为了形成一个独立的语义单元，我们可能要尽可能多地选取参加连接的词。这类情形主要有固定表达法、成语(或习惯用语)、格言和带有连字符的单词，另外还有粘合得非常紧密的结构，比如动词短语，带前缀或后缀的结构等。以下一对多的情形就是采用的最大匹配原则。

如：齐头并进

Keep abreast with

齐头并进〈—〉Keep　　齐头并进〈—〉abreast　　齐头并进〈—〉with

但是，如果可以找到词对词(内容词)的翻译，那么，对齐最好采用最小匹配(特别是针对那些直接从英语中借用过来的表达形式)

如：条条 大路 通 罗马。

All roads lead to Rome.

条条〈—〉All　　大路⟷ roads　　通〈—〉lead　　通〈—〉to　　罗马⟷ Rome

3.5 减少对齐到空

如果一种语言的某个片断在另一种语言中找不到对应片段，也不能采用伪对齐，那么就把这些片断对齐到空，但是要尽量审慎地对待这种情况，尽量减少对齐到空的情况。

4 工作步骤

1) 在真正对齐之前，标注者必须浏览源语言和目标语言句子，以获得句子的大意；

2) 发现少量录入错误，如果容易修改的话，要先修改再进行对齐，并添加句对注释。

如：他 说 **你们** 是 我 今年 会见 的 第一 个 国会 众议员

He said, “**They** are the first US Congressional delegation that I have met this year.”

应将“They”改为“You”后，再对齐，并添加句对注释。

3) 发现明显的切分错误后，要先修改后再进行对齐，并添加句对注释。

如：为此 我们 **现 在** 保证

应在原文中将“现 在”改为“现在”后再进行对齐。

如：外国 对 **越 南** 投资 增加

应在原文中将“越 南”改为“越南”后再进行对齐。

注意：同属于一个单词的音译汉字应该合在一起，如“约 翰”合成“约翰”，“普 京”合成“普京”；同属一个英文缩写的字母要合在一起，如“U. S.”合成“U. S.”

在进行切分判断时，要判断它是两字词还是分开的两个词，可遵循以下规则：

1) 共现频率高；

2) 在两个字之间不能再插入任何其他的字；

那么这两个字就很可能是一个两字词，而不是两个单独的词。

4) 以源语言句子为基础，从源语言句子开始，首先对实词进行连线对齐，实词全部处理完后，再对虚词进行连线对齐。

5) 当所有语义对等的连接都做完后，剩下的词或短语或者伪对齐到其他部分上或者对齐到空。对于既不能真对齐又不能伪对齐的词语必须全部对齐到空节点上，一个词不能既对齐到空，又连接到其他非空节点上。

5 真对齐

真对齐是词或词组在语义完全对等的基础上建立的对齐，这样的词或词组在一个较大的范围或某个专业领域内可以互译，可以构成一部内容非常详尽的双向翻译词典(包括专业词典)的词条内容，这样的词典包含同一个词或词组的多种不同的翻译方法，但每种译法必须具有较大的普遍性。

真对齐又分两种情况：

1) 强对齐

强对齐的两部分的语义在一个较大的范围或某个专业领域内是严格对等的。

如：表示⟷ said

如：专用⟷ special

如：其 产品 经 香港 **转口** 到 美国 。

Its products **transfer** from Hong Kong to the US.

转口⟷ transfer　　美〈—〉the　　美⟷ US

2) 弱对齐

弱对齐是专门为真对齐的词组的组成成分而设的。因

为在语义完全对等的词组中，很多组成成分之间的对齐既不是强对齐，又不是伪对齐，这时我们为了能抽提出整个词组，就必须将这些成分进行弱对齐。

如：中国 的

Chinese

中国⟷Chinese 的〈—〉Chinese

如：篮球队

basketball team

篮球队〈—〉basketball 篮球队〈—〉team

如：去年 秋天

last autumn

去年〈—〉last 秋天⟷autumn

真对齐中有以下现象值得注意：

1）强固定搭配和专有名词

在强固定搭配和专有名词中，建议尽可能全部采用真对齐，比如限定词或介词的对齐建议采用弱对齐而不是伪对齐。

如：这里 **将来** 要 盖 一 栋 楼 。

A building will be built here **in the future.**

将来〈—〉in 将来〈—〉the 将来⟷future

如：一个 接 一个

one by one

一个⟷one 接〈—〉by 一个⟷one

如：中国 银行

The Bank of China

中国⟷China 银行⟷Bank

银行〈—〉the 中国〈—〉of

2）并列同指

在并列结构中，当并列成分的共同部分被合在一起翻译时，并列成分的各个相应的原文可以都对齐到合并后的译文上，这种处理方法在汉英两个方向上都适用。

如：内地 的 **专家** 和 台湾 的 **专家**

the **experts** from the mainland and Taiwan

专家(第一个)⟷experts 专家(第二个)⟷experts

如：公务员 **本地化** ，法律 **本地化**

the localization of public servants and laws

本地化(第一个)⟷localization 本地化(第二个)⟷localization

如：**各国** **各** 地区

all countries and regions

各国〈—〉all 各国〈—〉countries

各(第二个)⟷all

如：**中国** **的** 经济 实力 和 综合国力...

the **Chinese** economy and **China's** overall national strength...

中国⟷Chinese 的〈—〉Chinese

中国⟷China 的⟷'s

3）动词短语

动词短语是一个动词联合一个介词或一个副词性小品词。

它们以两个或多个词联合的方式出现。由于形成一个固定的含义时，这些小品词粘着性强，和它们的动词不容易分开，所以在本任务中，我们将它们当作一个整体来处理。

如：指出

point out

指出〈—〉point 指出〈—〉out

6 伪对齐

伪对齐的前提条件：1）伪对齐是在无法找到真对齐的情况下进行的对齐，即找不到具有普遍性的语义对等部分的情况下进行的对齐；2）伪对齐必须语义相通或语法相关。

伪对齐与真对齐的区别：伪对齐不是语义在较大范围上的完全对等，只是语义相通或语法相关，伪对齐的部分完全不能用来作为翻译词典（包括专业词典）中的词条内容。

语义相通表现在：伪对齐有时用于一种指代表示，比如指代词与指代物之间的对齐；有时用于语义对等但另一边表达略有省略因而翻译不具备普遍性；有时用于仅局限于某一特定场合的翻译等，总之，虽然语义相通，但这类伪对齐的两个部分的翻译不具有普遍性，其语义对等关系只在当前语句或某种语境下成立。

如：张三 认为，

Zhang said

张三〈---〉Zhang 认为〈---〉said

注意：我们此处谈论的语义对等是双向的。如："张三"可以翻译成"Zhang"，但反过来的译文不具有普遍性，所以是伪对齐。

如：泽曼 **透露** ，

Zeman **said** that

透露〈---〉said（语义不完全对等，对等关系只局限在当前语句中）

语法相关表现在：该对齐是与特定语言的语法要求相关的对齐（主要是粘合式对齐），如找不到语义对等部分的汉语的量词、英语的系动词、介词的对齐等等。

6.1 语义相通

6.1.1 指代关系

6.1.1.1 职务与人名

如：美国 总统 抵达 大马士革

Clinton arrives in Damascus

美国〈---〉Clinton 总统〈---〉Clinton

6.1.1.2 指代词与被指代对象

如：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 ……

卢兹〈---〉he

如：我 买 了 张 椅子 ，**椅子** 很 贵 。

I bought a chair, **that** is very expensive.

