汉英篇章结构平行语料库的对齐标注研究*

冯文贺

(河南科技学院中文系，河南 新乡 453003；武汉大学计算机学院，湖北 武汉 430072)

摘要：篇章结构平行语料库是对具有对译关系的双语文本标注了平行篇章结构信息的语料库。对齐标注是汉英篇章结构平行语料库的核心理论基础。本文提出“结构对齐，关系对齐”的对齐标注策略，应用于切分对齐、层次结构对齐、关系对齐、中心对齐等环节，实现了对齐和标注并行、单位对齐和结构对齐共进的平行语料库工作模式。本策略辅之以相应标注平台和工作程序，及相应难点解决方案，被证明是一种高效的篇章结构平行语料库工作方式。

关键词：平行语料库；对齐标注；篇章结构

Alignment and Annotation of Chinese-English Discourse Structure Parallel Corpus

FENG Wenhe

(Department of Chinese Language and Literature, He Nan Institute of Science and Technology, Xinxiang, Henan 453003, China; School of Computer, Wuhan University, Wuhan, Hubei 430072, China)

Abstract: Discourse structure parallel corpus is a corpus annotated with parallel discourse structure information for bilingual text. This paper proposes such an alignment and annotation strategy, the structural and relational alignment, which is the theoretical basis of Chinese-English discourse structure parallel corpus. This strategy is applied to the corpus building process, including segmental, structural, relational, and central alignment, having achieved an operation mode of parallel corps along with alignment and annotation working together, as well unit alignment and structural alignment. The strategy with the help of corresponding annotation software and the solutions to the difficulties has been proved to be an effective operation mode for discourse structure parallel corpus.

Key words: parallel corpus; alignment; discourse structure

1 前言

篇章结构平行语料库是对具有对译关系的双语文本标注了平行篇章结构信息(含篇章单位和层次化结构及关系等)的语料库。例 1)给出了一个汉英篇章结构的平行标注文本。

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

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

平行语料库和篇章结构语料库近年来都有较大发展。国际上平行语料库 1990 年代以来快速发展，汉英平行语料库基本同步并取得较多进展[1-3]。然而整体上，现有汉英平行语料库除做了一般性段落、句子、短语对齐工作外，很少进行句法、语义等深度标注加工，特别是篇章结构的标注加工，还没见到相关工作。另一方面，国际上篇章结构语料库已有成熟工

* 基金项目：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) 、河南省教育厅人文社科规划项目“依存语法学流派研究”(2012-GH-080)。

作者简介：冯文贺（1976—），男，博士，博士后在站，研究方向：计算语言学，语言资源，语言理论。

作[4-6]，汉语方面也有一些理论探索和实践[7-9]，但至今未见到汉英(及其他双语)篇章结构平行语料库工作。篇章结构平行语料库的匮乏制约了基于篇章的机器翻译等技术的发展。我们在基本完成汉语篇章语料库 600 篇标注(CNDB1.0 版)工作基础之上，提出并开始汉英平行语料库的建设工作。本文的汇报基于已进行的标注实践。

对齐标注是汉英篇章结构平行语料库的核心理论基础。不同于一般平行语料库工作，它既要求单位对齐(篇章单位对齐)，还要求结构与关系对齐(篇章结构与篇章关系对齐)。不同于一般单语篇章结构语料库工作，它要在篇章结构标注同时考虑对齐问题。可以认为，汉英篇章结构语料库实质是对齐与标注合二为一的工作。由此，它富有挑战性和创新性；在机器翻译等领域将有独特应用价值，对于其他平行语料库工作也将有一定理论启示意义。

2 已有研究

首先，关于平行语料库的对齐和标注。就此问题，目前的平行语料库工作有以下主要特点：(1)理论上认为对齐和标注可以相对独立进行。通常对齐在前，然后单独进行各类标注，这也是平行语料库前期多对齐而少深层标注的原因。(2)对于对齐，多理解为单位对齐，如有段落、句子、小句、短语、词语等各级语言单位的对齐工作；一般不进行各层级的结构对齐工作。(3)由于标注独立于对齐，标注基本等同单语上的标注，并不考虑双语问题。

