汉语篇章依存结构的标注难点与处理*

冯文贺[1],徐钰仪[2],李青春[2]

（1.广东外语外贸大学语言工程与计算实验室，广州 510006；2.广东外语外贸大学中文学院，广州 510006）

摘要：篇章依存结构一般地表示为最小篇章单位（小句）间的支配与被支配关系，相比修辞结构等其可以有效刻画最小篇章单位间直接关系及其向心性。篇章依存结构的理论内涵及具体实践有待深入。本文结合汉语篇章依存结构语料库的标注实践，重点分析标注难点问题并给出处理方案，包括小句切分、小句关联、依存中心等重要分析任务。事实上，这些难点不仅是人工标注的难点，也是自动标注的难点，其处理不仅有利于指导构建高质量语料库，也有利于指导进一步的自动标注研究。

关键词：篇章结构；篇章依存结构；小句关联；中心

The difficulty and treatment of annotating the discourse dependency structure of Chinese text

FENG Wenhe[1]，XU Yuyi[2]，LI QingChun[2]

(1.Laboratory of Language Engineering and Computing, Guangdong University of Foreign Studies, Guangzhou 510006; 2. Faculty of Chinese Language and Culture , Guangdong University of Foreign Studies, Guangzhou 510006)

Abstract: The discourse dependency structure is generally expressed as the dominant relationship between the minimum discourse units (clauses), which can effectively depict the direct relationship between the minimum discourse units and its concentric nature compared with the rhetorical structure. The theoretical connotation and concrete practice of discourse dependency structure need to be deepened. Based on the annotating practice of Chinese discourse dependency structure corpus, this paper focuses on the analysis of the annotation difficult problems and gives the solutions, including clause segmentation, clause relevance, dependency head and other important analysis tasks. In fact, these difficulties are not only the difficulties of manual annotating, but also the difficulties of automatic analyzing. Their processing is not only helpful to guide the construction of high-quality corpus, but also to guide the further research of automatic analyzing.

Keywords: discourse structure; discourse dependency structure; clause relevance; head

1. 引言

篇章结构分析主要分析篇章单位间的结构关系，在自动摘要、自动问答及机器翻译等研究中有重要应用。比较著名的篇章结构表示模式主要有修辞结构[1,2]和宾州语篇树库模式[3]，前者将篇章表示为一颗层次化结构树，后者将篇章表示为一个个独立的篇章连接词的论元结构，其中论元大致相当于篇章单位。本质上，二者都是一种层次化结构，不同在于前者是完整篇章结构树，后者据连接词的论元辖域可推导出一定篇章结构树，但并不构造完整结构树。

* **收稿日期：** **定稿日期：**

基金项目：国家社科基金“汉语篇章结构的特征-依存描写机制及资源建设研究”（17BYY036）

层次化结构的问题在于不刻画最小篇章单位间的直接语义关联，特别是不能刻画非相邻和跨层级的最小篇章单位间的语义关联，而这种关联在篇章中客观而普遍存在。

研究者借鉴句法依存结构的思想，提出篇章依存结构[4,5,6]，为解决以上问题提供了形式表示的可能性。然而，篇章依存结构的深入内涵及具体语言分析实践如何，相关工作还不多见。而相应的篇章依存语料库工作或由修辞结构语料库简单转化而来[4]，或关系体系等略调面向特定领域[6]，且均非基于汉语。

本文根据文献[7,8]的“篇章特征-依存结构”理论机制，进行了汉语篇章特征-依存结构语料库的研制，其中篇章依存结构为篇章特征-依存结构的结构部分，另一部分为篇章关系特征结构部分（本文暂不涉及）。篇章依存结构的核心在于确定最小篇章单位间的支配与被支配关系，其基础任务包括小句切分、小句关联、中心确定等，本文根据我们的语料库标注实践，重点分析有关标注难点，并给出处理方案，以指导语料库标注。事实上，这些难点不仅是人工标注的难点，也是自动标注的难点，其处理不仅有利于指导构建高质量语料库，也有利于指导进一步的自动标注研究。

