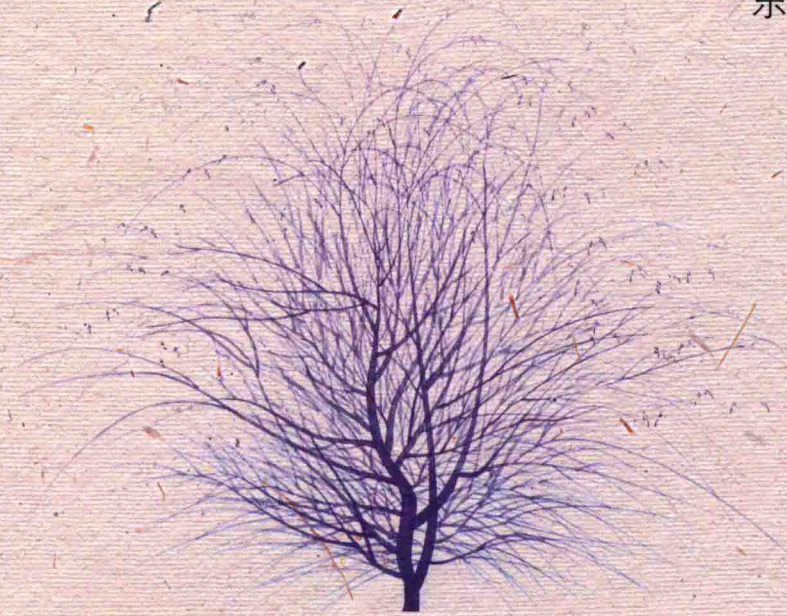
基于修辞结构树库的篇章衔接标记研究

乐　明◎著



Rhetorical Structure Theory
Rhetorical Structure Theory
Rhetorical Structure Theory
Rhetorical Structure Theory
Rhetorical Structure Theory

世界图书出版公司

本书出版承蒙浙江大学董氏文史哲研究奖励基金资助

基于修辞结构树库的篇章衔接标记研究

乐　明◎著

中国出版集团公司
世界图书出版公司
广州·北京·上海·西安

图书在版编目（CIP）数据

基于修辞结构树库的篇章衔接标记研究/乐明著. —广州：世界图书出版广东有限公司，2019.1

ISBN 978-7-5192-5878-8

Ⅰ. ①基… Ⅱ. ①乐… Ⅲ. ①汉语—修辞—语料库—研究 Ⅳ. ①H15

中国版本图书馆 CIP 数据核字（2019）第 005360 号

书　　名	基于修辞结构树库的篇章衔接标记研究 JIYU XIUCI JIEGOU SHUKU DE PIANZHANG XIANJIE BIAOJI YANJIU
著　　者	乐　明
责任编辑	冯彦庄
装帧设计	吴伟边
责任技编	刘上锦
出版发行	世界图书出版广东有限公司
地　　址	广州市新港西路大江冲 25 号
邮　　编	510300
电　　话	020-84452177
网　　址	http://www.gdst.com.cn
邮　　箱	wpc_gdst@163.com
经　　销	新华书店
印　　刷	广州市迪桦彩印有限公司
开　　本	787 mm×1 092 mm　1/16
印　　张	16
字　　数	300 千字
版　　次	2019 年 1 月第 1 版　2019 年 1 月第 1 次印刷
国际书号	ISBN 978-7-5192-5878-8
定　　价	58.00 元

《中国当代语言学文库》

编委会

序　言

1987 年，W. Mann 和 S. Thompson 在《修辞结构理论：一种文本组织的理论》（Rhetorical Structure Theory：A Theory of Text Organization）一文中，提出“修辞结构理论”（Rhetorical Structure Theory，简称为 RST）。这是一种基于文本局部之间关系的关于文本组织的描述理论。