这种“对齐和标注相对独立，有单位对齐而无结构对齐”工作模式的形成，与理论上认为双语的语言结构，特别是句法结构有巨大差异有关，由此，不可能有对齐的句法结构，也不可能有对齐的词性标注等，这就从根本上造成了目前的工作模式。由于对齐和标注独立，又由于有单位对齐而无结构对齐，平行语料库不能高效指导后续的语言技术。比如，在基于结构转换的机器翻译中[10]，结构对齐和转换不能在现有平行语料库中得到高效指导。

这种工作模式在篇章结构平行语料库中可能得到改变。在汉英篇章结构平行语料库中，将实现“对齐和标注并行，单位对齐和结构对齐共进”。这主要与客观上篇章结构的双语差异可能没有句法结构差异那么大、那么精细有关。其次也与主观上语言学理论对于篇章结构的认识还没有那么根深蒂固有关。

其次，关于篇章结构语料库标注。虽然目前的篇章结构语料库主要是单语工作，但有关的基本篇章单位定义、结构分析、关系体系及标注等工作，仍可作为平行篇章结构语料库的重要基础。然而，由于要考虑双语对齐，特别是结构对齐，双语平行语料库对于基本篇章单位、结构分析、关系分析等将有一些特别考虑，某些标注可能会和单语上的工作有很大不同。由于双语对齐视野，对于篇章结构及其分析我们将会有一些不同认识。

3 汉英篇章结构平行语料库的对齐标注策略

3.1 对齐标注总原则

汉英篇章结构平行语料库的对齐标注总原则是“结构对齐，关系对齐”。例 1)即是此原则下的对齐标注，该例结构层次和篇章关系完全相同。关于这一原则有几点需要说明：

第一，本原则的基本假设是具有对译关系的篇章，其内部的层次结构和结构关系一一对应。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。所以这里的“结构对齐、关系对齐”本质上是逻辑语义结构对齐。

第二，本原则没有明确体现单位对齐，并不意味着没有单位对齐，因为单位对齐是结构对齐的必然结果之一。标注过程中，主要着力于从上到下的层层结构对齐，其间及最终自然带来各级篇章单位、直至最小篇章单位的对齐。

第三，本原则在实现双语结构对齐、关系对齐的同时实现标注。所以它实质上是一个“标注中有对齐，对齐中有标注”的对齐与标注合二为一的过程。

汉英篇章结构的对齐标注，包括切分对齐、结构对齐、关系对齐、中心判定对齐等几个关键对齐标注任务，下面分述它们的具体处理。

3.2 切分对齐

切分对齐指篇章单位对齐，它用来解决某一语段能否切分或切分到何处的问题。其关键是基本篇章单位对齐问题。基本篇章单位是篇章结构从上到下切分的终点(在从下到上的结构组合中是起点)。汉英语的基本篇章单位有重要差异，要给出一个同时适合两种语言的基本篇章单位定义，并用以工程实践是困难的。在这个问题上，我们采用“源语优先”的对齐策略，即首先按既定的汉语基本篇章单位进行切分，然后以英语对齐(最终可根据结果归纳英语基本篇章单位)。例 1)的切分对齐就是在这一原则下实现的。对于汉语基本篇章单位，我们采用了一个操作性强的标准[11]：

“子句是篇章分析的基本单位，含传统单句和复句中的分句。结构上，子句至少包含一个谓语部分，至少表达一个命题；功能上，子句对外不作为其他子句结构的语法成分，子句和子句间发生命题关系；形式上，子句间一定有标点分割，通常是逗号、分号和句号等。实际语料中，一些与典型子句在结构、功能、形式上类似的传统所谓短语在特定条件下也作为子句处理。”

需要指出，汉英基本篇章单位的差异主要在内部结构，其对外语义功能是一致的，即均与其他篇章单位发生命题间“因果、转折”等关系，而非发生句法成分之间的语义关系。从处理结果上看，这种对齐切分的结果表现为：

(1)双语文本都是典型基本篇章单位

典型基本篇章单位既具备一定结构要素，又具备特定功能要素。其中结构要素一般包含谓语部分，功能要素是对外发生命题关系。例 2-3)中对齐的基本篇章单位都比较典型[1]。

2) 中国是世界上历史最悠久的国家之一。/中国各族人民共同创造了光辉灿烂的文化，//具有光荣的革命传统。

China is a country with one of the longest histories in the world. /The people of all of China’s nationalities have jointly created a culture of grandeur//and have a glorious revolutionary tradition.