2. 汉语篇章依存结构语料库标注

2.1. 篇章依存结构

以篇章为对象，切分出基本篇章单位（小句，clause）作为节点，建立其间关联（小句关联），确定小句关联中心项（关联中心），辨明小句关联的语义内容（篇章关系），明确连接词对相应小句关联的结构与语义表征关系，便确定篇章依存结构[7,8]。

图 1 为例（1）的篇章依存结构图式，数字“1、2、3”等代表小句序列，其间关联由连线表示，连线的实心端指向关联中心，连接词从结构与语义上反映相应小句关联的结构关系与直观语义内容，而其抽象语义内容则由特征结构表示（本文暂不涉及连接词及其语义分析）。

（1）[一1]浦东开发开放是一项振兴上海，建设现代化经济、贸易、金融中心的跨世纪工程，[2]因此大量出现的是以前不曾遇到过的新情况、新问题。[二3]对此，浦东不是简单的采取“干一段时间，等积累了经验以后再制定法规条例”的做法，[4]而是借鉴发达国家和深圳等特区的经验教训，[5]聘请国内外有关专家学者，[6]**积极、及时地制定和推出法规性文件**，[7]使这些经济活动一出现就被纳入法制轨道。[三8]去年初浦东新区诞生的中国第一家医疗机构药品采购服务中心，正因为一开始就比较规范，[9]运转至今，[10]成交药品一亿多元，[11]没有发现一例回扣。【001】[1]

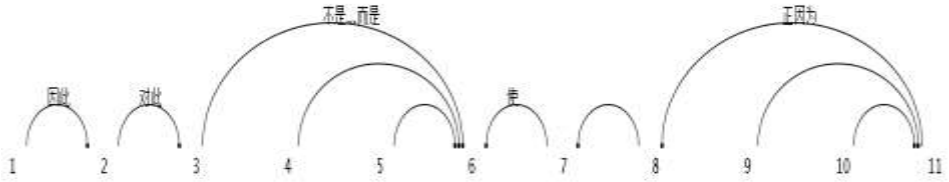

图 1　篇章依存结构[例（1）]

相比修辞结构等层次化结构，小句关联突破篇章单位的线性顺序和层级限制，准确刻画了基本篇章单位间直接关系，如 3-6、4-6、8-11、9-11 突破篇章单位线性顺序限制而直接关联，2-3、7-8 突破篇章单位不在同一大句（sentence。例中用上标的“一、二、三”等数字标记）的层级限制而直接关联。相应地，连接词则突破线性位置的连续性和单位层级限制，跨篇章单位，直接反映小句关联间结构关系，如“不是……而是”“正因为”分别反映非连续篇章单位 3-6、8-11 间关系，“对此”反映非同一大句的小句 2-3 间关系。而在语义上，连接词反映的小句关联关系，也比在层次化结构中反映的语义关系更具体、准确。在小句关联中心上，小句 6 因为与多个小句（3、4、5、7）直接关联及与一些小句相对间接关联，可推出其为相关关联对（3-6、4-6、5-6、6-7）的中心，进一步又可推导出其为全段中心。

[1] 【 】内标记例句在汉语宾州树库的篇章号，下同。

这种关联中心，相比传统依据关系对本身的关系类别确定的关系中心，更具全局性，也更具篇章本质。

2.2. 汉语篇章依存结构标注

根据文献[7,8]的有关篇章依存结构分析理论及标准，并基于相应标注平台[9]，在汉语宾州树库语料上[10]进行了汉语篇章依存结构语料库标注。以下根据标注实践，简要说明一般性处理准则，重点分析说明标注中遇到的难点，给出相应处理方案，以指导语料库标注工作。