例如，我们来研究下面的两个文本：

a. 小王喜欢收集中国的古典文学作品。他昨天买了曹雪芹的《红楼梦》。b. 小王喜欢收集中国的古典文学作品。他昨天买了托尔斯泰的《战争与和平》。

文本 a 是有意义的，它表达了小王昨天买了曹雪芹的《红楼梦》这个事实，这个事实很自然地紧接着他喜欢收集中国的古典文学作品的事实。而文本 b 则是有缺陷的。这种缺陷并不是单个句子的问题，文本 b 中的单个的句子孤立地看起来都是完美的，缺陷在于它们在意思上的结合不好，因为托尔斯泰的《战争与和平》不是中国的古典文学作品，而是俄罗斯的文学作品。不过，文本中两个句子顺序排列的事实暗示它们之间具有某种衔接关系，只不过文本 a 和文本 b 的衔接关系是不同的。对于文本 a 来说，这种关系具有详述（elaboration）关系的特征。而对于文本 b 来说，这种关系则具有对照（contrast 或 antithesis）关系的特征，因此，文本 b 应当更恰当地表示为：

小王喜欢收集中国的古典文学作品。然而，他昨天买了托尔斯泰的《战争与和平》。

这里，“然而”这个词明显地将对照关系的信号传递给读者，因此这个文本在意思上也就顺畅多了。

RST 描述了文本中篇章单元之间的衔接关系，建立了完整的理论，提出了具

体的方法。

Mann 和 Thompson 对语言使用的性质和如何解释这种性质持有这样一些基本观点：

1）如果想说明话语本身，就必需对说话者和听话者的参与有个明确的解释；

2）话语的结构比其他任何事物都更反映说话者的意图和目标，而意图普遍是有层次的；

3）注意（attention）和意图（intention）被认为是文本中相互独立又相互作用的方面；话语的形式、功能和结构三者之间的联系，并不是某种一一对应的映射关系，而是松散的相互制约关系。因此并不总有什么特定的词汇或语法形式唯一的标记结构特征。

RST 理论的核心是修辞关系的概念。修辞关系（Rhetorical Relation）是存在于两个互不重叠的文本跨段（Text Span）之间的关系（当然也有一些例外），这两个文本跨段一个叫“核心单元”（Nucleus，用 N 来表示），一个叫“卫星单元”（Satellite，用 S 来表示）。这种对核心和卫星的区分来自经验观察。例如，在上面的文本 a 中，“小王喜欢收集中国的古典文学作品”这个片断是核心单元，“他昨天买了曹雪芹的《红楼梦》”这个片断是卫星单元。核心单元与卫星单元的划分说明，许多修辞关系是非对称的，这里第二个片断是根据第一个片断来解释的。但并不是所有的修辞关系都是非对称的。RST 关系是根据它们施加于核心、外围、以及核心和外围的结合处的约束来定义的。

根据文本分析经验，Mann 和 Thompson 对文本结构作了如下的一些基本假设：组织性、整体性和连贯性、功能性、层级性、层级的同质性、关系的组合性、关系的不对称性、关系性质的“修辞”功能。

在对大量真实文本分析的经验基础上，Mann 和 Thompson 总结出了 25 种修辞关系，分为核心—卫星关系和多核心关系。核心—卫星关系可以表示为（N－S），多核心关系可以表示为（N－N（…N））：

核心－卫星关系有如下 21 种：

证据（Evidence）	辩理（Justification）
转折（Antithesis）	让步（Concession）
情景（Circumstance）	解决（Solution）
详述（Elaboration）	背景（Background）
使能（Enablement）	条件（Condition）
否则（Otherwise）	解释（Interpretation）

评估（Evaluation）　　重述（Restatement）

总结（Conclusion）　　动机（Motivation）

意愿性原因（Volitional cause）　　非意愿性原因（Non－volitional cause）

意愿性结果（Volitional result）　　非意愿性结果（Non－volitional result）

目的（Purpose）

多核心关系有如下4种：

序列（Sequence）　对照（Contrast）　联接（Joint）　列表（List）

Mann 和 Thompson 在多篇论文中反复强调，他们给出的修辞关系类型不是一个封闭的集合。核心—卫星关系所列的是 Mann 和 Thompson 在分析英语独白文本中发现的、能覆盖他们遇到的绝大部分语料的各种关系类型。