椅子〈---〉that

当英文中出现了比较具体的指示代词，对应的中文没有特指时，指示代词伪对齐到对等部分：

如：这次 演习 是 史上 **同类** 规模 最 大 、也 最 复杂 的 军事 实验 。

The exercise is the largest of **its kind** in history and also the most complicated war game .

同类〈---〉its　　同类〈---〉 kind

如：发展 **其** 在 华 业务

expand **its** business in China

其〈---〉its

如：**其中** 一个 人

One **of them**

其中〈---〉of　　其中〈---〉them

如；**及其**

and its

及其〈---〉 and　　及其〈---〉 its

如：**自己** 的

its own

自己〈---〉 its　　的〈---〉 its

自己〈—〉 own　　的〈—〉 own

6.1.2　语义一致的省略或增加

为了更好地理解语义一致的省略或增加，我们区分三种情形：词一级、句子级和谈话级

词一级：

如：他 带来 了 **书**

He brought **his books.**

书〈---〉his　　书⟷ books

此例中，原文中只有一个词“书”，省略了“他的”。仔细检查这种省略，我们可以得出如下结论：“他的”在源语言中只是在词汇上省略了，但在语义上还是表达出来了。根据语义对等原则，我们可以将这种类型的省略或增加采用粘合的方法进行伪对齐。

句子级：

虽然词在译文或原文中被明显地省略或增加了，但是从一个句子级上进行考察，这种省略或增加是充分和正确的。

如：屋子 很 **大** 。

The room is **big in dimension.**

大⟷ big　　大〈---〉in　　大〈---〉dimension

在词汇或词一级，“in dimension”相对于“big ⟷大”显得多余，但是，短语“in dimension”加进来是为了表达与“a room being big”稍有不同的意思，是“big”意义的一个扩展。所以，在这个意义上，当考虑到主语的时候，“in dimension”从句子级的角度是可接受的，同样可以与“big”粘合后伪对齐到“大”上。

如：厦门 加强 **城市** **建设**

Xiamen speeds up **construction**

建设⟷ construction　　城市〈---〉construction

如：最惠国 待遇

Most-favored nation **trade status**

待遇〈---〉trade　　待遇⟷ status

如：**德国** 经济 开始 回升

Economy in **Western Germany** begins recovery

德国⟷ Germany　　德国〈---〉Western

如：会谈 将 于 明日 **继续** 在 北京 **举行**

The talks are scheduled to **continue** tomorrow

继续⟷ continue　　举行〈---〉continue

谈话级：

比如，在一篇文章的开头，我们谈到了“中国篮球队”，在后面的句子里我们省略了这个短语中的“中国”，虽然很明显从谈话的角度“篮球队”和“中国篮球队”是指一回事，但是，如果从词一级或局部来考察，我们决不可以说这两个短语在语义上是等价的，它们只有通过增加谈话线索才是等价的。这时，我们也可以这样对齐：篮球队〈---〉Chinese

篮球队〈—〉basketball　　篮球队〈—〉team

注意：被省略或增加的部分只有在当前句对中与对应部分的语义一致，才允许建立伪对齐。如果二者隐含的语义并不一致，就只能对齐到空。

6.1.3　特殊语境下的翻译

只有在特定的语言环境下（如标题），对齐的两个部分的语义才是对等的。

如：美国 代表 **谈** 难民 问题

American Delegate **on** Refugee Issues

谈〈---〉on

如：墨西哥 总统 **访问** 上海

Mexican President **in** Shanghai

访问〈---〉in

6.1.4　中文原文中出现英文

当中文原文中出现英文时，将中文中的英文部分进行伪对齐。

如：布鲁斯·史宾斯汀（ **Bruce Springsteen** ）

Bruce Springsteen

B r u c e〈---〉 Bruce　　S p r i n g s t e e n〈---〉 Springsteen

6.2　语法相关

这种伪对齐是针对特定语言的语法要求而建立的对齐（主要是粘合式对齐）。比如，汉语中有丰富的量词，但英语中没有这一词类，当遇到量词的对齐时，不论词性如何，只要能找到语义对等的部分，我们都可以进行真对齐，如“一杯茶 a cup of tea”，“杯”可以强对齐到“cup”（虽然 cup 是名词），但绝大多数情况下，汉语中的量词在英语译文中都找不到语义对等部分，这时我们就采用粘合的方式进行伪对齐。

汉英语法现象对比(★表示为某种语言特有)

	冠词	量词	句尾语气词	时态助词	从句先行词	不定式标记
汉语		★	★			
英语	★			★	★	★

以上这些★部分如果找不到语义对等部分,都采用粘合的方式进行伪对齐。

6.2.1 不定词"to"

1) 不定词"to"若能找到其对等的翻译,直接对齐就可以了。

如:**为了** 解决 问题 ,我们 讨论 了 一 下午 。

To solve this problem, we discussed the whole afternoon.

为了↔to

如:以 设法 解决

To try to solve

以↔to

2) 不定词"to"处在固定搭配中,应弱对齐到对应部分。

如:继续

continue to

继续↔continue 继续〈—〉to

注意:to 如果隶属于某个较固定的搭配最好放在搭配中处理,而不是粘着在后面的动词上。

如:坚决

is determined to

坚决〈—〉is 坚决↔determined 坚决〈—〉to

3) 其他情况:粘合在其后的动词上。有时,一些词可能插在它们中间,这时需要格外小心。

如:我们 希望 他 继续 努力 ,迅速 有效 地 **解决** 问题 。

We hope he could make further efforts to quickly and efficiently **solve** the problem.

解决〈---〉to 解决↔solve

6.2.2 句尾语气词

包括帮助在句尾形成感叹的词,如:"呀"、"啊"、"哇"等,帮助在句尾形成疑问的词,如"吗"、"么"、"呢"等,可以与后面的标点符号粘合伪对齐到对应英文的标点符号上。

如:天气 真 棒 啊 !

What a wonderful day !

啊〈---〉! ! ↔!

但是,它们出现在句中表示犹豫或改变想法时,如果能在目标语言中找到对应的片段,就真对齐,否则,对齐到空。

6.2.3 量词

虽然量词的选择是由其后的名词决定的,但考虑到量词与前面的数词或代词位置更加接近,粘合更加紧密,所以量词一般伪对齐到其前的数词或指示代词的对等部分上。

如:他 买 了 一 **本** 书

He bought a book.

一↔a 本〈---〉a

如:他 是 第一 个 提出 此 议案 的 人

He is the first to propose such a motion.

第一↔first 第一〈—〉the 个〈---〉first

如:这 张 椅子

this chair

这↔this 张〈---〉this

注意以下三种情况:

当表示单个事物、汉语中省略了数词时,将量词直接伪对齐到英文中对等的数词上。

如:他 买 了 **本** 书

He bought a book.

