汉英篇章结构平行语料库对齐标注的难点与对策

冯文贺[1]　李艳翠[2]　周国栋[3]

1.河南科技学院 中文系　新乡 453003 ；2.河南科技学院 信息工程学院 新乡 453003 ；3.苏州大学 计算机学院 苏州 215006

E-mail: {wenhefeng,yancuili}@gmail.com；{gdzhou}@suda.edu.cn

摘要：汉英篇章结构平行语料库是标注了对齐篇章结构信息的汉英平行语料库，对齐标注是其核心工作，基本原则是“结构对齐，关系对齐”。由于汉英双语差异等，在篇章单位切分对齐、结构对齐、关系对齐、中心对齐等不同层面的对齐标注工作中都有一些难点问题，该文基于标注实践，系统总结各个层面上的主要难点问题，并给出相应解决策略。实践表明，该研究对于提高对齐标注的质量和效率有重要意义。

关键词：汉英篇章结构平行语料库；对齐标注；结构对齐；篇章结构

Difficulties and Countermeasures of Structural alignment annotation of Chinese - English Discourse Treebank

Wenhe Feng[1]　Yancui Li[2]　Guodong ZHou[3]

1.Department of Chinese Language and Literature,Henan Institute of Science and Technology, Xinxiang 453003, China;2. School of Information Engineering, Henan Institute of Science and Technology, Xinxiang 453003; 3.Department of Computer Science and Technology, Soochow University, Suzhou 215006;

E-mail: {wenhefeng,yancuili}@gmail.com；{gdzhou}@suda.edu.cn

Abstract: *Chinese-English Discourse Treebank (CEDT) is a parallel corpus annotated with alignment discourse structure information for Chinese and English. Its core task is alignment annotation under the basic principle of structural and relational alignment. Because of the Chinese-English bilingual differences, there are some difficult issues in the annotation of discourse segmental, structural, relational and central alignment. Based on annotation practices, this paper summarizes the major difficult issues on all levels, and proposes the corresponding solution strategy. Practice shows that the corresponding difficulty Countermeasures can effectively improve the quality and efficiency of the corpus annotation.*

Key words: *Chinese-English Discourse Treebank; alignment annotation; structural alignment; discourse structure*

1. 引言

汉英篇章结构平行语料库（Chinese–English Discourse Treebank,简写为 CEDT）是对具有对译关系的汉英双语文本标注了对齐篇章结构信息的语料库[冯文贺，2013]。例 1 给出了一个汉英篇章结构的对齐标注文本，图 1 给出了其对应的图结构。

例 1. 少年姓孙，[A]//[并列]属马，[B]/[并列]比小水小着一岁，[C]///[并列]个头也没小水高，[D]//@[转折]人却本分实诚[E]。(贾平凹《浮躁》)

This boy, a member of the Sun family,[1] //@[并列]had been born in the year of the horse. [2]@/[并列]Although he was a year younger[3] @///@[并列]and a head shorter than Water Girl,[4] //[转折]he was honest and sincere[5] . (Goldblatt, 1991)

（说明：例中上标的字母和数字分别标明汉英小句及其顺序，“/”多少表明篇章结构层次高低，篇章关系用[ ]标记，连接词用下划线标记，@标明每一个关系中心项所在的位置）

可以看出，这种对齐不仅仅是一般的语言单位对齐，它在要求语言单位对齐的同时也要求语言结构对齐。结构对齐是汉英篇章结构平行语料库的核心理念。标注了结构对齐信息的双语篇章结构语料库可以为机器翻译等提供较为直接的双语篇章结构转换知识。

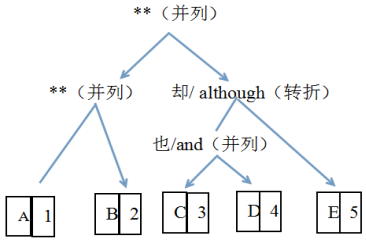

图 1：例 1 的汉英篇章结构对齐标注

（说明：箭头指向关系中心项，表示无显式连接词）**

现有的一些汉英平行语料库工作[柏晓静等，2002；王克非，2004；刘泽权等，2008]，一般仅进行段落、句子等语言单位对齐，并不提供双语结构对齐标注信息，更无篇章层面的结构对齐信息，这使得其在汉英篇章结构对齐知识的提供上作用相当有限。而现有篇章结构语料库主要还是单语上的工作，英文的如修辞结构树库[Carlson *et al.*，2003]、宾州篇章树库[Prasad *et al.*，2008]，汉语的如财经篇章修辞结构树库[乐明，2008]、宾州篇章模式树库[Xue，2005；Zhou & Xue，2012]、哈工大篇章树库[张牧宇等，2014]、连接依存篇章树库[Li *et al.* ，2014]等，这些工作篇章结构标注体系不尽一致，也没有基于互译关系的平行文本，由此，难于提供直接的汉英篇章结构转换知识。可以说，至今双语对齐篇章结构知识资源还相当匮乏，这也直接制约了基于篇章结构的机器翻译等研究的进展。在这样的背景下，汉英篇章结构平行语料库的对齐标注具有十分重要的理论和实际意义。