3) 一九一一年孙中山先生领导的辛亥革命，废除了封建帝制，//创立了中华民国。/但是，中国人民反对帝国主义和封建主义的历史任务还没有完成。

The Revolution of 1911, led by Dr.Sun Yat-sen, abolished the feudal monarchy//and gave birth to the Republic of China. / But the historic mission of the Chinese people to overthrow imperialism and feudalism remained unaccomplished.

(2)源语是典型基本篇章单位，目的语不是典型基本篇章单位

注意对照例4-5)中英文的划线部分的内部结构。

4) 人民依照法律规定，通过各种途径和形式，管理国家事务，/管理经济和文化事业，//***管理社会事务***。

The people administer State affairs /*and* manage economic and cultural undertakings //***and social affairs*** through various channels and in various ways in accordance with the provisions of law.

5) 在维护民族团结的斗争中，要反对大民族主义，//***主要是大汉族主义***，/也要反对地方民族主义。

In the struggle to safeguard the unity of the nationalities, it is necessary to combat big-nation chauvinism, // ***mainly Han chauvinism,*** /and to combat local national chauvinism.

3.3层次结构对齐

层次结构对齐要求双语的篇章层次结构分析一致。层次结构是篇章单位语义亲近程度的反映，具有一定客观性，通常双语的篇章层次结构会自然对应，如例2-3)。这种情况下各自独立标注双语，也会得到双语篇章层次结构对齐。但由于双语差异和篇章层次结构的理

---

[1] 下文各例均自《中国宪法》(中英文)，语料来源为中国人大网 http://www.npc.gov.cn/。

解主观性，目的语中会加入特定语言特征和翻译者的理解主观性，并进而影响目的语的层次结构。这种情况下，使用目的语优先原则进行层次结构对齐。对比例6-8)的 A、B 两种可能处理，其中 B 为目的语优先原则下的处理。

6) A．人民依照法律规定，通过各种途径和形式，管理国家事务，/管理经济和文化事业，/管理社会事务。

B．人民依照法律规定，通过各种途径和形式，***管理***国家事务，/***管理***经济和文化事业，//***管理***社会事务。

The people ***administer*** State affairs /***and manage*** economic and cultural undertakings //**and** social affairs through various channels and in various ways in accordance with the provisions of law.

7) A. 一九四九年，以毛泽东主席为领袖的中国共产党领导中国各族人民，在经历了长期的艰难曲折的武装斗争和其他形式的斗争以后，终于推翻了帝国主义、封建主义和官僚资本主义的统治，///取得了新民主主义革命的伟大胜利，//建立了中华人民共和国。/从此，中国人民掌握了国家的权力，//成为国家的主人。

B. 一九四九年，以毛泽东主席为领袖的中国共产党领导中国各族人民，在经历了长期的艰难曲折的武装斗争和其他形式的斗争以后，终于推翻了帝国主义、封建主义和官僚资本主义的统治，//取得了新民主主义革命的伟大胜利，//建立了中华人民共和国。/从此，中国人民掌握了国家的权力，//成为国家的主人。

After waging protracted and arduous struggles,armed and otherwise,along a zigzag course,the Chinese people of all nationalities led by the Communist Party of China with Chairman Mao Zedong as its leader ultimately, in 1949,overthrew the rule of imperialism, feudalism and bureaucrat-capitalism,//won a great victory in the New-Democratic Revolution //***and*** founded the People's Republic of China. /Since then the Chinese people have taken control of state power and become masters of the country.