本〈---〉a

当数词有多个时,量词粘合伪对齐到多个数词上。

如:一千六百万 **名**

16 million soldiers

名〈---〉16 名〈---〉million

如:二百四十多 **家**

More than 240

家〈---〉more 家〈---〉than

家〈---〉240

如:成千上万 **名**

tens of thousands of

成千上万〈—〉tens 成千上万〈—〉of 成千上万〈—〉thousands 成千上万〈—〉of

名〈---〉tens 名〈---〉of 名〈---〉thousands 名〈---〉of

与中文量词有关的英文"NP of"结构,of 粘合伪对齐到量词上。

如:一 **例** 回扣

one **case of** kickback

例↔case 例〈---〉of

如:一 **系列**

a **series of**

系列↔series 系列〈---〉of

如:第一 **轮** 投票

the first **round of** voting

轮↔round 轮〈---〉of

6.2.4 系表结构中多出来的系动词

系表结构中系动词有对等部分则直接对齐到对等部分,没有对等部分则伪对齐到离它最近的表语上。

如:她 又 高 又 苗条 。

She **is** tall and slender.

高〈---〉is

如：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.

愿〈---〉**is**

如：他 在 。

He **is** in.

在〈---〉is　　在←→ in

6.2.5 其他

具体情形，请参考 11.

7 全连线

1) 真对齐的全连线

全连线的条件：词组与词组整体上语义完全对等，但是词组内部有的词根本找不到可以独立对齐或粘合的部分(如词组一边有省略的情况)。真对齐的全连线全部采用弱对齐。

专有名词词组

如：华约 集团 ("集团"译文省略)

the Warsaw Pact

华约〈—〉**the**　　华约〈—〉**Warsaw**

华约〈—〉**Pact**

集团〈—〉**the**　　集团〈—〉**Warsaw**

集团〈—〉**Pact**

如：高盛 (投资) 公司

Goldman Sachs

高盛〈—〉Goldman　　高盛〈—〉Sachs

(〈—〉Goldman　　(〈—〉Sachs

投资〈—〉Goldman　　投资〈—〉Sachs

)〈—〉Goldman　　)〈—〉Sachs

公司〈—〉Goldman　　公司〈—〉Sachs

如：新墨西哥 州

New Mexico

新墨西哥〈—〉**New**　　新墨西哥〈—〉**Mexico**

州〈—〉**New**　　州〈—〉**Mexico**

普通词组

如：以 …为 基础

on the basis of

以〈—〉on　以〈—〉the　以〈—〉basis　以〈—〉of

为〈—〉on　为〈—〉the　为〈—〉basis　为〈—〉of

基础〈—〉on　基础〈—〉the　基础〈—〉basis　基础〈—〉of

注意：如果一个专有名词或强固定搭配不用全连线，每个部分单独抽出来语义也对等，就不必全连线。但是如果有些部分单独抽出来不合适，最好还是采用全连线。

2) 伪对齐的全连线

主要用于：整个词组到词组的对齐本身属于伪对齐、但其中词和词之间不能全部一一对齐的情况。全连线时全部采用伪对齐。

如：**拉脱维亚** 与 **中国** 将 开展 积极 的 对话 与 合作 。

The two countries will have active dialogue and cooperation.

拉脱维亚〈---〉the　　拉脱维亚〈---〉two

拉脱维亚〈---〉countries

与〈---〉the　　与〈---〉two　　与〈---〉countries

中国〈---〉the　　中国〈---〉two　　中国〈---〉countries

如：**刘 华章** 介绍

The **Financial Planner** explained

刘〈---〉The　　刘〈---〉Financial　　刘〈---〉Planner

华章〈---〉The　　华章〈---〉Financial　　华章〈---〉Planner

8 对齐到空

对于找不到语义对等部分、且无法进行伪对齐的词，我们统一将之对齐到空。

8.1 句法性虚词

对于"it"、"there"、"here"等词，当它们在句子中只起一个句法作用，对句子的意思没有贡献时，如果没有对等部分，多出来的"it""there"或"here"可以对齐到空。

如：努力 准备 考试 很 重要 。

It is important that you work hard for the exam.

()〈～〉It

如：汽车 来 了 。

There comes the bus.

()〈～〉There

注意：当 there 表"存在"出现在"there be"句型中时，**there** 与 **be** 都参加对齐(弱对齐到 有/无/没有/…)：

如：那 项 报导 **无法** 证实 。

there is no way to verify that report

无法〈—〉there　　无法〈—〉is

无法〈—〉no　　无法〈—〉way

如：从 那时 起 ，欧洲 政局 **无** 一 日 安宁 ，危机 重重 。

Henceforth, **there has not been** a single day of peace in the political situation of Europe , a crisis-ridden continent.

无〈—〉there　　无〈—〉has　　无〈—〉not

无〈—〉been

如：**有** 八十 美元 左右 的 包 吗 ？

Are there any bags aroud eighty dollars?

有〈—〉are　　有〈—〉there　　吗〈—〉are

吗〈—〉there

8.2 个人用词风格

由于个人用词风格产生的插入或省略，与流派有关的插入和其他实际中的或谈话中的特点，比如，在广播对话中，有纯粹的口语式的插入，表示犹豫、改变话题或交谈中个人的用词习惯，它们在语义上不重要，没有对应译文，应对齐到空。

如：那么 敌后 战场 **呢** ，日本 为了 要 夺 权 ，巩固

他 **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.

呢〈~〉() 这个〈~〉() 那个〈~〉()

8.3 对话中的省略

如: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, finally lost in the selection.

()〈~〉since

8.4 翻译丢失

当词或词串及其信息的译文都丢失时,如果在句子中占的比重不大,还不足以放弃对齐,修改起来又不方便时,可以对齐到空。

如:他 买 了 牛奶 **和** **大米** 。

He bought milk.

和〈~〉() 大米〈~〉()

8.5 译文选择

在对齐时,有时一个词的译文在不同地方出现了两次或多次,如果不是"并列同指"的话,要避免将同一个词对齐到不同的地方,而是要对齐到一个位置相应的地方,而剩下的不能对齐的部分就对齐到空。

如:他们 相信 通过 逐步 **微调** 办法 ,把 银行 利率 调整 到 … ,将 …

They believe that the strategy to fine tune the rate to … will …

应:微调〈~〉() 调整〈---〉 fine

调整⟷ tune

不应:微调〈—〉fine 微调〈—〉 tune

调整⟷ tune

如果这样连线,tune 这个词既在"微调(fine tune)"词组中,又在"调整(tune)"中,势必造成抽词组困难,所以这种情况要避免。

如:这个 问题 是 应 由 中国 人 自己 解决 **的** **问题**

This issue is up to the Chinese themselves to resolve

的〈~〉() 问题(第二个)〈~〉()

9 放弃对齐

针对空白的句子、有一半或一半以上不匹配的句子、翻译了一半或不到一半的句子、两边都是纯英文的句子、完全意译的句子,可使用"放弃对齐"按钮放弃对齐,并在"句对注释"栏中添加放弃的原因(空白、不匹配、未翻完、纯英文、意译等),有其他放弃原因也请注明。

如:这 实在 微不足道 ,不过 尽力 而已 。

It is the least I can do.