3. 小句切分的难点处理

小句切分是篇章依存结构分析的第一步，其分析性能直接影响进一步的分析。汉语小句切分的一般标准主要采用文献[11,12]的标准：①结构上，小句一般是主谓结构，可以是完整主谓结构，也可以是不完整主谓结构。不完整主谓结构一般是主语省略，而省略主语能够明确补出。②功能上，小句一般可独立成句，其中一部分小句在结构上需增加或删除篇章衔接成分方可独立成为自然小句,小句与其他线性相邻的语言单位没有句法结构关系（主谓、动宾、状中、定中等），当前小句删除不影响相邻小句的成立。③形式上，小句间一般有停顿（书面上以逗号、句号等为标志），没有停顿的结构（含多动词短语）一般不作为小句。下列情况下，小句切分容易出现标注不一致，以下说明处理方案。

3.1. 主谓结构作句法成分不划分为小句

小句是主谓结构，但并非所有主谓结构都是小句。主要从功能上审查，主谓结构如果与相邻的语言单位有句法结构关系，不作小句切分。

3.1.1. 主谓结构作主语

(1) [产品、项目水平高，][2]是该区的重要特点。【011】

(2) [重大科技成果迅速转化为现实生产力，]是这个开发区的突出特点。

例（1）“产品、项目水平”是主语，“高”是谓语，形成完整主谓结构。但由于该主谓短语与相邻的“是该区的重要特点”构成句法关系，即主谓关系，并充当主语，对外不具有语法独立性，这个主谓短语不可划分为小句。例（2）的“重大科技成果迅速转化为现实生产力”也是与相邻的“是这个开发区的突出特点”构成主谓关系，充当主语，不划分为小句。

3.1.2. 主谓结构作定语

(3) [由浙江医科院院长、中国科学院士毛江森主持在世界上率先研究成功，]并具有国际先进水平的甲肝减毒活疫苗，去年经卫生部批准正式投入生产和使用，/[3]目前该区生产此疫苗的普康公司已形成年产五百万人份的生产规模，/这对有效地控制甲肝流行具有重大意义。【011】

(4) [这是上海海关为进一步推进市郊外向经济的发展，]继奉贤、莘庄、嘉定、松江、青浦、金山之后设立的第七个海关机构。【007】

例（3） 单独看，“由浙江医科院院长、中国科学院士毛江森主持在世界上率先研究成功”可视为一个主语（“甲肝减毒活疫苗”）省略的主谓结构，但该短语与相邻的“并具有国际先进水平”（也可视为主语省略的主谓结构）并列，并共同作它们之后的“甲肝减毒活疫苗”的定语，因此不具有语法独立性，不能划分为小句。

例（4）“这是上海海关为进一步推进市郊外向经济的发展”表面上也可以分析为主谓结构，但实际上，该语言片段中“这是”与后面的“第七个海关机构”构成真正的主谓结构，而该语言片段的主体“上海海关为进一步推进市郊外向经济的发展”作为主谓结构，与相邻其后的“继奉贤、莘庄、嘉定、松江、青浦、金山之后设立”（也可视为主语省略的主谓结构）并列，并共同作它们之后的“第七个海关机构”的定语，因此不具有语法独立性，不能划分为小句。

[2] 例中容易被错化为小句的部分用“[ ]”标出，下同。

[3] “/”标明小句切分，下同。

3.1.3. 主谓结构作介词宾语

(5) [[这对进一步改善崇明县的投资环境，][加快吸引外资，][方便快捷地办理海关手续，][把崇明建设成对外高度开放的大型贸易港口，][带动出口加工、航运中转等外向型经济的发展，]]将起到积极的作用。【007】

例（5）“进一步改善崇明县的投资环境”“加快吸引外资”“方便快捷地办理海关手续”“把崇明建设成对外高度开放的大型贸易港口”“带动出口加工、航运中转等外向型经济的发展”均可视为主语省略的主谓结构，它们合起来构成复句结构，各自独立来看，它们均可划分为小句，但从更大范围看，它们与相关成分有句法结构关系，即在语法上，它们作介词“对”的宾语，而整个介宾结构又充当相邻语段“将起到积极的作用”的状语，由此各个主谓结构及它们构成的复句结构并不具有语法独立性，也就不能划分出小句。