结构对齐是 CEDT 的核心思想。对齐标注基本原则是“结构对齐，关系对齐”，基本依据在于具有对译关系的篇章，其内部的层次结构和结构关系也一一对应。本质上篇章结构是一种逻辑语义结构，对于一个优质的翻译文本，源语中的因果、转折、并列等逻辑语义关系必然在目的语中得到反映，而且该逻辑语义关系的结构层级等也会得到较好反映。所以“结构对齐、关系对齐”本质上是逻辑语义结构对齐。

如图 1 所示，CEDT 整体采用类修辞结构的连接依存树篇章模式[li *et al.*，2014]，设计了统一的标注体系同时应用于双语标注。依据篇章结构标注任务，相应对齐标注工作主要包括篇章单位切分对齐、结构对齐、关系对齐、中心对齐等，[冯文贺，2013]提出了这些对齐标注的一般性策略，但对于其具体可行性还没有进行进一步研究。由于汉英两种语言属于不同语系，存在众多差异，再加上翻译者和标注者的翻译及理解主观性等，都给对齐标注带来一些难点问题。正确认识和合理解决这些难点问题就成为汉英篇章结构平行语料库创建中的关键问题。本文以[冯文贺，2013]的双语篇章结构对齐标注策略为理论基础，结合标注实践，系统总结和分析汉英篇章结构对齐标注中的难点问题，并给出了相应的解决方案。

2. 切分对齐的难点与对策

切分对齐是指语段该不该切分、在何处切分的问题，它要求汉英对译文本的切分位置保持一致。切分对齐是标注工作的基础，切分得当层次结构才可能准确对齐，关系对齐和中心对齐才能更好实现。切分对齐的关键在基本篇章单位（称为“子句”）对齐，对此，CEDT 一般采用“汉语优先”的策略，即原则上总是保证汉语的最小切分是一个篇章单位，并以汉语为标准对应得到相应的英语切分。例如：

例 2. 消化要素成本上涨压力，/正确引导市场预期，//***坚决抑制价格上涨势头。***

We need to cushion the upward pressure for costs of factors of production,/ and correctly guide market expectations// ***to resolutely curb price rises.***（以下例子，除特别注明，均出自《2011

年政府工作报告》汉英双语版）

例 2 中汉语根据子句标准[李艳翠等，2013]可切分出三个子句，相应的英语也切分为三个子句。注意，其中英文“to”引导的结构并非是一个典型的小句，这一篇章单位是在“汉语优先”的策略下对齐切分而来的。这种策略可以保证结构对齐总是在篇章结构范畴内。由于汉英差异的复杂性，切分对齐仍有一些难题。

2.1 汉语句法结构转化为英文篇章结构问题

汉语的句法结构有可能转化为英文篇章结构，而汉语的句法成分会转化为英文一般子句形式等，这时按照“汉语优先”有可能带来英文整体结构性错误。例如：

例 3. “十一五”前期，针对投资增长过快、贸易顺差过大、流动性过剩，以及结构性、输入性物价上涨等问题，//*采取正确的政策措施，/有效防止了苗头性问题演变成趋势性问题、局部性问题演变成全局性问题。*

In the early stages of the Eleventh Five-Year Plan period, we adopted correct policies and measures //to address overheated investment growth, the excessive trade surplus, excess liquidity, and structural and imported inflation**; /*effectively prevented emerging problems from evolving into trends; and prevented problems in any one area from becoming general problems.***

例3’. “十一五”前期，针对投资增长过快、贸易顺差过大、流动性过剩，以及结构性、输入性物价上涨等问题，采取正确的政策措施，***/有效防止了苗头性问题演变成趋势性问题、/局部性问题演变成全局性问题***。（X）

In the early stages of the Eleventh Five-Year Plan period, we adopted correct policies and measures to address overheated investment growth, the excessive trade surplus, excess liquidity, and structural and imported inflation***; /effectively prevented emerging problems from evolving into trends;/and prevented problems in any one area from becoming general problems.***

例 3 中“局部性问题演变成全局性问题”在汉语中作为并列宾语的一部分出现，是句子的一个句法成分；相应英译部分“and prevented…… general problems”在英语中却是复杂句的一个子句，与其它子句呈现明显的并列关系，整体呈现为“；；and”的典型并列模式。按照“汉语优先”标准，英语的最后一个并列子句不能切分，如此，英语最后一个并列子句将没用“and”连接，这显然不合英文规范。