8) A. 中国人民和中国人民解放军战胜了帝国主义、霸权主义的侵略、破坏和武装挑衅，/维护了国家的独立和安全，/增强了国防。

B. 中国人民和中国人民解放军战胜了帝国主义、霸权主义的侵略、破坏和武装挑衅，维护了国家的独立和安全，增强了国防。

The Chinese people and the Chinese People's Liberation Army have defeated imperialist and hegemonist aggression, sabotage and armed provocations /***and*** have thereby safeguarded China's national independence and security //***and*** strengthened its national defence.

这种处理在目的语中往往有形式标志。如例6)英文谓词 administer 和 manage 所引导的篇章单位首先构成第一层并列，而中文原有的后一个并列项为第二层并列，因为英文中后一个并列项与前一并列项共享一个谓词 manage。例7)中，逻辑上"终于……统治""取得……胜利"前二分句的关系比与后一分句"建立……共和国"的关系近一点，但对应英文采用",，and"一般并列结构的连接形式，故采用 B 的结构分析。而例8)，直观上中文的三个分句可构成并列，但对应英文采用的"and…and"并不是英文连接同层并列的一般方式，分析后可知，第一个 and 的地位要高于第二个 and，故相应结构划分采用 B。这种"注重形式，目的语优先"的层次结构对齐方式，有利于指导机器翻译中的结构转换等工作。

3.4 关系对齐

关系对齐要求双语对应结构的篇章关系类别判定要一致。篇章关系本质上是逻辑关系，由于逻辑关系的客观性，通常判定一种语言的篇章关系，同时运用于两种语言即可。不过，篇章关系的理解具有主观性，特别是翻译文本中会加入翻译者的主观理解，从而会影响到目的语。这种情况下按照目的语优先原则进行关系对齐。例9-10)所标记关系即为目的语优先原则下的对齐标注。目的语优先通常要求目的语有形式标志，例9)的连接词

“and…thereby”，例10)的“to”提示了相应关系。目的语优先的关系对齐有利于指导机器翻译的关系翻译等。

9)中国人民和中国人民解放军战胜了帝国主义、霸权主义的侵略、破坏和武装挑衅，/[**递进，因果**]维护了国家的独立和安全，//增强了国防。

The Chinese people and the Chinese People’s Liberation Army have defeated imperialist and hegemonist aggression, sabotage and armed provocations /**[递进；因果]*and*** have ***thereby*** safeguarded China’s national independence and security //and strengthened its national defence.

10)各少数民族聚居的地方实行区域自治，/设立自治机关，//行使自治权。

Regional autonomy is practised in areas where people of minority nationalities live in concentrated communities; /in these areas organs of self-government are established//**[目的]** ***to*** exercise the power of autonomy.

3.5 中心对齐

中心通常是关系项的主旨或重点，中心对齐要求双语文本对于关系项主次地位的判定一致。中心项的确定有客观性，但也有理解主观性，翻译中会加入翻译者的理解，进而影响目的语的语言结构，我们使用目的语优先原则进行对齐。这时候目的语一般有形式标志，如例11)下划线所示英文篇章单位的不定式形式提示该项在相应关系中的非中心地位，例 12)下划线英文篇章单位的名词短语限定形式、定语从句形式和主要谓语形式提示相应项的主次地位。采用目的语优先的中心对齐标注，对于机器翻译中主从结构转换等会有一定指导意义。

11)各少数民族聚居的地方实行区域自治，/设立自治机关，*//行使自治权。

Regional autonomy is practised in areas where people of minority nationalities live in concentrated communities; //in these areas organs of self-government are established//**to exercise the power of autonomy.** (注：这里用*标记相应层次结构的中心项，下同)

12) 中国人民政治协商会议是有广泛代表性的统一战线组织，*//过去发挥了重要的历史作用，/*今后在国家政治生活、社会生活和对外友好活动中，在进行社会主义现代化建设、维护国家的统一和团结的斗争中，将进一步发挥它的重要作用。

The Chinese People’s Political Consultative Conference, **a broadly based representative organization of the united front * //which has played a significant historical role**, /* **will play a still more important role** in the country’s political and social life, in promoting friendship with other countries and in the struggle for socialist modernization and for the reunification and unity of the country.