3.1.4. 主谓结构连带宾语

(6) [统计资料显示，]过去五年广西对外贸易和利用外资规模迅速扩大，/进出口贸易额累计达到一百亿美元，/其中出口六十八点七亿美元，/分别比“七五”时期（一九八六至一九九〇年）增长一点七八倍和一点四三倍【008】

(7) [海关统计表明，]“八五”期间（一九九〇年—一九九五年），中国外商投资企业的进出口呈直线上升之势，/出口年均增长百分之四十三点二，/进口年均增长百分之三十八点六。/去年实现进出口总值达一千零九十八点二亿美元，/占全国进出口总值的比重由上年的百分之三十七提高到百分之三十九。【002】

例（6）“统计资料显示”、例（7）“海关统计表明”均为主谓结构，但其后的语段语法上作二者谓语动词“显示”“表明”的宾语，二者由此也不具有语法独立性，不能划分为小句。但考虑到二者之后的语段一般是复句结构，包含丰富篇章语义信息，整体语段仍作小句切分，只是把“统计资料显示”“海关统计表明”划分到相邻的小句中，如例（6）“统计资料显示，过去五年广西对外贸易和利用外资规模迅速扩大”为一个小句，例（7）“海关统计表明，“八五”期间（一九九 0 年—一九九五年，中国外商投资企业的进出口呈直线上升之势”为一个小句。

3.2. 框式连词包孕的主谓结构“状语”划分为小句

(8) **正值**中国与韩国双边贸易额大幅增长**之际**，/一项大型的中韩经贸研讨会将于今年四月十八日至十九日在北京举行。/这是中韩双方首次举办的专门讨论两国经贸交流与合作的大型研讨会。/业内人士认为，它将为中韩两国经贸界提供一次扩大交流与合作的良机。【014】

(9) **随着**崇明海关办事处**的**设立，/崇明县内的单位足不出岛就可以办理一切海关手续，/这对进一步改善崇明县的投资环境，加快吸引外资，方便快捷地办理海关手续，把崇明建设成对外高度开放的大型贸易港口，带动出口加工、航运中转等外向型经济的发展，将起到积极的作用。【007】

(10) 分析表明，**在**机遇良多、国际形势十分有利**的情况下**，/中国今年经济发展仍面临严峻挑战。【016】

例（8）“正值中国与韩国双边贸易额大幅增长之际”可分析为：框式连词“正值……之际” +主谓结构“中国与韩国双边贸易额大幅增长”。例（9）“随着崇明海关办事处的设立”可分析为：框式连词“随着……的” +主谓结构“崇明海关办事处设立”。例（10）“分析表明，在机遇良多、国际形势十分有利的情况下”可分析为：框式连词“在……的情况下” +主谓结构“分析表明，机遇良多、国际形势十分有利”。这些结构传统被分析为“状语”，但分析为“框式连词+主谓结构”后，有利于篇章语义的分析。篇章依存结构中小句本身不含连接词，这样的话，框式连词包孕的主谓结构就是正常的小句。一般情况下，框式连词包孕的主谓结构在篇章依存结构居于非中心地位，从这个意义上，框式连词是汉语小句从属化的一种重要标志。

3.3. 动介歧义结构区别对待

汉语中有一些词语动介同形，如“根据”“据”等既可作动词，也可作介词，要区别对待它们引导的结构。

3.3.1. 作为动词引导的结构作小句

(11) 为规范建筑行为，/防止出现无序现象，/新区管委会**根据**国家和上海市的有关规定，/结合浦东开发实际，/及时出台了一系列规范建设市场的文件。【001】

(12) **根据**建设部的规定，/凡属于国际金融组织贷款并由国际公开招标的工程全部由外国投资或赠款建设的工程，以及国内企业在技术上难以单独承包的中外合资建设工程，/境外建筑企业在取得中国审批的外国企业承包工程资质证后，/皆可进入中国境内承包建设项目。【004】