如:三 个 臭 皮匠,顶 个 诸葛 亮。

Two heads are better than one.

10 句对注释

1) 如果发现任何潜在的问题,请在"句对注释"栏中添加注释。

2) 以下情况必须添加"句对注释":录入错误、切分错误、放弃对齐。主要是标明错在哪里、怎么改的、放弃对齐的原因(空白、不匹配、未翻完、纯英文、意译等)。

如:巴西 宣布 **基** **磋** **施** 投资 计划

Brazil Announces Investment Plan for 1998

宜将"基 磋 施"改为"基础 设施",再进行对齐,并在"句对注释"栏中记录改动情况:"原:基 磋 施 现:基础 设施"

11 各种具体情况

11.1 介词

1) 对于介词来说,如果有对等部分,直接对齐即可。

2) 对于英文部分有两个或多个介词,而中文部分只有一个对等介词,可以将两个介词粘合,将它们对齐到另一边的对等介词上;同样,当中文中有介词后缀时,也和前面的介词粘合,真对齐到对等的英文介词上。

如:**对** 美国 **来** **说**

To the **US**

对⟷ to 来〈—〉to 说〈—〉to

如:在 近年 来

In recent years

在⟷ in 来〈—〉in

如:**在** 桌子 上面

On the table

在⟷ on 上面⟷ on

如:遍布 **在** 这 片 热土 **上**

have spread out all **over** this stretch of hot turf.

遍布〈—〉spread 遍布〈—〉out

遍布〈—〉all 遍布〈—〉over

在〈—〉over 上〈—〉over

3) 对于一边有一个介词,而另一边没有对等部分的情形,分以下几种情况:

1' 强固定搭配:介词使用真对齐中的弱对齐

如:**我** 同意 你 的 意见 。

I agree **with** you.

同意⟷ agree 同意〈—〉with

如:我 同意 你 的 意见 。

I am **in** agreement **with** you.

同意〈—〉in 同意⟷ agreement

同意〈—〉with

如:健康 状况 良好

in good health

健康⟷ health 状况〈—〉health

状况〈—〉in 良好⟷ good

如:对 香烟 加以 规范(注意识别中文词组)

regulate cigarettes

对〈—〉regulate　　加以〈—〉regulate

规范←→regulate

2’弱固定搭配：将介词伪对齐到与它搭配的词的对等部分。

如：尼加拉瓜 总统 结束 访问 芬兰

Nicaragua president ends visit to Finland

访问←→visit(名词)　　访问〈---〉to

如：目前 ，警方 正在 调查 事故 原因

Investigation **into** the cause of the accident is underway

调查←→Investigation　　调查〈---〉into

对结构比较松散的词组：

如：芬兰 向 埃及 提供 环保 援助

Finland to Provide Egypt **With** Environmental Aid

向〈---〉provide　　提供←→provide

提供〈---〉With

3’介词后接地点、范围、方面、领域、时间类的词，介词没有对等部分且不与其他词构成固定搭配

如：广州 成立 破产 法庭

Bankruptcy Court Established in Guangzhou

广州〈---〉in　　广州←→Guangzhou

如：世界 科学 大会

World Conference on Science

科学〈---〉on　　科学←→Science

如：他 1997年 毕业 。

He graduated in 1997.

1997年〈---〉in　　1997年←→1997

如：他 9日 来。

He will come **on** the 9th.

9日〈---〉on　　9日〈—〉the　　9日←→9th

4’对齐到空

对等部分未出现，对齐到空

如：做 实验 可以 得出 结果 。

Results can be obtained **by** doing experiments.

()〈～〉by

汉语中的介词和副词没有合适的对齐部分时，对齐到空。

如：他 把 鸡蛋 给 吃 了 。

He has eaten the eggs.

把〈～〉()　　给〈～〉()

如：国家 每年 都 拨 专款

Every year the nation allocates special funds

都〈～〉()

如：就 在 人们 议论纷纷 之际 ，…

When people were commenting on this, …

就〈～〉()

4）介词及限定词的伪对齐部分(量词的请见规范6.2.3)

• 如果有一个完整的主名词，则直接将介词或限定词伪对齐到主名词对应的译文上，即使这个主名词有时与它的限定词相隔较远。

如：社会主义 制度

the socialist system

制度〈---〉**the**

如：新区 管委会

the new region ’s management committee

新区〈---〉the (注意：此处the对齐的不是“管委会”)

如：澳门 特别 行政区

in the Macao Special Administrative Region

行政区〈---〉in　　行政区〈---〉the

• 如果主名词是一个由“and”或”or”连接的联合结构，则只要粘合在离它最近的主名词上，伪对齐到其对等部分。

如：中国 的 发展 和 进步

the development and progress of China

发展〈---〉the

如：中国 和 美国 …

In China and U. S. , …

中国〈---〉In

• 如果很难分离出明显的主名词，则一一伪对齐。

如：三 到 五 年

for 3 to 5 years

for〈---〉三　　for〈---〉到　　for〈---〉五

for〈---〉年

• 主名词有强对齐部分，又有伪对齐部分时，介词或限定词伪对齐到强对齐部分。

如：欧洲 股票 市场 创下 了 新 低 价位 。

European stock markets recorded **a** new low.

低〈---〉a　　低←→low　　价位〈---〉low

11.2 英语助动词

助动词提供了关于它后面的主动词的更多的语义信息。助动词可以有多种形式，它用来表示以下功能：被动、进行、完成、情态或虚拟。只要两边都出现了对等部分，就可以直接进行对齐。

如：他 会 来 。

He **will** come.

会←→will

如：谈判 也 能 产生 重要 进展 。

negotiations **will** also make major progress.

能←→will

当一边没有对等部分时，多出来的助动词分以下几种情况：

1）否定句：将助动词伪对齐到否定词上

如：他 不 知道 。

He does not know.

不〈---〉does　　不←→not

2）疑问句：起语法作用的助动词因与中心动词相隔较

远，所以对齐到空，但含语义信息（否定）的否定词仍进行对齐。

如：你 有….

Do you have…

()〈～〉Do

如：你 **不**….

Do not you

()〈～〉Do 不⟷not

3）与主动词粘合伪对齐到对应的中心动词上

如：蜡烛 吹灭 了 。

The candles **were** blown out.

吹灭〈---〉were 吹灭〈—〉blown

吹灭〈—〉out 了〈---〉blown

如：早日 达成 双边 协议

An agreement **can be** reached

达成〈---〉can 达成〈---〉be 达成⟷reached

11.3 汉语的时态

因为缺乏屈折形态，汉语动词不改变形式，用诸如“着，了，过，一直，正，正在”这样的词表示时态和时间，这些词应伪对齐到英文中的相似表达（如表完成、进行的助动词）上，这是一种特殊的语义相通又语法相关的伪对齐，如果没有对等部分，则与中心动词粘合伪对齐到后者的对等部分上。对“曾、已经、将、会、能”等类似的词，如果有语义对等部分的话，可以进行真对齐，否则也作类似处理。

如：图书馆 **已经** 建成 。

The library has been completed.