一般而言，双语篇章结构标注十分注重标点符号、连接词等显性标记模式的提示作用，在这种情况下，我们曾试图按照“英语优先”来对齐切分汉语，将得到例 3’。然而，如此一方面将切分出“局部性问题演变成全局性问题”这样的汉语非子句单位，另一方面汉语的子句单位“针对投资……等问题”也将不能得到切分，这样就会破坏整个篇章结构分析。可以说，相比“汉语优先”，“英语优先”的切分将带来更多问题。考虑到这种情况在汉-英翻译中是少数情况，我们最终仍采用“汉语优先”的切分对齐策略，即采用例 3 的切分。

就双语转换事实而言，这里反映的问题是，汉语的句法结构转化为英语的篇章结构的情况。为简化对齐分析中的问题及矛盾，汉英篇章结构平行语料库暂不正面解决这一问题。在汉英篇章结构平行语料库中，“汉语优先”在根本上是为了保证汉语的每一个篇章结构问题都可以得到对应的英语结构转换。而对于英语的每一个篇章结构问题转化为汉语的情况，将在进一步的英汉篇章结构平行语料库（语料为英-汉翻译方向）研制中进行系统性解决，届时将按照“英语优先”来对齐切分汉语。

2.2 一些传统汉语“介词短语”结构的语法性质问题

汉语介词由动词演化而来，一些介词还有一定词汇意义，如“以”、“通过”、“针对”等，由这些词引导的“介词短语”被逗号切分后，具有相当强的成句性，可以与相关子句发

生命题间关系。这种结构在单纯汉语篇章结构子句认定时就有一定争议。对于这些结构，当其转换为英语时，也方便划分子句的，即确定为子句。例如：

例 4. ***我们一定要以对国家和人民高度负责的精神，//通过艰苦细致的工作和坚持不懈的努力，/***加快解决这些问题，//让人民满意！

We must therefore have a strong sense of responsibility toward the country and the people //***and work tirelessly and painstakingly*** /to solve these problems more quickly //to the satisfaction of the people.

例 4 中，汉语“以”和“通过”引导的句法结构，分别对应英文的两个主句结构，并以“and”连接。而汉语“加快…问题”和“让…满意”两句则对应英语的两个“to”结构。这种对齐切分是比较合理的切分，不仅考虑到了双语子句对应的情况，也考虑到了进一步的结构对齐、关系对齐及中心对齐的问题。

汉语“针对”等词引导的结构与其对应英语往往也很方便作类似的对齐切分，如：

例 5. ***针对工程建设、土地使用权出让和矿产资源开发、国有产权交易、政府采购等重点领域存在的问题，/***加大查处违法违纪案件工作力度，//坚决惩处腐败分子。

In regard to problems in key areas such as construction, sale of land-use rights, exploitation of mineral resources, trading of state-owned property rights, and government procurement, /we will intensify investigations and prosecutions of violations of the law or discipline //and resolutely punish corruptionists.

例 5 中，汉语“针对”引导的结构转换为英语的介词结构，而在汉英翻译中由汉语句子转化为英语介词结构的情况是很普遍的情况。这种分析符合汉英篇章结构转换的一般情况。

切分对齐以“汉语优先”为基本策略，对于个别汉语子句标准不清晰的问题，如介词结构问题的判断问题，从根本上看，是从汉英篇章转换的角度进一步明确了汉语子句的标准。

3. 层次结构对齐的难点与对策

层次结构对齐指汉英篇章层次结构的判断要一致。层次结构是篇章单位间逻辑语义亲近关系程度的反映，又体现为篇章关系的管辖范围的大小。一般而言双语篇章层次结构会较好的对齐。这特别反映在较高层次的篇章结构上。双语篇章结构转换的关键在底层篇章结构的转换问题上（复句或复杂句内的结构转换问题）。对此，CEDT 原则上采取“英语优先”策略，本质上是因为这种策略可以反映双语的翻译结构，另一方面也因为在底层篇章结构上英语有较明显的形式标志。例如：

例 6. 加快建设国家创新体系，//实施知识创新工程和技术创新工程，//突破了一批产业发展急需的前沿技术、核心技术和关键装备技术，/一大批科研成果实现了产业化。

We accelerated the development of the national innovation system**;** //carried out knowledge innovation projects and technology innovation projects**;** //**and** made breakthroughs in urgently needed cutting-edge technologies, core technologies and key equipment technologies**.** / A large number of research results have been applied in industrial production.

例 6 汉语是一个复句，内部子句间均以逗号隔开，不便从形式上判断其内部的层次结构。而相对应的英文分别以句号和分号及连接词“and”分割和连接，根据这些结构形式很容易对英语进行结构分析，进而对齐切分到汉语的层次结构。不过，本质来看这种结构分析也反映了双语翻译结构。由于双语差异等，“英语优先”策略下依然存在一些难点问题。

3.1 汉语有明显层次结构标记问题

汉语句子带有可以显示篇章结构的某种标记，然而相应英语又呈现另一种结构标记，并且提示其结构与汉语层次结构不同。这时会给双语结构对齐分析带来一定困扰，例如：

例 7. 现在，我代表国务院，//向大会作政府工作报告，///**请**各位代表审议，/**并请**全国政协委员提出意见。

On behalf of the State Council, //I now present to you my report on the work of the government ///for your deliberation and approval**.** /I **also** invite the members of the National Committee of the Chinese People's Political Consultative Conference (CPPCC) to submit comments and suggestions.