这里“根据”等构成的结构有一些特点：①“根据”等词语不可删除，说明其为句法中心，作动词。②与“根据”等组合的短语不能分析为动词性短语，而只能分析为名词性短语，如例（11）的“国家和上海市的有关规定”，例（12）的“建设部的规定”都只能分析为名词性短语，而不能分析为动词性短语。

3.3.2. 作为介词引导的结构不作小句

(13) [**据**中国人民银行西藏自治区分行行长索朗达吉介绍，]“八五”期间，西藏自治区分行在全国率先撤销了人民银行县支行，/中国农业银行西藏自治区分行于去年七月一日正式对外挂牌营业，/实现了金融体制在框架上与全国一致。【005】

这里“根据”等构成的结构有一些特点：①“据”等词语可删除，说明其非句法中心，作介词。②与“据”等组合的短语可以分析为动词性短语，而不能分析为名词性短语，如例（13）的“中国人民银行西藏自治区分行行长索朗达吉介绍”可以分析为动词性短语（主谓结构），而不能分析为名词性短语（定中结构）。

4. 小句关联的难点处理

小句关联分析的一般准则：①直观判断上，两个关联小句能够组成自然的复句；②语义连贯上，两个小句的关联以小句间的词汇语义关联为基础；③形式衔接上，依附于小句的连接词或指代成分等，应从关联小句上得到解释；④形式限制上，篇章中所有小句联通但无环，以保证语篇的整体性和语义关联表示的根本性、简约型；⑤分析程序上，小句关联分析受大句内、大句间等层级域优先性约束一般按照大句内、大句间顺序进行分析；⑥模糊处理上，小句关联的距离就近不就远。有时候要综合运用以上准则，就会遇到一些困难。以下说明小句关联的难点处理。

4.1. “并列”的处理

4.1.1. 有无共同关联对象

语段内并列项有共同的关联对象，则仅构建并列项和共同关联对象之间的关联，而不建立并列项间的关联，主要是考虑小句关联的无环性和语义精简性。如例（14）中建立并列项（小句 2、3）与共同关联对象（小句 1）的关联，而不建立并列项（小句 2、3）的关联，见图 2。**主要考虑是无环限制**，而小句 1 中的“进出口”分别与小句 2、3 的“出口”“进口”有着内在的词汇语义关联。

(14) [1] 海关统计表明，“八五”期间（一九九〇年—一九九五年），中国外商投资企业的进出口呈直线上升之势，/[2] 出口年均增长百分之四十三点二，/[3] 进口年均增长百分之三十八点六。/[4] 去年实现进出口总值达一千零九十八点二亿美元，/[5] 占全国进出口总值的比重由上年的百分之三十七提高到百分之三十九。【002】

图 2 “并列”的处理：有共同关联对象[例（14）]

语段内并列项没有共同关联对象，则直接建立并列项间的小句关联。如例（15）的小句1、2、3 并列，小句 4、5 并列，但它们均无共同关联项，则直接建立并列项间的关联。并列项间关联一般是**就近关联**，由此有 1-2、2-3 关联，进一步又有 3-4、5-7 并列关联。

(15)[1] 在开放开发布局上，广西以北海、钦州、防城为对外开放重点，/[2] 充分发挥首府南宁对外开放城市的作用，/[3]促进沿海、沿边、沿江进一步开放；/[4]办好柳州市城市综合改革试验区、玉林地区城乡综合改革试验区、桂林旅游开发试验区，/[5]建设右江河谷扶贫经济开发带、红水河水电为主的扶贫综合开发带。/[6]并投资一千三百多个亿，/[7]加强基础设施和基础产业建设，/[8]为扩大对外开放创造良好环境。【008】

图 3 “并列”的处理：无共同关联对象[例（15）]