从这些作用于核心、卫星和核心卫星间组合的限制以及与每个关系相联系的总体效果中，可以得出文本的衔接性。在 RST 中，文本衔接性是从一套限制以及与每个关系相联系的总体效果中得出的。这些限制作用于核心、卫星和核心卫星间的组合。

用来描写文本结构的 RST 只识辨结构的三种主要类型：整体结构、关系结构和句法结构，并主要研究中间一层的关系结构。RST 提供了一般的方法来描写文本各组织元素之间的结构关系，不论这些关系是语法上标记的，还是词汇上标记的。因此，RST 是联系各连词的各种意义、小句组合的语法和无标记平行结构的一个有用的理论框架。同时，RST 提供了研究关系命题（Propositional Relations）的框架，关系命题是在解释文本的过程中从文本结构得出的、未经陈述但可以引申出来的命题。因为文本的衔接性部分地依赖于这些关系命题，所以 RST 在文本衔接性研究中也很有用。除了衔接性的研究外，RST 还被用于研究小句间关系、连接词语、暗示性交际、小句组合、作品风格和文体等。

RST 总共给出的 25 种修辞结构关系已经足以描述各式各样文本的修辞结构。在实践中，研究人员倾向于从这些修辞结构关系中选出适合他们各自应用领域的子集。

RST 明确地提出了文本的树结构模型。RST 的树结构模型要满足完整性、联系性和唯一性三个条件，从根结点开始的树形图（tree graph）可以代表整个文本的修辞结构关系。

Mann 和 Thompson 认为，文本的修辞结构都可以用这样的树形图来表示，因此，我们应当花力气来构建修辞结构树库（tree bank），在树库中，树形图的各子结点代表各个互不重叠但又相互衔接的文本跨段。一个文本跨段是任何一部分

从文本组织的角度上看有功能整体性的一个文本片段。修辞结构关系存在于两个不重叠的文本跨段之间。

乐明博士依据 Mann 和 Thompson 的修辞结构理论，并吸取英语和德语的修辞结构树库在修辞关系标注和相关表层信息标注方面的经验，为一定规模的汉语财经评论语篇构建了汉语修辞结构树库，定义了汉语修辞关系集、篇章单元切分原则、关系优选原则和标注流程。汉语财经评论的篇章修辞结构树库有 400 篇新闻财经评论（CJPL400），总计 70 万字。乐明本人标注了所有篇章在分号句颗粒度上的局部修辞结构树，并为 150 篇文章构造了完整的修辞结构树；由她的合作者对其中较短的 86 篇文章（CJPL86）完成了多个版本的修辞结构标注，并对其中的 20 篇较短的文章（CJPL20）完成了逗号子句层次上的修辞结构标注。乐明博士在这些标注的基础上作了篇章修辞关系分布的研究，并对篇章语料所含的多个衔接标记进行了自动和人工相结合的多维度的标注和个案研究。

由于面对真实的语料，这些工作数据量大，难度高，计算技术也比较复杂。经过多年不懈努力，乐明博士在这些工作的基础上，写成了《基于修辞结构树库的篇章衔接标记研究》这本专著。在这本书出版之际，我对她表示热烈的祝贺。

在计算语言学中，篇章之间的衔接关系的研究一直是一个比较薄弱的环节，在中文信息处理中，汉语的篇章衔接标记用法的研究也不多见。我相信，本书的出版将有力地推动这个领域的研究。

希望乐明博士再接再厉，在计算语言学的艰苦探索中，创造出更好的成绩。

冯志伟

2017 年 7 月 15 日于北京

目　录

缩略语 …… 1

第一章　篇章连贯研究概论 …… 3
1.1　研究对象 …… 4
1.2　研究背景 …… 7
1.2.1　篇章连贯的理论研究 …… 8
1.2.2　篇章剖析技术的发展 …… 10
1.2.3　标注篇章语料库的开发 …… 12
1.2.4　中文信息处理的相关研究 …… 14
1.3　研究问题 …… 16
1.4　研究方法及主要成果 …… 16
1.5　本书结构 …… 17