3.6 角色分布对齐

角色指篇章关系中关系项的角色地位，如因果关系中，一个关系项为“原因”项，一个关系项为“结果”项。角色分布指关系项的位置分布或顺序，比如汉语“因果关系”通常“原因”在前，“结果”在后。我们以汉语的角色分布常规作为角色分布的对齐标准。对于一个“原因”在后，“结果”在前的文本，无论中英文，均认为其“不合常规”。这种对齐对于机器翻译中的语序调整将起一定作用。

4 对齐标注实现

4.1 标注平台

为了获得高效、一致的标注，我们开发了一个汉英篇章结构的辅助对齐标注平台。实现的功能包括双语导入、篇章单位切分、层次结构标注、连接词标注、关系标注、角色分布标注、中心标注。标注平台工作界面见图 1。为了便于结果直观对比，中英双语的对齐标注均给出树图显示，见图 2，图式例子为例 3)。直观上双语篇章结构对齐，树图结构完全一致。

4.2 标注操作

为了保证对齐标注，我们制定了对齐标注操作流程规范。主要有：

第一，从汉到英，从英到汉，形式优先。从汉到英，指切分首先从汉语判定，以汉语为标准切分对齐，这主要与本工作是“汉-英”方向的平行语料库有关。从英到汉，指层次结构、篇章关系、中心等由英语而汉语进行判定，这一方面与英语有较多的形式结构可把握有关，另一方面也与这首先是一项服务于机器翻译的工作有关。