4.1.2. “并列”的层级范围

这里“层级范围”指小句关联是在大句（句号层级的句）内、大句间等层次，一般是先建立大句内的小句关联，才建立大句间的小句关联。“并列”的处理也遵循这一准则。例(16)中，小句 1、2、3 并列，并且处于同一个句子内，需首先建立其间的并列关联，即有1-2、2-3。而小句 1、2、3 虽然均语义上与小句 4 共同有关，但并不分别建立其与小句 4 的关联（见图 4），这和例（14）语义上共同相关句在同一大句的处理（图 2）不同，虽然二者都为了避免形成环。例（16）中仅在并列项小句 1、2、3 中选择比较重要的并列项（中心项）与小句 4 关联（见图 4）。

(16)[1] 目前，辽宁省拥有省、市、县各类外贸进出口公司一百多家，/[2] 有进出口经营权的大中型生产企业二百多家，/[3]有经营进出口业务的外商投资企业一千五百多家。/[4]去年全年共完成进出口总额一百零九点九亿美元。【029】

图 4 “并列”的处理：层级范围[例（16）]

同理，例(17)中（见图 5），小句 5 和 7 是并列项，且处于同一大句内，故应先建立 5-7 关联。而并列项的共同关联项（小句 3），仅和一个并列项（小句 5）关联，即 3-5。注意，例（17）是分号层级的并列，而例（16）是逗号层级的并列，层级上虽有异，但都是大句内的并列，故处理是一致的。

(17)[1] 据海关总署提供的统计数据，/[2] 今年一至二月份中国对外贸易进出口继续保持增长势头，/[3]进出口总值达三百六十一亿美元，/[4]比去年同期增长百分之十三点九。/[5]其中出口为一百七十八点三亿美元，/[6]比去年同期下降百分之一点三；/[7]进口一百八十二点七亿美元，/[8]增长百分之三十四点一。【025】

图 5 “并列”的处理：层级影响[例（17）]

4.1.3. 有无连接词

在构建小句关联时，连接词首先要得到满足，之后再有其他关联建立。因此，在处理

并列项时，如果并列项中含有明显的标志并列关系的连接词，则先按连接词来构建关联。即使是有共同关联对象，也需要先满足含有连接词语的小句关联。

例(18)中，小句 1 和 2 是并列项，在语义上小句 3 是它们的共同关联对象。但是由于在小句 1 和 2 内存在并列连接词“还”，所以需要先满足并列连接词所体现的的小句关联，即有 1-2，而不让并列项分别关联共同对象。相比之下，例（14）无并列连接词，但有共同关联对象，则让并列项分别连接共同关联对象（见图 2）。

(18)[1]特别是为正大公司、华兴铝业公司等外商投资企业开办了常规的运输保险、资产保险、汽车保险等保险业务，/[2]还适时进行雇主责任险、投资保险、利润损失险等新险种，/[3]满足了外商投资需求，/[4]使外商投资企业投保率达到百分之九十以上。【009】

图 6 “并列”的处理：有并列连接词[例（18）]

4.2. 小句间词汇语义关联强弱难判

小句关联以小句间的词汇语义联系为基础，但有时某些小句间的语义关联强弱难以区分，以致很难从语义上判断与取舍小句关联。如例（19）究竟是 1-3 还是 2-3，例（20）究竟是 1-4 还是 1-5，例（21）究竟是 2-3 还是 2-6，很难从语义上取舍出关联强弱来。

这时候可利用共指依赖和距离远近来判断，一般是有共指依赖的要从当前小句关联中得到解读，而距离上除非有特别强的语义提示一般是就近关联。据此，例（19）取 2-3，不取 1-3，因为前者小句 2、3 距离近，且主语共指，而小句 1、3 距离远，主语不共指；例（20）取 1-4，不取 1-5，因为 1-4 距离近，主语各自独立明确，而 1-5 距离远，5 缺主语但并不能从 1 中得到解释（只能从 4 中得到解释）；例（21）取 2-3，不取 2-6，因为 2-3 距离近，连接词“对此”也能就近得到解释，且 3 的主语独立明确，相比之下，2-6 距离远，6 的主语指代不能从 1 中得到解释（仅能从 3 中得到解释）。见图 7、8、9。