第二章　研究方法 …… 19
2.1　语料库设计 …… 20
2.2　语料标注 …… 21
2.2.1　标注理据 …… 21
2.2.2　标注原则 …… 22
2.2.3　标注格式 …… 23
2.2.4　标注手段 …… 23
2.2.5　标注质量的控制和检验 …… 24
2.3　统计分析和机器学习 …… 26
2.3.1　描述性统计 …… 26
2.3.2　推断性统计 …… 27
2.3.3　机器学习 …… 28
2.4　小结 …… 31

第三章　构建汉语篇章修辞结构树库 …… 32
3.1　修辞结构理论 …… 32
3.1.1　内容简介 …… 33
3.1.2　应用情况 …… 36
3.1.3　主要争议 …… 36
3.1.4　与汉语传统理论的比较 …… 39
3.2　构建汉语篇章语料库 …… 41
3.2.1　语料选择 …… 41
3.2.2　语料预处理 …… 42
3.3　汉语篇章基本单元切分 …… 43
3.3.1　篇章基本单元的定义 …… 43
3.3.2　切分的方法 …… 45
3.3.3　自动切分的形式标记 …… 46
3.3.4　自动切分的处理结果 …… 47
3.4　汉语篇章修辞关系集的设定 …… 48
3.4.1　基于特征的定义方法 …… 49
3.4.2　汉语修辞关系集的简化分类 …… 52
3.4.3　一些说明 …… 57
3.5　修辞关系标注及篇章修辞结构树的构造 …… 58
3.5.1　分号句层级及以上的修辞结构标注 …… 59
3.5.2　逗号子句层级的修辞结构标注 …… 63
3.6　工具和流程 …… 68
3.7　质量控制和检验 …… 70
3.8　小结 …… 72

第四章　标注篇章衔接标记特征 …… 74
4.1　标注对象 …… 75
4.2　理论框架 …… 76
4.2.1　连接词 …… 77
4.2.2　指代词 …… 78
4.2.3　语气词 …… 80

4.2.4 标点符号 …… 81
4.3 自动标注 …… 81
4.4 人工标注 …… 82
4.4.1 标注内容(通用部分) …… 83
4.4.2 软件工具 …… 84
4.4.3 标注流程和质量控制 …… 85
4.5 统计分析和参数选择 …… 87
4.6 小结 …… 88

第五章 数据分析 …… 89
5.1 财经评论语料库篇章结构特点 …… 89
5.1.1 修辞结构树特征 …… 89
5.1.2 分号句层级上的 RR 概率分布 …… 90
5.1.3 小结 …… 93
5.2 但、但是 …… 93
5.2.1 前人成果 …… 93
5.2.2 标注方案 …… 95
5.2.3 数据结果 …… 96
5.2.4 讨论 …… 100
5.2.5 小结 …… 103
5.3 这、那 …… 103
5.3.1 前人成果 …… 104
5.3.2 标注方案 …… 104
5.3.3 数据结果 …… 105
5.3.4 讨论 …… 107
5.3.5 小结 …… 111
5.4 吗、? …… 112
5.4.1 前人成果 …… 112
5.4.2 标注方案 …… 113
5.4.3 数据结果 …… 114
5.4.4 讨论 …… 118
5.4.5 小结 …… 122

第六章　应用测试——以“因为”为例 …… 123
6.1　研究背景 …… 123
6.2　任务分析 …… 125
6.3　学习工具 …… 127
6.4　数据处理、结果及讨论 …… 128
6.5　小结 …… 130

第七章　讨论 …… 132
7.1　修辞结构树的理论和工程价值 …… 132
7.1.1　修辞结构树和修辞关系的性质 …… 132
7.1.2　修辞结构的层级同质性 …… 134
7.1.3　修辞结构树与依存句法树 …… 136
7.2　对 RR 与 DCM 的理解 …… 138
7.2.1　RR 的多义性和模糊性 …… 138
7.2.2　DCM 的变异和多义性 …… 141
7.2.3　RR 与 DCM 间的多对多映射 …… 142
7.2.4　多义性、多因分析与多任务学习 …… 143
7.3　对语料库人工标注的理解 …… 144
7.3.1　人工标注的准确性和挑战 …… 144
7.3.2　一致性指标的解读 …… 146
7.3.3　可能的解决方案 …… 148
7.4　数据挖掘技术在语言工程和理论研究的应用 …… 149