例7' 现在，我代表国务院，//向大会作政府工作报告，/**请**各位代表审议，//**并请**全国政协委员提出意见。

On behalf of the State Council, //I now present to you my report on the work of the government /for your deliberation and approval**.** //I **also** invite the members of the National Committee of the Chinese People's Political Consultative Conference (CPPCC) to submit comments and suggestions. （X）

单看汉语，"请……，并请……"格式中既有连接词"并"提示并列关系，又有共同的谓语动词"请"引出的兼语结构句，很容易认定该格式的并列关系，并处理为例 7'。然而，这种分析与英文结构有不同，英文中"I also ……"前的句号和该句中的"also"都提示了另外的结构分析，即例 7 的分析。注意，直观上汉语的连接词"并"连接的是两个汉语子句，而英文的"also"连接的是两个英文句子。这时会给结构对齐标注造成一定困扰。

有时，汉语有分号和明显排比句式，呈现显著的并列结构模式，而英语翻译在结构上又完全不同，分析困扰就更大，例如：

例 8. 我们战胜各种严峻挑战，靠的是发展；//各领域取得的一切成就和进步，靠的是发展；/解决前进道路上的困难和问题，仍然要靠发展。

We have relied on development to overcome all types of severe challenges,// and all our achievements and progress in every area come from development. /We must therefore continue to rely on development to resolve the difficulties and problems on the road ahead.

例 8'. 我们战胜各种严峻挑战，靠的是发展；/各领域取得的一切成就和进步，靠的是发展；/解决前进道路上的困难和问题，仍然要靠发展。

We have relied on development to overcome all types of severe challenges,/ **and** all our achievements and progress in every area come from development**.** /We must **therefore** continue to rely on development to resolve the difficulties and problems on the road ahead. （X）

按照汉语分号和排比句式的一般结构对应情况，很容易认定例 8'三个汉语子句为并列。然而对应的英文却采用了另外的结构形式，其中明显与汉语不同的是用一个句号切分出了两个句子，这就在结构上和汉语结构有了根本性不同。

对于上述情况，我们依然采用"英语优先"的结构对齐策略。根本依据是这可以反映双语的翻译结构，就例 7 来看，依据英语切分汉语，让汉语连接词"并"连接"现在……各位代表审议"和"请全国政协委员提出意见"，在内容上可能更有依据，因为前项主要是对人大代表的事情，而后项主要是对政协委员的事情。这样处理也较好的反映了"also"和"并"在管辖范围上的翻译关系。而如果按"汉语优先"的例 7'，让"also"连接介词短语"for your deliberation and approval"和句子"I invite……suggestions"是完全不合英语语法规范的。

3.2 英语的状语结构管辖问题

汉语的某些子句转化为英语的状语，英语状语结构的语义和语法结构不完全一致，在采用"英语优先"策略时，采用何种英语结构分析会给标注带来一定困扰。例如：

例 9. 抓紧建立保障性住房使用、运营、退出等管理制度，/提高透明度，/加强社会监督，//[目的]***保证符合条件的家庭受益***。

We will promptly establish an administrative system for the use, operation and return of low-income housing; /increase transparency; /and strengthen public oversight//[目的] ***to ensure that eligible families benefit from low-income housing***.

例 9'. 抓紧建立保障性住房使用、运营、退出等管理制度，//提高透明度，//加强社会监督，/[目的]***保证符合条件的家庭受益***。

We will promptly establish an administrative system for the use, operation and return of low-income housing; //increase transparency; //and strengthen public oversight/ [目的]***to ensure that eligible families benefit from low-income housing***.（X）

从语法结构上看，例 9“to”引导的不定式结构只是最后一个分句的状语，语法管辖决定了其只能在最后一个句内划分。但从语义关系看，“to”引导的不定式结构可以管辖整个前面三个子句，与前三个子句构成目的关系，相应的其结构地位要高，结构分析为例 9'。选择语义管辖还是语法管辖就成了结构分析的一个困扰。再看例 10。

例 10. 近两年，***面对百年罕见的国际金融危机冲击***，//我们沉着应对，/科学决策，/果断实行积极的财政政策和适度宽松的货币政策。

In the last two years, we responded coolly// ***to the impact of the global financial crisis - a crisis of a severity seldom seen in the last century,*** / made decisions scientifically /and resolutely followed a proactive fiscal policy and moderately easy monetary policy.