(19)[1]“八五”（一九九一至一九九五年）期间，西藏金融体制改革坚持与全国框架一致、体制衔接的方针，/[2]顺利完成了西藏各级人民银行的分设工作，/[3]实现信贷资金使用从粗放型经营方式向集约型经营方式转变。/[4]去年，全区各项存款首次突破了年净增二十亿元大关。【005】

图 1 小句间词汇语义关联强弱难判[例（19）]

(20)[1]海关统计表明，“八五”期间（一九九O年—一九九五年），中国外商投资企业的进出口呈直线上升之势，/[2]出口年均增长百分之四十三点二，/[3]进口年均增长百分之三十八点六。/[4]去年实现进出口总值达一千零九十八点二亿美元，/[5]占全国进出口总值的比重由上年的百分之三十七提高到百分之三十九。【002】

图 8 小句间词汇语义关联强弱难判[例（20）]

(21)浦东开发开放是一项振兴上海，建设现代化经济、贸易、金融中心的跨世纪工程，/因此大量出现的是以前不曾遇到过的新情况、新问题。对此，浦东不是简单的采取“干一

段时间，等积累了经验以后再制定法规条例”的做法，/而是借鉴发达国家和深圳等特区的经验教训，/聘请国内外有关专家学者，/积极、及时地制定和推出法规性文件，/使这些经济活动一出现就被纳入法制轨道。/去年初浦东新区诞生的中国第一家医疗机构药品采购服务中心，正因为一开始就比较规范，/运转至今，/成交药品一亿多元，/没有发现一例回扣。【001】

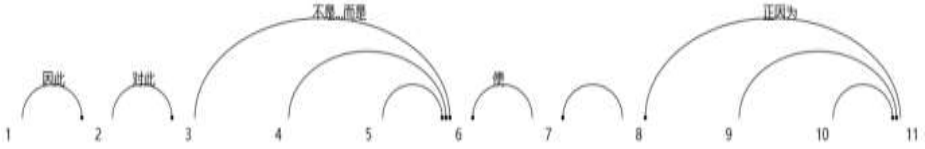

图 9 小句间词汇语义关联强弱难判[例（21）]

5. 中心确立的难点处理

中心分析一般准则：①分析程序与原则上，小句关联中心一般在语篇的所有小句关联建立后确定，通常根据小句的全局相对关联度大小来确定中心，对于同一个关联对中的两个小句，忽略当前小句关联后，比较两个小句在整个语篇（段落）中构成的其他关联的数量多少，关联数量大的一方为此关联对的中心。②形式限制上，一个小句关联对有且仅有一个中心，一个小句（除全篇中心外）指向并仅指向一个中心。③中心是语篇的，但在词汇、句法上有一定表现。以下是一些小句关联中心的难点及处理。

5.1. 当前关联度不足

目前，篇章依存结构的工作对象为段落，在段落范围内，关联度相对较大的可能并非真正中心，而真正中心需要利用段间的关系或段间的关联来确定中心，也可以利用标题与小句间的关联来确定中心。如例(22)有三个小句，构成了两个关联对。按照段内的全局关联度，应该是小句 3 为关联对 1-3 的中心，但是结合更大上下文，如标题“中国十四个边境开放城市经济建设成就显著”，可确定真正的中心应该是小句 1。例（2）有二个小句，构成一个关联对，当前无从根据关联度确定中心，结合标题及更大上下文，可确定中心为小句 1。

(22)[1]中国十四个边境对外开放城市一九九五年经济建设取得可喜成果。/[2]据统计，这些城市去年完成国内生产总值一百九十多亿元,/[3]比开放前的一九九一年增长九成多。【003】

图 2 当前关联度不足[例（22）]