第八章　结　语 …… 153

第九章　附　录 …… 156
9.1　中文 RR 关系集(按中文名称排序) …… 156
9.2　中文 RR 定义及示例(按英文名称排序) …… 157
9.2.1　addition－n(附加－n) …… 158
9.2.2　addition－s(附加－s) …… 159
9.2.3　antithesis－s(转折－s) …… 160
9.2.4　attribution－m(引述－m) …… 161

9.2.5 attribution – s(引述 – s) …… 163
9.2.6 background – s(背景 – s) …… 164
9.2.7 circumstance – s(环境 – s) …… 165
9.2.8 concession – n(让步 – n) …… 166
9.2.9 concession – s(让步 – s) …… 167
9.2.10 condition – s(条件 – s) …… 168
9.2.11 conjunction – m(并列 – m) …… 169
9.2.12 contrast – m(对立 – m) …… 169
9.2.13 disjunction – m(析取 – m) …… 171
9.2.14 elaboration – s(详述 – s) …… 172
9.2.15 enablement – s(使能 – s) …… 173
9.2.16 evaluation – m(评价 – m) …… 173
9.2.17 evaluation – n(评价 – n) …… 174
9.2.18 evaluation – s(评价 – s) …… 176
9.2.19 evidence – s(证据 – s) …… 176
9.2.20 interpretation – n(解释 – n) …… 177
9.2.21 interpretation – s(解释 – s) …… 178
9.2.22 joint – m(联接 – m) …… 179
9.2.23 justify – n(论证 – n) …… 181
9.2.24 justify – s(论证 – s) …… 183
9.2.25 list – m(罗列 – m) …… 183
9.2.26 means – s(方式 – s) …… 184
9.2.27 motivation – s(动机 – s) …… 184
9.2.28 nonvolitional – cause – s(非意愿性原因 – s,NVC) …… 184
9.2.29 nonvolitional-cause-result-m(非意愿性因果-m,NCR) …… 185
9.2.30 nonvolitional – result – s(非意愿性结果 – s,NVR) …… 186
9.2.31 otherwise – s(否则 – s) …… 187
9.2.32 preparation – s(准备 – s) …… 188
9.2.33 purpose – s(目的 – s) …… 190
9.2.34 question – m(问答 – m) …… 191
9.2.35 question – s(问答 – s) …… 191
9.2.36 restatement – s(重述 – s) …… 192

9.2.37 same - unit - m(同一单元 - m) …… 193
9.2.38 sequence - m1(主观语用序列 - m) …… 194
9.2.39 sequence - m2(客观语义序列 - m) …… 195
9.2.40 solution - m(解答 - m) …… 196
9.2.41 solution - n(解答 - n) …… 197
9.2.42 solution - s(解答 - s) …… 198
9.2.43 summary - s(总结 - s) …… 199
9.2.44 topic - comment - m(话题 - 评论 - m) …… 200
9.2.45 unconditional - s(无条件 - s) …… 201
9.2.46 unless - s(除非 - s) …… 201
9.2.47 volitional - casue - s(意愿性原因 - s,VC) …… 202
9.2.48 volitional - cause - result - m(意愿性因果 - m,VCR) …… 203
9.2.49 volitional - result - s(意愿性结果 - s,VR) …… 204
9.2.50 ??? - s(未知关系) …… 205
9.3 **汉语 RR 标注关系优选协议及相关说明(2.0)** …… 205
9.4 CRST - DT **树形构造规范(2017.6)** …… 210
9.4.1 树形选择 …… 210
9.4.2 N 的确定 …… 212
9.4.3 非树结构的处理 …… 214
9.5 **标注样例(CJPL2224725)** …… 215
9.5.1 原文 …… 215
9.5.2 标注结果 …… 216