例 10 中，从汉语看，“面对”引导的小句也有管辖范围问题，即管辖其后的一个子句，还是三个子句。但与例 10 不同的是，从对应英文看，由于“to”引导的介词短语在第一个子句后面，语法上“to”引导的介词短语结构只与第一个子句发生关系，并且由于语法限制很难从语义上与后两个子句发生关系。这样就只有例 10 的分析才是唯一合理的方案。这种选择的实质是在“英语优先”策略下，进一步确定句法优先，而非语义优先。

参考例 10 后，我们进一步确定了前一例选择例 9 的分析，而非例 9'的分析。其依据在于，本质上这种结构分析体现了双语间的翻译结构关系，双语篇章结构平行语料库的根本目的在于服务于翻译，而不仅仅是一般的篇章语义分析。另外，这种选择也可以使得例 9、例 10 的获得统一的分析，因为一般它们都是状语管辖的问题，从而简化问题。

3.3 同类结构的层次问题

英语使用多个同类结构表达同类关系，而其间层次结构并不相同，这时结构对齐分析会有一定难度。例如：

例 11. 我们一定要以对国家和人民高度负责的精神，//[并列]通过艰苦细致的工作和坚持不懈的努力，/[目的]*加快解决这些问题，*//[目的]*让人民满意*!

We must therefore have a strong sense of responsibility toward the country and the people //[并列]and work tirelessly and painstakingly /[目的]***to solve these problems more quickly*** //[目的]**to the satisfaction of the people**.（多个 to 结构表目的）

例 12. 我们有效应对国际金融危机冲击，/保持经济平稳较快发展，//*胜利完成“十一五”规划的主要目标和任务*，/国民经济迈上新的台阶。

We effectively warded off the impact of the global financial crisis**,** //maintained steady and rapid economic development //[并列]***and fulfilled the major objectives and tasks of the Eleventh Five-Year Plan,*** /[并列]**and the economy scaled new heights.**（多个 and 表并列）

例 13. *要以经济和法律手段为主*，[1]//[并列]*辅之以必要的行政手段*，[2]/[条件]全面加强价格调控和监管[3]。

We need to comprehensively strengthen our work controlling and monitoring prices [3]/[条件]***mainly through economic and legal methods supplemented*** [1]//[并列]***by administrative means when necessary***[2].（多个介词结构表条件）

对于这类情况，需要综合语义、语法等多种因素，来确定相关结构的层次。例 11 中，从语法和语义上可以确定第一个“to”引导的是整个不定式结构“to solve……the people”，在第 1 个层次上，与前面的语段构成目的关系。而第二个“to”引导的介宾结构仅仅与语段“solve these ……more quickly”构成目的关系。

例 12 需要从并列结构的一般表达模式上来判断整体并列结构及其间两个“and”的结构地位。根据多项并列一定要在最后一项前使用 and，而前项可以使用逗号或分号连接的结构

模式，即“,/;……(,/;)and”，可以确定后一个“and”的并列前项为“We effectively……financial crisis”和“maintained steady……economic development”，由此其地位较第一个“and”的地位要高。而第一个“and”仅仅连接“maintained steady……economic development”和“fulfilled the ……Five-Year Plan”，其地位较后一个“and”要低。

例 13 的情况更为复杂，主要是英语的两个介词结构相对于汉语发生了较大语序结构变化。按照语法，英语的两个介词结构是逐层附加到主干部分的，但由于翻译时语段顺序改变，不能对齐进行汉语的结构分析。这种情况下，把两个介词结构划分为并列结构，共同修饰核心语段，以方便进行双语结构对齐分析。这种分析在语言事实上也有一定支持，这里英语的两个介词结构间也可以添加一个连接词“and”来连接二者。

4. 关系对齐的难点与对策

关系对齐指双语相应篇章结构间的因果、转折等篇章关系要标注一致。一般而言篇章关系是逻辑语义关系，具有客观性，双语篇章关系通常自然对应。但是，汉英语言结构的差异性，翻译者和标注者的主观性，都会带来双语篇章关系分析的差异。对此，整体上，CEDT 采取了“英语优先”的策略，本质上是因为这样可以体现翻译结构，另一方面也与英语有较多关系标记容易判断有关。如例 14 的英语关系词“thus”可以较好提示因果关系。

例 14.我们成功举办北京奥运会、上海世博会，/[因果]实现了中华民族的百年梦想。

We successfully hosted the Beijing Olympics and Shanghai World Expo, /[因果]**thus** fulfilling dreams the Chinese nation had cherished for a century.

4.1 关系词的一词多义问题

“英语优先”的关系对齐标注中一个重要依据是英语有关系词作为关系标记。然而，英语的关系词有相当一部分是多义的，其关系判定就有一定困难。最突出的是“and”，其多用于表并列关系，但也可以表示顺承、递进、因果、目的等多种关系。

例 15. 政府的一切权力都是人民赋予的，/[因果]必须对人民负责，//[并列]为人民谋利益，//接受人民监督。

All the government’s power is entrusted by the people,/[因果] **and** the government must **therefore** be responsible for the people,// work to benefit them, //[并列]**and** accept their oversight.