已经〈---〉has 建成〈---〉been

建成⟷completed

如：他 去 北京 **了** 。

He **has** gone to Beijing.

了〈---〉has 去⟷gone 去〈—〉to

如：他 听 着 音乐 。

He **was** listening to music.

听⟷listening 听〈—〉to 着〈---〉was

如：他 去 **过** 北京 。

He visited Beijing.

去⟷visited 过〈---〉visited

如：不 曾 遇到 **过**

have not been encountered

不⟷not 曾〈---〉have 遇到〈---〉been

遇到⟷encountered 过〈---〉have

如：爱尔兰 总统 **将** 正式 访问 英国（标题）

Irish President **to** Visit Britain

将〈---〉to

更复杂的情况（含被动语态）：

如：**受到** 俘虏 的 对待

is being treated as a prisoner.

受到〈---〉is（表进行时态，与 being 粘合）

受到〈---〉being 对待⟷treated 对待〈—〉as

如：**遭** 人 下 毒

have **been** poisoned

遭〈---〉been 遭〈---〉have（表完成时态，与 been 粘合）下〈—〉poisoned 毒〈—〉poisoned

11.4 “的”、“地”、“得”

11.4.1 “的”

11.4.1.1 表所有格的“的”

所有格用来表示所有关系。英语中所有格可以采取两种形式：“’s”（或’）和“of”。

如果两边都出现所有格，直接对齐即可。

如：中国 的 领土

China ’s territory 的⟷’s

如：中国 的 领土

The territory of China

的⟷of

如果英文复数的所有格只是一个省略符“’”，那么将省略符“’”按和“’s”同样的方式处理，因为它们在语义上是等价的。

如：学生 的 教科书

Students ’ textbooks

的⟷’

关于“of”和“的”的对齐请注意：

1）英语中的“of”经常与“的”进行连接，使用机器自动对齐的结果中基本上所有的“of”都会和某个“的”进行连接，但是应看清是否确为对应。

如：她 个子 高 ，能 看到 墙 的 那 一边

she can see over the wall because of her height

机器对齐的结果是“的”和“of”对应，但这里显然不是。

2）固定搭配中 of 和其他部分一起连到对应的译文上

如：去年 初

at the beginning of last year

初〈—〉at 初〈—〉the 初⟷beginning

初〈—〉of

如：数以万计 **的**

tens of thousands **of**

数以万计〈—〉tens 数以万计〈—〉of

数以万计〈—〉thousands 数以万计〈—〉of

的〈～〉()

如：百分之五十 的 女性

50 percent **of** women

百分之五十〈—〉50 百分之五十〈—〉percent

百分之五十〈—〉of

的〈～〉()

3）语义一致的省略

如：关塔那摩 **的** 囚犯

Guantanamo prisoners

关塔那摩⟷Guantanamo 的〈---〉Guantanamo

如：中国 领土

territory **of** China

中国⟷China　　中国〈---〉of

应该理解为：在“中国”后插入“的”字，整个词组“中国的领土”读得通顺，才可以将 of 视为省略粘合伪对齐到“中国”上。所以：

如：石油 业 将 从 此 蓬勃 发展

robust development **of** the oil sector

上例中的 of 不应再伪对齐“业”

4) city of/town of/capital of 等后接地名，当地名和 town、city、capital 是同位语关系时，of 不应当再当作对应的“的”省略的情况来处理，而应伪对齐到 city、town、capital 等词上。

如：雅夫纳 市

city **of** Jaffna

市⟷city　　雅夫纳⟷Jaffna　　市〈---〉of

如：印尼 首都 雅加达

the Indonesian capital of Jakarta

首都⟷capital　　雅加达⟷Jakarta

首都〈---〉of

但下列情况是“的”省略：

泰米尔纳德邦 首府 钦奈

Chennai , the capital **of** Tamil Nadu State

泰米尔纳德邦〈—〉Tamil　　泰米尔纳德邦〈—〉Nadu

泰米尔纳德邦〈—〉State　　首府⟷capital

钦奈⟷Chennai　　泰米尔纳德邦〈---〉of

11.4.1.2　形容词后的“的”

1) 英语中有相应的形容词，则将“的”与中文形容词一并真对齐到英文形容词上

如：雪白 的 雪 。

White snow

雪白⟷white

的〈—〉white（注意：此处是真对齐的弱对齐）

2) 与之类似，以下的“的”也可以弱对齐到对应的作为前置定语的英文分词形式上

如：随行 的 记者

the accompanying reporters

随行⟷accompanying

的〈—〉accompanying

如：以 美国 为 首 的 国家

US-led countries.

的〈—〉US-led

11.4.1.3　“所…的”结构

如：他 描述 了 他 所 看到 的 。

He described what he saw.

所〈---〉what　　的〈---〉what

11.4.1.4　其他定语后的“的”

1) 定语从句：“的”如果有关系代词与之相应，可直接与之进行伪对齐，否则对齐到空。详见 11.12 中的定语从句部分。

如：目前 在 当兵 的 姜敏元 …

Jiang Mingyuan **who** is serving in the army now…

的〈---〉who

如：分享 价值 25 万 美元 的 20 根 金条

Sharing 20 gold bars worth 250,000 US dollars

的〈～〉()

如：中央台 报道 的 新闻 很 有趣 。

The news reported by CCTV is very interesting.

的〈～〉()

2) 如果介词引导的介词结构作定语，且中文中出现了“的”，那么“的”的对齐可以分成两种情况：

• 介词有对应的中文译文时，介词和对应部分可以强对齐，“的”对齐到空：

来自 美 、英 、法 三 国 的 士兵

Soldiers from three countries , namely , the US , UK and France

的〈～〉()

加 熏肉 的 煎 鸡蛋。

Scrambled eggs with bacon, please .

的〈～〉()

我 在 去 迈阿密 海滩 的 路上 。

I'm on the way to Miami Beach .

的〈～〉()

去 华盛顿 的 十二 号 航班 是 哪个 登机门 ？

Which boarding gate is flight twelve to Washington?

的〈～〉()

对 问题 的 洞察

its penetration of the problem

的〈～〉()

• 英文介词没有对应的中文译文时，“的”可以伪对齐到英文介词上：

如：大 地主 的 投降

capitulation **by** the big landlords

的〈---〉by

如：五月 的 第三 个 星期日

the third Suday **in** May

的〈---〉in

如：美 、英 、法 三 国 的 士兵

Soldiers **from** three countries , namely , the US , UK and France

的〈---〉from

如：一 个 21 世纪 的 共同 方案

A common program **for** the 21st century

的〈---〉for

如：北京 的 胜利

a victory **for** Beijing

的〈---〉for

如：约旦河 西岸 **的** 居民

Residents **in** the West Bank

的〈---〉in

3）其他对齐到空的情况

如：他 是 个 二十 岁 **的** 青年 。

He is a 20 year old youth.