第二，从上到下，从左至右，步步对齐。从上到下，从左至右，指标注中层次结构的划分遵循从上到下、从左至右的结构切分流程，并且要求汉英篇章结构平行分析，步步对齐。

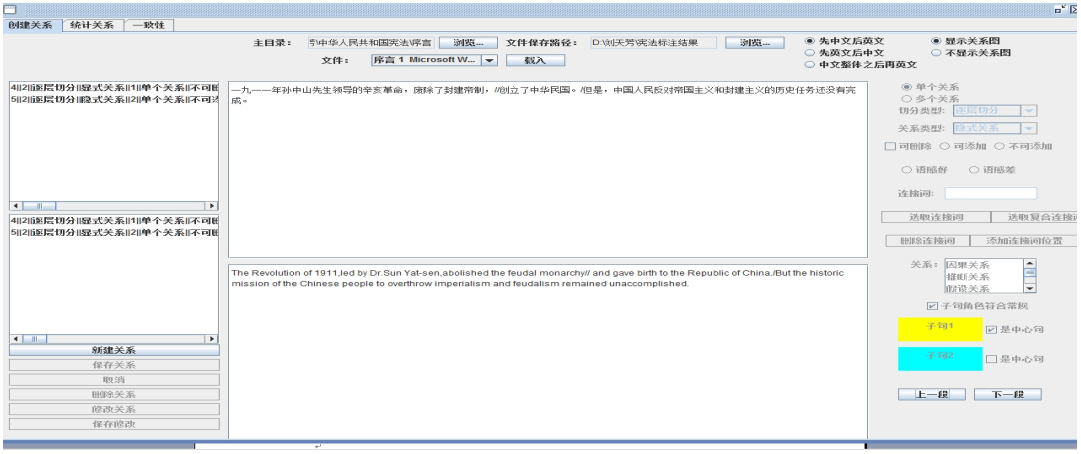

图 1：汉英篇章结构平行语料库标注平台界面

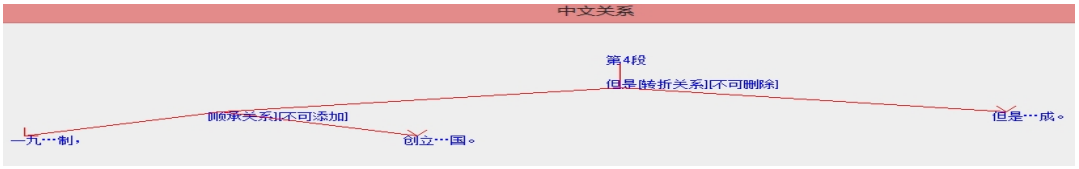

图 2 A：中文篇章结构标注结果直观图式

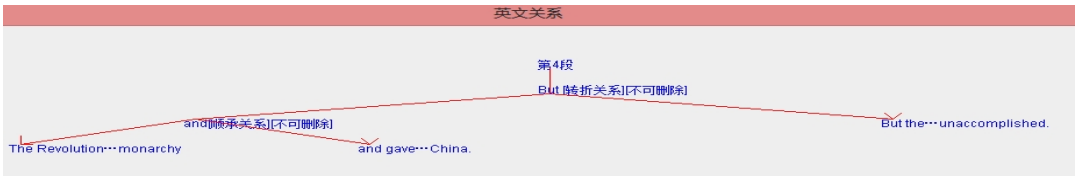

图 2 B：英文篇章结构标注结果直观图式

4. 3 标注结果

标注结果保存为 XML 格式，双语标注结果各自独立保存，二者的对齐关系通过段落号(P ID)、关系号(R ID)等体现。下面给出例 3)的标注结果实例。

中文标注实例：

<P ID="4">

<R ID="4" StructureType="逐层切分" ConnectiveType="显式关系" Layer="1" RelationNumber="单个关系" Connective="但是" RelationType="转折关系" ConnectivePosition="35…36" ConnectiveAttribute="不可删除" RoleLocation="normal" LanguageSense="true" Sentence="一九一一年孙中山先生领导的辛亥革命，废除了封建帝制，创立了中华民国。|但是，中国人民反对帝国主义和封建主义的历史任务还没有完成。 " SentencePosition="1…34|35…63" Center="2" ChildList="5" ParentId="-1" UseTime="20"/>

<R ID="5" StructureType="逐层切分" ConnectiveType="隐式关系" Layer="2" RelationNumber="单个关系" Connective="" RelationType="顺承关系" ConnectivePosition="" ConnectiveAttribute="不可添加"

RoleLocation="normal" LanguageSense="true" Sentence="一九一一年孙中山先生领导的辛亥革命，废除了封建帝制，|创立了中华民国。" SentencePosition="1…26|27…34" Center="3" ChildList="" ParentId="4" UseTime="72"/>

英文标注实例：

<P ID="4">

<R ID="4" StructureType="逐层切分" ConnectiveType="显式关系" Layer="1" RelationNumber="单个关系" Connective="But " RelationType="转折关系" ConnectivePosition="116…119" ConnectiveAttribute="不可删除 " RoleLocation="normal" LanguageSense="true" Sentence="The Revolution of 1911,led by Dr.Sun Yat-sen,abolished the feudal monarchy and gave birth to the Republic of China.|But the historic mission of the Chinese people to overthrow imperialism and feudalism remained unaccomplished. " SentencePosition="1…115|116…225" Center="2" ChildList="5" ParentId="-1" UseTime="25"/>

<R ID="5" StructureType="逐层切分" ConnectiveType="显式关系" Layer="2" RelationNumber="单个关系" Connective="and" RelationType="顺承关系" ConnectivePosition="76…78" ConnectiveAttribute="不可删除 " RoleLocation="normal" LanguageSense="true" Sentence="The Revolution of 1911,led by Dr.Sun Yat-sen,abolished the feudal monarchy| and gave birth to the Republic of China." SentencePosition="1…74|76…115" Center="3" ChildList="" ParentId="4" UseTime="14"/>

5 难点问题及其解决

5.1 基本篇章单位问题

对齐切分以汉语标准为优先原则，汉语切分中篇章结构和复杂句结构的区分是个难点。如例 13)，如果认为“在…以后，终于”是表顺承关系的连接词，可以认为划线部分就是一个基本篇章单位。不过，传统语法一般把其分析为状语，作为句法结构的一部分。这是篇章结构和句法结构有过渡地带的反映。我们暂按传统语法，把划线部分的分析留给句法结构。

13) 一九四九年，以毛泽东主席为领袖的中国共产党领导中国各族人民，***在***经历了长期的艰难曲折的武装斗争和其他形式的斗争***以后，终于***推翻了帝国主义、封建主义和官僚资本主义的统治，取得了新民主主义革命的伟大胜利，建立了中华人民共和国。