例 16. 改革开放是实现国家强盛、人民幸福的必由之路，/[因果]必须贯穿社会主义现代化建设全过程。

The only way to make the country strong and prosperous and to ensure the people’s happiness is through reform and opening up, /[因果]**and** we must implement them throughout the course of socialist modernization.

例 15 中，第一个“and”表因果关系，可以据“therefore”得到判断；而第二个“and”表并列，可以据“,,and”的并列模式得到判断。但对于例 16 则只能依据“and”连接的前后项内在的逻辑关系来判断了。

4.2 which 句的关系问题

英语中“which”一词引导非限制性定语从句，与主语之间属于主从关系，直观而言，可以将其归为解说关系，即定语从句是对主语的进一步说明。然而，“which”引导的从句，统一解释为解说关系，并对齐到汉语是有一定困难的。例如：

例 17. 制定并实施国家中长期科学和技术发展规划纲要,/中央财政科技投入 6197 亿元，///年均增长 22.7%，//[因果][*解说]***取得了一系列重大成果***。(注：用*表明错误的关系分析)

We formulated and implemented the National Medium- and Long-Term Plan for Scientific

and Technological Development./ The central government allocated 619.7 billion yuan for science and technology, ///an average annual increase of 22.7%, //[因果][*解说]***which resulted in a series of major achievements.***

根据英文的“resulted in …… achievements ”和汉语的“取得……成果”等，更适合认定相关语段间为因果关系，而非解说关系。对于“which”引导的从句和相关主句的关系，全面考察后，我们认为：“which”一词只是作为一个指代词，指代前面的某个短语或句子，而相关主从句间的关系则宜于根据其间的实际逻辑关系判断。

4.3 英语句法结构的关系问题

汉语的篇章结构有的转换为英语的句法结构，这个时候不便于从英语句法结构间的关系确定篇章关系，这样就成为一个难点问题。

例 18. ***国内生产总值达到 39.8 万亿元，//[条件] [*并列]年均增长 11.2%，***/财政收入从 3.16 万亿元增加到 8.31 万亿元。

GDP grew at an average annual rate of 11.2%//[条件][*并列] to reach 39.8 trillion yuan./ Government revenue increased from 3.16 trillion yuan to 8.31 trillion yuan.

从汉语看，两个子句之间的关系似乎是并列关系，然而英语“to reach 39.8 trillion yuan”明显在语法上不能与前面的语段构成并列关系，很难进行关系对齐。就语义上看，作为增长到的某种程度，“to reach 39.8 trillion yuan”与前面的语段可以有两种关系成立：因果关系和条件关系。这里选择条件关系，是因为：汉英文本中两个语段的顺序发生了变化，“at an average annual rate of 11.2%”（年均增长 11.2%）作为插入语出现在“GDP grew to reach 39.8 trillion yuan.”这个句子中间，表明以什么样的程度增长着，这个语段可以移至句首，将句子变成：At an average annual rate of 11.2%，GDP grew to reach 39.8 trillion yuan.这样更适合认定其间是条件关系。可以看出，对于这类问题，既需要考虑双语关系分析的适合性，还要考虑逻辑语义关系的各种可能性及其在句法关系上的适应性。这类问题判断是比较复杂的。

5. 中心对齐的难点与对策

中心对齐是指双语文本对相应的关系项的主次地位的判断要一致。CEDT 一般采取“英语优先”的策略，以英语的结构形式来判断关系项的主次。如：

例 19. 我们一定要以对国家和人民高度负责的精神，@//@[并列]通过艰苦细致的工作和坚持不懈的努力，@/ [目的]加快解决这些问题，//让人民满意！

We must therefore have a strong sense of responsibility toward the country and the people @//@[并列]and work tirelessly and painstakingly @/[目的]to solve these problems more quickly @//to the satisfaction of the people. （注：这里@标明中心所在的关系项）

目的语在句子中以状语成分出现，结构上与句子主干明显有主次之分，因此目的语在句中作为非中心项。类似判断还有：which 引导的定语从句、for 引导的目的语、with 引导的方式状语等。中心对齐分析主要是服务于汉英翻译的主从句及句法结构转化。由于双语等，在“英语优先”策略下，中心对齐标注也会遇到一些困难。

5.1 并列关系的中心问题

并列关系的关系项一般地位平等，即都是中心项。然而，也存在并列关系项并不平等的情况。例如：

例 20. 我们必须不断完善社会主义市场经济体制，/充分发挥市场在资源配置中的基础性作用，激发经济的内在活力，@//[并列]***同时，科学运用宏观调控手段，促进经济长期平稳较快发展。***

We must constantly improve the socialist market economy, /and make full use of the basic

role of the market in allocating resources to stimulate the internal vitality of the economy @ //[并列]***<u>while</u> using macro-control tools scientifically to promote long-term, steady and rapid economic development.***

英文“while”及相应汉语的“同时”都提示相关语段是并列关系，然而“while”引导的是分词状语结构，相比另一个主句结构的并列项，分词结构居于非中心地位。

一些特殊认定的并列关系也有类似情况，例如：

例 21．中央财政科技投入 6197 亿元，@//[并列]***年均增长 22.7%***，@/[因果]取得了一系列重大成果。

The central government allocated 619.7 billion yuan for science and technology,@ //[并列]***an average annual increase of 22.7%***, @/[因果]which resulted in a series of major achievements.