的〈～〉()

4）与“的”类似的“之”

如：他 结束 东盟 5 国 **之** 行

He Shi concludes visit to 5 ASEAN countries

之〈---〉to

11.4.1.5 中心词省略的“的”字结构

如：实行 拍卖 **的** ，可 减免 有关 税收 ；

for those implementing auctions , related taxes can be reduced or eliminated ;

的〈---〉those

如：有 白色 **的** 吗?

Do you have a white one?

白色〈---〉a 的〈---〉a 白色⟷ white 的〈---〉one

11.4.1.6 句尾的“的”

在中文“是…的。”强调句型中，“是”和句尾的“的”均对齐到空。

如：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 .

是〈～〉() 的〈～〉()

如：鲁宾 一行 **是** 应 财政部 **的** 邀请 访华 **的** 。

Rubin and his delegation are visiting China at the invitation of the Ministry of Finance .

是〈～〉() 的〈～〉()

11.4.2 “地”、“得”

如：他 深深 **地** 爱上 了 她 。

He loves her deeply .

深深⟷ deeply 地〈—〉deeply

如：他 干 **得** 好 。

He did well .

干⟷ did 得〈---〉did

注意：“他们”“的”都要与“their”真对齐；“逐渐”“的”都要与“gradual”真对齐；“气冲冲”“地”都要与“angrily”真对齐。

11.5 重复

中文中重复很常见且形式不一。一般情况下，如果在英文中没有找到重复的话，被重复的字要真对齐。但个别情况除外。

11.5.1 名词性重复

如：人人 都 到 了 。

Everyone arrived .

人人⟷ everyone

11.5.2 动词性重复

如：让 他 **试 一 试** 。

Let him **have a try** .

试（第一个）〈—〉have —⟷ a 试（第二个）⟷ try

如：让 他 **试 试** 。

Let him **have a try** .

试（第一个）〈—〉have 试（第二个）〈---〉a

试（第二个）⟷ try

如：他 去 **散散步** 。

He went to **take a walk** .

散散步〈—〉take 散散步〈—〉a

散散步〈—〉walk

如：他 **解释 了 又 解释** 。

He **explained and explained** .

解释（第一个）⟷ explained（第一个）

解释（第二个）⟷ explained（第二个）

了〈---〉explained（第一个） 又⟷ and

如：他 **做** 作业 **做** 得 好 。

He did his homework well .

做（第一个）⟷ did 做（第二个）〈---〉did

得〈---〉did

11.5.3 形容词性重复

如：太阳 红彤彤 的 。

The sun is red .

红彤彤⟷ red 的〈—〉red

如：他 傻不拉几 的 。

He is stupid .

傻不拉几⟷ stupid 的〈—〉stupid

如：屋子 干干净净 。

The room is clean .

干干净净⟷ clean

11.5.4 量词性重复

如：花 一 朵朵 地 凋凌 。

The flowers withered gradually .

一〈—〉gradually 朵朵〈—〉gradually

地〈—〉gradually

注意：“地”字必须参加真对齐。

11.6 离合动词

汉语离合动词的不同部分在对齐时遵循：动词部分伪对齐，名词部分真对齐。

如：他 深深 地 **鞠** 了 一 **躬** 。

He made a deep bow .

鞠〈---〉made 了〈---〉made —⟷ a 躬⟷ bow

如：睡觉 前 洗 个 热水 澡 。

Take a hot bath before going to bed .

洗〈---〉take　　澡←→ bath

11.7　专有名词

11.7.1　一般情况

1）人名、组织名、机构名、国名、地名及其首字母缩写形式等都被处理为一个完整的单位。

如：联合国

the United Nations

联合国〈—〉the　　联合国〈—〉United　　联合国〈—〉Nations

如：联合 航空局

the Joint Aviation Authorities （ JAA ）

联合←→ Joint　　航空局〈—〉Aviation　　航空局〈—〉Authorities

航空局〈—〉the　　联合〈—〉JAA　　航空局〈—〉JAA

联合〈—〉(　　航空局〈—〉(　　联合〈—〉)

航空局〈—〉)

如：中国　共产党

the CPC

共产党〈—〉the　　中国〈—〉CPC　　共产党〈—〉CPC

如：苏联

the USSR

苏联〈—〉the　　苏联←→ USSR

2）当专有名词内部的词不能一一对齐时，一般采用全连线弱对齐的方式。详见 7 全连线

11.7.2　人名的特殊情况

1）第一种情形：

如：张 衡

Zhang

张←→ Zhang　　衡〈---〉Zhang （伪对齐）

如：若斯潘

Lionel Jospin

若斯潘←→ Jospin　　若斯潘〈---〉Lionel （伪对齐）

2）第二种情形：

如：诸葛亮

Zhu

诸葛亮〈---〉 Zhu

3）第三种情形：

如：欧萨玛 · 宾拉登

Osama Bin Laden

欧萨玛←→ Osama　　·＜～＞()(对齐到空)

宾拉登〈—〉 Bin　　宾拉登〈—〉 Laden

11.8　前缀和后缀

11.8.1　前缀

中文中的前缀“本，该，此”(“本人”除外)通常应直接真对齐到其对等部分而无需粘合到别的词上，因为它们都有实义。

如：**本人** 姓 李。

My family name is Li .

本人〈—〉My

构成词组：本人姓(My family name is)的一部分

如：我 遇到 过 **此** 人 。

I met this person .

此←→ this

如：**该** 书 描叙 了 他 的 一生 。

The book describes his whole life .

该←→ The

如：**前** 总统 克林顿

Former president Clinton

前←→ Former

11.8.2　后缀

后缀一般应伪对齐在它前面的词的对齐部分的中心词上，因为它们通常没有实义。

11.8.2.1　名词后缀

象“者，儿，员…”一类的名词后缀应伪对齐到名词的对等部分的中心词上。

如：汤姆 是 个 小青年 **儿** 。

Tom is a young man .

小青年〈—〉young　　小青年〈—〉man　　儿〈---〉man

(注意：此处“儿”只伪对齐到中心词上)

11.8.2.2　动词后缀

动词后缀主要指趋向动词，对齐分以下两种情况：

1)无对等部分的对齐，粘合在中心动词上伪对齐。

如：从 … 发展 起来 的

developed from

发展←→ developed　　起来〈---〉developed

有对等部分的对齐，直接对齐即可。

如：他 走 **下来** 。

He walked **down** .

走←→ walked　　下来←→ down

11.9　限定词

11.9.1　一般情况

定冠词和不定冠词“a/an/the”以及代词“his/her/their/its/your/our/my”都属于限定词。

以冠词为例，分以下三种情况处理：

1）限定词 〈〉 限定词

如：卡斯特罗 称赞 中国 的 成功 是 一 个 奇迹

Castro praised as a miracle China's success.