5.2 篇章关系问题

由于目前的篇章关系体系还不是一个严格逻辑体系，以及篇章关系理解的主观性，当缺少明显关系标记的时候，关系对齐标注就比较困难。我们采取两种策略解决这个问题。

第一，制定形式策略，保证篇章关系判定的客观性。常用方法如下。

添加连接词法：为当前关系添加某类关系的典型连接词，如果连贯顺畅，该关系可能即为当前关系的所属类别。如例14)通过添加“但是”测试，可以判定相应关系为转折关系。

提问回答法：用适合于某类关系的提问方式测定当前关系，如果当前关系的前后项比较适合该提问方式则认定当前关系即为该类关系。如例14)对前项提问“怎样区域自治”，而后项适合作为该项回答，可以认定当前关系为解释关系。

14)各少数民族聚居的地方实行区域自治，//**[解释](提问：“怎样区域自治？”)**设立自治机关，///**[目的](添加连接词：“以”)**行使自治权。**/[转折](添加连接词：“但是”)**各民族自治地方都是中华人民共和国不可分离的部分。

第二，允许多种篇章关系存在，但一般不超过三种。从不同角度看，可能同时存在多种关系。见例 15)。这即可减少关系判断的困难与分歧，也较真实的反映了篇章关系事实。

15) 平等、团结、互助的社会主义民族关系已经确立，**/[顺承;并列;递进]**并将继续加强。

5.3 中心问题

由于中心的理解主观性，在缺少一定形式标志的时候，中心对齐就成为困难问题，通过两种策略解决。

第一，制定形式策略，保证中心判定的客观性。通常可用删除法测试。见例 16)。

删除法：关系中的中心项不可删除，非中心项可以删除。二者的区别在于非中心项删除后句子仍然保持原有连贯关系，而中心项对外具有代表性，删除后不能保持原有连贯关系。

16) 各少数民族聚居的地方实行区域自治，*//~~设立自治机关，~~*///~~行使自治权~~。*/*各民族自治地方都是中华人民共和国不可分离的部分。

第二，允许多个中心存在。当无法利用形式标志和既定策略判定中心项的时候允许多个中心存在。如例 16)第一层前后项均为中心。另外，并列结构一般是多中心结构。

值得指出，以上的一些难点问题，大多是单语篇章结构标注中就存在的问题。

6 结语

对齐标注是汉英篇章结构平行语料库的核心理论基础，本文提出“结构对齐，关系对齐”的对齐标注策略，应用于切分对齐、层次结构对齐、关系标注对齐、中心对齐等环节，实现了“对齐和标注并行，单位对齐和结构对齐共进”的平行语料库构建模式。本策略辅之以相应工作平台和工作程序，和相应难点解决方案，被证明是一种高效的篇章结构平行语料库工作方式。下一步工作中，我们将不断完善本标注策略，进一步扩大标注实验，形成完整的对齐标注规范和其他相关篇章结构标注规范，最终研制一个大规模的汉英篇章结构平行语料库供学界和工业界使用。

致谢：常伟开发了辅助标注平台，郭海芳、王筱铮、王玉梦、胡炎磊参与了项目标注工作，李艳翠为项目设计和本文提出了宝贵意见。

参考文献

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

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

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

[4] Carlson L, Marcu D, Okurowski M E. Building a discourse-tagged corpus in the framework of rhetorical structure theory [C]// Jan van Kuppevelt and Ronnie W.Smith (eds.),Current and New Directions in Discourse and Dialogue, Kluwer Academic Publishers,2003,85-112.

[5] Wolf F, Gibson E. Representing discourse coherence: A corpus-based study [J]. Computational Linguistics, 2005, 31(2): 249-287.

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

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

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

[9] ZhouY, Xue N. PDTB-style Discourse Annotation of Chinese Text[C]//Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics. 2012:69-77.

[10] 刘群. 汉英机器翻译若干关键技术研究[M].北京：清华大学出版社.2008.

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

作者联系方式：冯文贺　河南省新乡市河南科技学院中文系 453003，　电话 15225901522，邮箱 459775647@qq.com, wenhefeng@gmail.com