例 21 中的并列关系前后项具有明显的主次关系，“an average annual increase of 22.7%”只是一个名词性成分，结构上处于次要地位。但例 20 中名词性成分与前面的动词没有任何关系，删除后语法结构不会受到影响。

5.2 结构形式不明显问题

一些英语复杂句内的两个子句在形式上难于区分主次，通常认定两项同等重要。例如：

例 22. 过去五年，我们是一步一个脚印走过来的，@ /@[因果]中国人民有理由为此感到自豪！

We worked steadily and made solid progress, @/@[因果]and the Chinese people have every reason to take pride in this.

而对于句子间关系（句群关系），对于英语从形式上难于区分主次也是普遍的情况。对此，按照单语上的中心分析原则处理即可，即对于一个关系看哪个关系项可以代表所在整体与外界发生关系，便认定其为中心项[冯文贺，2013]。

6. 尚不能对齐的一些情况

6.1 无对应篇章单位

例 23. 与此同时，政府听到社会人士对外籍家庭佣工雇主缴付的雇员再培训征款（“外佣征款”）有不同意见。<u>**我决定豁免缴付“外佣征款”的安排**，于今年 7 月 31 日届满后，*取消征收“外佣征款”*，减轻雇用外佣家庭的负担</u>。

Meanwhile, the Government notes that there are different views in the community on the Employees Retraining Levy imposed on employers of foreign domestic helpers (FDHs). To ease the burden on families employing FDHs, I have decided to abolish the FDH levy when the suspension of its collection expires on 31 July 2013. （香港特区 2013 年梁振英施政报告）

汉语“我决定我决定豁免缴付‘外佣征款’的安排”，在英语中并无对应的篇章翻译单位。对于这种情况，不能进行双语对齐分析。由此，也不再对汉语进行此篇章单位的切分。这种情况主要是翻译者进行了一定语义整合，如例 23 中即整合了“我决定我决定豁免缴付‘外佣征款’的安排”和随后的“取消征收‘外佣征款’”。

根据我们的实践，无对应翻译单位的情况在公文、法律、新闻题材中都非常少，在一些文学翻译中略多，主要是因为采用了意译的方法。例如：

例 24. 这位老向导就住在西山脚下，/早年做过四十年的向导，//胡子都白了，///<u>**还是腰板挺直，**硬朗得很</u>。（《香山红叶》）

This old man happened to live right at the foot of the West Hill. / He had worked as a tourist guide for forty years //and although he had become a man bearing a white beard,/// he was still quite strong.

6.2 无法进行结构对齐

首先，语序跨句调整导致无法结构对应。

例 25. 在医保计划方面，我们已成立了工作小组和咨询小组，//[A]负责就推行计划提出具体建议，[B]/包括研究提供合理而又必要的财务诱因或公帑资助，[C]///[例证]例如以税务减免方式鼓励市民购买医疗保险，[D]//[目的]**以配合推行计划**[E]。

A working group and a consultative group have been set up[1] //to make specific recommendations on the implementation of the Health Protection Scheme (HPS)[2]. /They are also studying the provision of reasonable and necessary financial incentives or public subsidies[3] ///[目的]**to facilitate implementation of the scheme,**[4] //[例证]such as tax breaks, to encourage people to purchase health insurance[5].

例 26. 除了将九龙东打造为低碳小区以外[A]，我已责成环境局局长领导跨部门的督导委员会[B]，/加强部门间的协调[C]，议定具体实施策略及行动计划[D]，并与业界和持份者紧密交流合作[E]，**推动绿色建筑**[F]。

Apart from our plan to develop Kowloon East into a low-carbon community,[1] I have asked the Secretary for the Environment to lead an inter-departmental steering committee[2] **to promote green building**[3]**. /**The committee will strengthen the co-ordination among departments[4] to formulate implementation strategies and action plans[5], while maintaining close dialogue and co-operation with the relevant sectors and stakeholders[6]. （香港特区 2013 年梁振英施政报告）

例 25 中，汉语小句 E，对应英语小句 4，两者在双语中的语序和直接组合关系都有一定差异。英语中，小句 4 与 3 组合，构成目的关系，它们进而和 5 组合构成例证关系。但相应汉语中，小句 CD 组合构成例证，进而 E 才能与 CD 的组合，构成目的关系。这样双语的分析才是合理的分析，但如此，双语的结构就不能够对齐，这种不能对齐是由于双语中具有对译关系的单位有较大位置差异，由此相应篇章单位间的直接组合关系也有了一定差异。