—←→ a

2）() 〈〉 限定词 (或者反过来)

一边有限定词而另一边限定词省略的情况很常见，这时可将多出来的限定词伪对齐到它后面的主名词对应的译文上，具体对齐到哪个词上，和介词的对齐类似，请详见 11.1 介词

3）限定词 〈〉 不同的翻译(或者反过来)

有时，一边出现了一个限定词，而另一边出现的是一个与它对等的非限定词译文。当一个限定词被翻译成其他形式时(或反过来)，我们按它们表达的意思进行对齐(不局限

于同一词性)。

如:我 给 了 他 一 本 书 ,**这** 本 书 很 有趣 。

I gave him a book , **the** book is interesting .

这⟷the

11.9.2 关于"the"的几种特殊的对齐

1) 最高级的对齐

如:最 大 的 苹果

The largest apple

最〈—〉the 最〈—〉largest 大〈—〉largest

的〈—〉largest

如:他 最 高 。

He is the tallest .

最〈—〉the 最〈—〉tallest 高〈—〉tallest

如:以 萨克森 邦 最为 严重

Sachsen is the hardest hit state

最为〈—〉the 最为〈—〉hardest 严重〈—〉hardest

附:比较级的对齐

如:他 更 快 。

He is faster .

更〈—〉faster 快〈—〉 faster

如:它 更 重要 。

It is more important .

更⟷more 重要⟷important

2) 序数词的对齐

如:首 次

the first time

首〈—〉the 首⟷first

如:第二 天

the second day

第二〈—〉the 第二⟷second

如:他 31 日 去 的 。

He went on the 31th .

31 日〈┄〉on 31 日〈—〉the 31 日⟷31th

如:第五 个 国家

The fifth country

第五〈—〉the 第五⟷fifth 个〈┄〉fifth

如:第 5 个 国家

The fifth country

第〈—〉the 第〈—〉fifth 5〈—〉fifth

个〈┄〉 fifth

3) "年代"的对齐

如:七十 年代

the 1970s

七十〈—〉1970s 年代〈—〉1970s 七十〈—〉 the

年代〈—〉 the

4) 在固定搭配中的对齐

如:同 期

The same period

同〈—〉the 同⟷same 期⟷period

如:最后 仪式

the last ceremony

最后〈—〉the 最后⟷last 仪式⟷ceremony

11.10 连接词

11.10.1 带"and"的连接

以下情况只有"和(/以及 等)⟷and"、",⟷,"是真对齐,其他都是伪对齐或对齐到空:

1) 和〈〉and:直接强对齐

如:美国 政府 **和** 美国 国会

the US government **and** the US Congress

和⟷and

2) ,〈〉and (或者:和〈〉,):将逗号与"and"伪对齐

如:他 挥手 再见 ,开 车 走 了 。

He waved good-bye **and** drove away.

,〈┄〉and ,

3) ,〈〉,and (或者:,和〈〉,):将逗号与逗号强对齐,多出来的"and"和"和"伪对齐到逗号

如:中方 希望 美国 谨慎 妥善 处理 台湾 问题 ,不要 让 这个 问题 干扰 中 美 关系 的 稳定 发展

China hopes that the US will carefully and appropriately handle the issue, **and** not allow the Taiwan issue to interfere with the stable development of Sino-US relations.

,⟷, ,〈┄〉and

4) 和〈〉,and (或者:,和〈〉and):将"和"与"and"强对齐,多出来的",伪对齐到"和"或"and"

如:鲁宾 说 , 美国 非常 重视 发展 与 中国 的 经济 合作 , 特别 是 在 美 中 两 国 经济 持续 增长 **和** 双边 投资 流动 不断 增加 的 情况 下 更 是 如此

Rubin said the US emphasizes economic cooperation with China , especially at a time when the economies of both countries are experiencing sustained growth , **and** the exchange flow of mutual investments is increasing .

和⟷and 和〈┄〉,

5) ()〈〉and(或者:和〈〉()):将多出来的"and/和"对齐到空

如:他 抬头 对 我 说 ……

He looked up and said to me that……

()〈~〉and

11.10.2 其他连接

"and"结构以外的连接可以以下面的例子来说明:

如:他 **既** 不 吃 **也** 不 喝 。

He **neither** eats **nor** drinks .

既〈—〉neither 不(第一个)〈—〉neither 也〈—〉nor 不(第二个)〈—〉nor

如:**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 re-

sponsibilities .

或者(第一个)⟷ Either　　是(第一个)〈—〉Either

或者(第二个)⟷ or　　是(第二个)〈—〉or

其他的典型结构还有：Not only…but also / Not only…but …as well 等等

11.11 被动句

如果在原文和目标译文中都出现了被动语态，那么做对齐很容易。然而对于一边是被动语气，另一边是主动语气的情况，词序非常不同，不容易将对应部分挑出来，要做出正确的对齐需要格外小心。

1) "by"有对齐部分，直接对齐，没有的话，多出来的"by"对齐到空

如：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.

由⟷ by

如：桥 被 洪水 冲垮 了 。

The bridge was destroyed **by** the flood.

被⟷ by

如：这 是 中央台 报道 的 。

This is reported **by** CCTV.

()〈～〉by

2) Be 动词在没有可对齐部分时，一般伪对齐到对应动词上

如：能否 妥善 处理 台湾 问题

Whether or not the issue of Taiwan can **be** appropriately dealt with

处理〈---〉be

如：桥 冲垮 了 。

The bridge **was** destroyed.

冲垮〈---〉was　冲垮↔destroyed　了〈---〉destroyed

但有对齐部分时，直接对齐即可。

如：桥 被 洪水 **给** 冲垮 了 。

The bridge **was** destroyed by the flood.

给〈---〉was

如：四十 名 外国人 **被** 挟为 人质

40 foreigners were **being** held hostages

被〈---〉being　被〈---〉were(表进行时态，与 being 粘合)

11.12 从句

当句子被置于层次结构中，最重要的是主句，次重要的被叫作从句，因为它们不能独立存在。一个从句经常被一个关系标记(从属连词或关系代词)所引导。

从属连词有：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 等，在汉英对齐中，这些连词通常可以找到与它们对等的词汇。

但是，对于关系代词，很难总是在目标语或源语中找到与其匹配的部分。关系代词有：that, which, whichever, who, whoever, whom, whose, whosever, whomever 等。

汉语的关系代词经常被省略，也可能以其它形式呈现出来，比如以"的"字结构的形式。

按照功能，从句可充当主语、宾语、补语和状语。

1) 主语从句

多出来的关系标记可以对齐到空。

如：他 是 学生 是 个 事实 。

That he is a student is a fact .

()〈～〉that

2) 宾语从句

在汉英词语对齐中，多出来的关系标记可以对齐到空或伪对齐到源语言的逗号。

如：他 指出 这个 问题 关系 到 中国 的 主权

He pointed out **that** the this question is a matter of Chinese sovereignty .

()〈～〉that

如：他 指出 ，这个 问题 关系 到 中国 的 主权

He pointed out **that** the Taiwan question is a matter of Chinese sovereignty .

，〈---〉that

3) 状语从句

关系标记经常可以找到它们的对等部分。

如：**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 .