(23)中国进出口银行最近在日本取得债券信用等级ＡＡ－，/这是日本金融市场当前对中国银行的最高债券评级。【010】

图 3 当前关联度不足[例（23）]

5.2. “并列”结构的中心

一个关联对仅有一个中心。当出现小句“并列”关联，要考虑到中心对关联的限制，选取其中一个更加适合的小句作为中心，进行对外的关联。例(23)中，关联 1-2 即为通常所谓“并列”关联，“还”表明这种关系。通常认为“并列”项同等重要，但篇章依存结构中，一个关联对仅有一个中心。例（24）从“特别是”“还适时”等可以看出小句 1 在语义上更

为重要，确立其为中心。例(25)中，关联对 2-3 里，小句 2 比小句 3 拥有更加完整的主语，依赖性更弱，语义更完整，因此对外联系的程度越高，是该关联对中的中心。

(24)[1]特别是为正大公司、华兴铝业公司等外商投资企业开办了常规的运输保险、资产保险、汽车保险等保险业务，/[2]还适时进行雇主责任险、投资保险、利润损失险等新险种，/满足了外商投资需求，/[3]使外商投资企业投保率达到百分之九十以上。【009】

图 4 “并列”结构的中心[例（24）]

(25)针对甘肃旅游业的发展需求，/人保公司积极推出海外游客保险，/还在国内首家推出海外散客保险办法，/使“八五”期间凡到甘肃观光游览的海外游客全部得到保险保障。【009】

图 5 “并列”结构的中心[例（25）]

6. 结语

篇章依存结构相比修辞结构等层次化篇章结构直接刻画了最小篇章单位间的直接关联，有效刻画了非连续和跨层级的最小篇章单位间直接关系及篇章向心性关系，对篇章结构分析有其独有价值。我们正在构建汉语篇章依存结构语料库，本文总结了我们在汉语篇章依存结构语料库标注中的一些难点及处理方案，主要用于指导我们的语料标注实施，也可用以指导进一步自动标注中的一些问题解决。目前，我们标注了新闻语料 300 篇，并进行了一些初步的自动分析研究。下一步，我们将完善和扩大语料标注，进行更深入的自动分析研究。

参考文献

[1] Mann W ,Thompson S. Rhetorical structure theory. Toward a functional theory of text organization. Text, 1988, 8(3):243–281.

[2] Carlson L, Marcu D, Okurowski M E. Building a discourse-tagged corpus in the framework of rhetorical structure theory. In Current and new directions in discourse and dialogue, Springer, 2003: 85–112.

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

[4] Sujian Li, Liang Wang, Ziqiang Cao, Wenjie Li. Text-level discourse dependency parsing. Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (ACL), 2014:25–35.

[5] An Yang, Sujian Li. SciDTB: Discourse Dependency TreeBank for Scientific

Abstracts .Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL) , 2018: 444–449.

[6] Yoshida Y, Suzuki J, Hirao T, et al. Dependency-based discourse parser for single-document summarization. In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). 2014:1834–1839.

[7] 冯文贺.汉语篇章结构的特征-依存描写机制及资源建设，国家社科基金，2017.

[8] 冯文贺,陈伊琳.汉语篇章小句关联结构的资源标注与计算模型研究,The 27th Annual Conference of International Association of Chinese Linguistics (IACL-27),2019.

[9] 冯文贺,黄熙.篇章特征-依存结构标注软件,国家版权局软件登记号：2019SR0043051。

[10] Nianwen Xue, Fei Xia, Fu dong Chiou, and Martha Palmer. The Penn Chinese TreeBank: Phrase Structure Annotation of a Large Corpus. Natural Language Engineering, 2005, 11(2):207–238.

[11] 李艳翠,冯文贺,周国栋,等.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, 49(1):7-14.

[12] Yancui Li, Wenhe Feng , Fang Kong, et al. Building Chinese discourse corpus with connective-driven dependency tree structure. Proceedings of the Conference on Empirical Methods in Natural Language Processing(EMNLP), 2014:2105-2114.