例 26 中，汉语小句 F，对应英语小句 3，两个小句在双语中的语序和直接组合关系都发生了较大改变。而且，在英语中，小句 3 在第一个句子中，但在汉语中，相应的 F 却在整个句子的末尾，相对于在英语的第二个句子中。这样就很难进行二者的双语结构对齐分析。

其次，结构融合导致无法结构对应。

例 27. 计划可大幅减少整体汽车粒子排放物达八成，/氮氧化物排放也减少达三成。

The scheme will significantly reduce the overall emissions of particulates and nitrogen oxides by 80% and 30% respectively. （香港特区 2013 年梁振英施政报告）

例 27 中，汉语的并列结构，在英语中发生了融合性改变，导致不能切分出对应的篇章单位，进而也不能进行相应的结构对齐分析。

对于以上情况，目前我们采用的篇章结构机制还不能较好进行结构对齐分析，这是因为现有的篇章结构分析机制主要适合描写连续的组合关系和语序差异不大的情况。要想解决这些问题，还需引入新的篇章结构描写机制及其对齐分析机制。不过，汉英篇章的整体语序差异不大，以上所分析的结构不能对齐的情况，在我们分析的政府报告、法律、新闻文本中所占比例非常小，所以现有的篇章结构描写及对齐分析机制整体还是比较适合的。

7 结语

对齐标注是汉英篇章结构平行语料库的核心工作理念，“结构对齐，关系对齐”是对齐标注的基本原则，由于汉英双语差异及译者、标注者的主观性等，在切分对齐、结构对齐、

关系对齐、中心对齐等工作上，都遇到了一定难题，系统总结和解决这些问题就成了保证汉英篇章结构语料库质量的关键性问题。本文基于标注实践系统总结了这些问题，并提出了针对性解决办法。目前我们在对齐标注平台上手工标注了约 60 余字/词的汉英双语法律、公文、新闻及文学语料，并进行了对齐标注一致性和效率评估研究（另文），工作实践表明，通过系统总结这些问题和提出相应策略可以提高标注质量和效率。我们将在下一步的工作中进一步完善相关标注方案，给出标注结果及其分析，并提供公开的汉英篇章结构平行语料库。目前部分汉英篇章结构对齐标注语料，可致信 wenhefeng@gmail.com 获取。

致谢：国家自然科学基金“汉语篇章结构分析的资源建设与计算模型研究”（61273320）、教育部人文社科项目“汉英篇章结构平行语料库构建研究” (13YJC740022) 、中国博士后基金“基于依存结构和特征结构的篇章结构描写机制”（2013M540594）。苗孟华、温晓颖、原苏洁、林光伟、李巧丽、王文庆、曹登霞等参与了标注工作。

参考文献

柏晓静, 常宝宝, 詹卫东, 等.2002. 构建大规模的汉英双语平行语料库. 机器翻译研究进展——2002 年全国机器翻译研讨会论文集.

Carlson L., D.Marcu, and M E.Okurowski.2003. Building a discourse-tagged corpus in the framework of rhetorical structure theory . Jan van Kuppevelt and Ronnie W.Smith (eds.),Current and New Directions in Discourse and Dialogue, Kluwer Academic Publishers,pages85-112.

冯文贺.汉英篇章结构平行语料库的对齐标注研究.中文信息学报,2013(6):158-165.

Li Y, W.Feng. F.Kong,J.Sun,and G.ZH . Building Chinese Discourse Corpus with Connective-driven Dependency Tree Structure.In *Proceedings of EMNLP 2014*.

李艳翠,冯文贺,周固栋,朱坤华.基于逗号的汉语子句识别研究. 北京大学学报: 自然科学版, 2013 (1):7-14.

刘泽权,田璐,刘超朋.《红楼梦》中英文平行语料库的创建. 当代语言学, 2008, 10(4): 329-339.

Prasad R., N.Dinesh ,A. Lee , et al. 2008.The Penn Discourse Treebank 2.0.In Proceedings of the 6th International Conference on Language Resources and Evaluation.

王克非. 2004.双语对应语料库：研制与应用.北京：外语教学与研究出版社.2004

Xue N. 2005.Annotating discourse connectives in the Chinese Treebank.In Proceedings of the Workshop on Frontiers in Corpus Annotations II: Pie in the Sky. Association for Computational Linguistics, pages 84-91.

乐明. 汉语篇章修辞结构的标注研究. 中文信息学报, 2008, 22(4): 19-23.

张牧宇,宋原,秦兵,刘挺.中文篇章级句间语义关系体系及标注.中文信息学报,2014(2):28-36

ZhouY,and N.Xue. PDTB-style Discourse Annotation of Chinese Text.In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics*. 2012:69-77.