因为⟷ because

4) 定语从句

1'"的"字结构：将关系标记伪对齐到"的"上。

如：你们 是 我 今年 会见 **的** 第一 个 美国 国会 众议员 代表团

They are the first US Congressional delegation **that** I have met this year

的〈---〉that

如：特别 是 在 美 中 两 国 经济 持续 增长 **的** 情况 下 。

especially at a time **when** the economies of both countries are experiencing sustained growth .

的〈---〉when

2' 非"的"字结构：

• 关系标记有对等部分：

如：防止 霍乱 、痢疾 等 疫情 爆发，**这些 疾病** 可能 造成 数以万计 的 人 死亡 。

prevent the outbreak of diseases such as cholera and dysentery , **which** have the potential to kill tens of thousands of people .

这些〈---〉which 疾病〈---〉which

• 关系标记没有对等部分：

关系标记要粘合到它们的先行词上进行伪对齐。特殊情况下，非限制性定语从句的先行词为整个主句，当关系标记 which 找不到对等部分时，可将它对齐到空。

如：印度 可能 会 在 选 后 出现 一个 成员 庞大 复杂 的 联合 **政府**，导致 未来 五 年 的 执政 困难重重

India could see a big and complex coalition **government that** would face difficulty in ruling for the five-year term to come .

政府⟷ government 政府〈---〉 that

如：该 委员会 将 在 基本法 实施 **时** 设立

the committee will be set up at the **time when** the basic law goes into force

时⟷ time 时〈---〉when

如：他 表示 **相信** 她 的 访问 会 推动 友好 关系 的 发展（宾语从句）

He expressed his **belief that** her visit will push forward the development of friendly relations（定语从句）

相信⟷ belief 相信〈---〉that

如：一个 中年 女子 杀害 了 自己 的 丈夫 ，令 我 十分 恐惧 。

A middle-aged woman killed her husband, **which** frightened me very much

which〈---〉()

注意：定语从句前后的逗号当没有对等部分时，也要伪对齐到"的"字或者先行词上。

如：这 位 在 华 工作 已经 一 年 多 **的** 大使 说

the ambassador , **who** has been working in china for more than one year , said

的〈---〉,（第一个） 的〈---〉who 的〈---〉,（第二个）

如：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.

马尼拉⟷ Manila 马尼拉〈---〉,（第二个） 马尼拉〈---〉where

对比：

如：比如 通货膨胀率 还 比较 高 ，为 百分之六 。

They included the higher inflation rate , **which** reached 6 percent .

通货膨胀率〈—〉 inflation 通货膨胀率〈—〉rate

,⟷, 通货膨胀率〈---〉 which

5）同位语从句

同位语从句的关系标记也要粘合在它们的先行词上进行伪对齐。

如： 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.

立场⟷ position 立场〈---〉that

11.13 标点

所有对等的标点都可以对齐，比如逗号对逗号，句号对句号等，多出来的标点可以对齐到空。

中文中的特殊情况如下：

1）：⟷,（强对齐）

如： 他 说 ："……"。

He said , "…".

：⟷,

2）,⟷. ()（强对齐）

一边是逗号，另一边是句号，将逗号强对齐到句号。

3）","〈---〉"and"（或者反过来）：请参看 11.10.1 小节

4）「⟷" 和 」⟷"（强对齐）

如：「消息显示，他们可能试验蓖麻毒气。」

"Information indicated they might have experimented the toxic gas of ricin . "

「⟷" 」⟷"

5）英文缩写词后紧跟的点，如果所起的作用是表明其前的单词是缩写的话，请确保缩写部分和这个点切分成一个词，然后对齐到中文的对等部分。更特殊的：",**etc** .":

如：经 香港 **等** 地 转口 到 美国 **等** 传统 市场 。

transferring from places like Hong Kong , **etc** . to traditional markets such as the US **etc** .

等（第一个）〈---〉， 等（第一个）⟷ etc. （"."先合并再对齐）

等（第二个）⟷ etc （etc. etc 都可以表示"等"）

。⟷.

6）各种标点〈～〉()（或者反过来）

对一边出现了标点，另一边标点消失的情况，多出来的标点符号可以对齐到空。

如：新华社 23 日 电 。（ 记者 ：王玮 ）

Xinhua News Agency, 23rd, by reporter Wang Wei.

(〈～〉() :〈～〉())〈～〉()

如：他 对 菲律宾 政府 坚持 " 一个 中国 " 的 立场 表示 感谢

He expressed gratitude for the support of the Philippine government towards the one China stance

"〈～〉() "〈～〉()

参考文献：

[1] Linguistic Data Consortium. Guidelines for Chinese-English Word Alignment [EB/OL]. Version 1.1. http://projects. ldc. upenn. edu/gale/Alignment/specs/

GALE_Chinese_alignment_guidelines_v1. 1. pdf, 2006.

[2] Melamed, D. Annotation style guide for the Blinker project [EB/OL]. Version 1. 0. 4. IRCS Technical Report # 98-06. University of Pennsylvania, Philadelphia. http://arxiv.org/abs/cmp-lg/9805004, 1998.

[3] Jean Véronis. ARCADE Tagging guidelines for word alignment [EB/OL]. Version 1. 0. http://aune.lpl.univ-aix. fr/projects/arcade/2nd/word/guide/index.html, 1998.

[4] 朱德熙. 语法讲义[M]. 北京：商务印书馆，1982.

北京大学计算语言学教育部重点实验室建设计划通过论证

2009 年 4 月 27 日，教育部科技司组织专家对“计算语言学教育部重点实验室”的建设计划进行了可行性论证。论证会由教育部科技司明炬处长和李武副处长主持，教育部语言文字信息管理司陈敏处长参加了论证会。

论证会专家组由清华大学张钹院士、中国中文信息学会常务副会长曹右琦研究员、中国科学院心理所杨玉芳研究员、中国科技信息研究所王惠临研究员、社科院语言所顾曰国研究员、教育部语用所冯志伟研究员、北京语言文化大学张普教授组成。

专家组听取了王厚峰教授代表实验室所做的建设计划报告，并针对实验室的研究方向、队伍建设、条件建设、运行管理机制等方面的建设计划进行了论证。专家组经过认真讨论，一致认为计算语言学重点实验室建设计划合理，目标明确，符合教育部重点实验室建设要求，同意通过建设计划论证；同时建议实验室进一步加强跨学科、跨部门之间的合作。

教育部 2009 年 1 月批准建设的“计算语言学教育部重点实验室”由北京大学承建。实验室研究人员由北京大学信息科学技术学院计算语言学研究所以及中文系、软件与微电子学院语言信息工程系、心理系、计算机技术研究所和外语学院的相关研究人员构成。

计算语言学教育部重点实验室将围绕如下五个方向开展研究：

(1) 中文计算的基础理论与模型；

(2) 大规模多层次语言知识库构建的方法；

(3) 国家语言资源整理与语音数据库建设；

(4) 海量文本内容分析与动态监控；

(5) 多语言信息处理和机器翻译。

计算语言学教育部重点实验室的成立是我国计算语言学发展史上的一件大事，对中文信息处理事业的发展将产生积极的影响。

北京大学计算语言学研究所　王厚峰供稿