参考文献 …… 217

索 引 …… 234

后 记 …… 236

表目录

表2-1 机器学习的类型 …… 29
表2-2 机器学习与统计分析的部分对应术语 …… 30
表3-1 汉语RR特征矩阵 …… 55
表3-2 汉语RR分组表 …… 56
表3-3 CJPL标注工作分配示意 …… 68
表3-4 CJPL树库两名标注者间一致性检验数据提取表(Kappa系数) …… 71
表3-5 CJPL修辞结构树库最终两名标注者间一致性检验结果(Kappa系数) …… 72
表4-1 与其篇章用法相关的“但”的特征 …… 87
表5-1 CJPL修辞关系与修正右截尾Zipf-Alekseev分布的拟合 …… 92
表5-2 CJPL400中“但”“但是”与紧跟其后的词语 …… 95
表5-3 “但”“但是”在CJPL语料库中的分布 …… 96
表5-4 CJPL语料库“但”“但是”经正交旋转后的因子载荷矩阵 …… 101
表5-5 CJPL(分)句首词“这$_{指}$和这$_{代}$”的独立样本T检验结果汇总 …… 107
表5-6 CJPL(分)句首回指词“这”所在EUDA的rr分布 …… 108
表5-7 “这”用法特征与RR间的Pearson相关性系数汇总 …… 108
表5-8 “吗”字句在CJPL400中的分布 …… 115
表5-9 修辞关系在CJPL400“吗”字句中的分布 …… 116
表5-10 “吗”字句上层结构在语篇中的修辞结构作用RRabove分布 …… 117
表5-11 篇章、段落位置与“吗”字句结构、语义及语用类型间的相关性分析 …… 119
表9-1 汉语修辞关系优选代码表 …… 208

图目录

图 1 - 1　本书所用术语的语义范围界定示意 …… 4
图 1 - 2　基于修辞结构树库的 DCM 用法研究示意图 …… 17
图 3 - 1　修辞结构子树的基本类型:A 单核关系;B 多核关系 …… 35
图 3 - 2　篇章连贯结构中的交叉依存关系示例(Wolf and Gibson,2005) …… 38
图 3 - 3　各类标点符号在 CJPL400 中的使用比例 …… 46
图 3 - 4　经典 RST 文献对修辞关系的定义格式(Mann and Thompson,1987) …… 50
图 3 - 5　CJPL3145038 - 14.25 承上单元处理示例 …… 60
图 3 - 6　CJPL2089327 - 12.17"当然"句式的处理图示 …… 62
图 3 - 7　反映在分号句层次的篇章结构树图(CJPL2184103.12) …… 62
图 3 - 8　句内顶层分析得到的结构(CJPL2184103.12) …… 62
图 3 - 9　逗号子句层次的修辞结构子树(CJPL2184103.12) …… 63
图 3 - 10　逗号子句作状语时的修辞结构子树示例 …… 65
图 3 - 11　逗号子句层级的 preparation - s 和 topic - comment - m 关系示例 … 66
图 3 - 12　CJPL 多层次语言学信息标注系统的主要构件 …… 68
图 3 - 13　CJPL 树库开发中期两名标注者间的一致性结果(Holsti 系数) …… 71
图 4 - 1　修辞距离示例(CJPL1995119.2) …… 80
图 4 - 2　CJPL 多层次语言学信息标注方案相互关系 …… 83
图 4 - 3　CJPL 多层次语言学信息人工标注流程图 …… 86
图 5 - 1　两种螺旋型篇章子树图示 …… 90
图 5 - 2　CJPL rr 分布图(16 大类,三层结构) …… 91
图 5 - 3　CJPL rr 分布的右截尾 Zipf - Alekseev 拟合图(Yue and Liu,2011) …… 93
图 5 - 4　"但"和"但是"的显著差异特征 …… 100
图 5 - 5　决定"但""但是"用法的主要成分分析结果(碎石图) …… 101

图 5－6　CJPL400(分)句首词“这”的先行语结构类别和句内功能分布 …… 105
图 5－7　CJPL400(分)句首词“这”所在回指语的句法功能和语义类型分布 …… 107
图 5－8　篇章衔接标记“吗”的修辞结构树示例(CJPL1995119.2) …… 114
图 6－1　篇章衔接标记“因为”的修辞结构属性示例 …… 127
图 6－2　基于多任务学习的篇章修辞结构信息提取流程 …… 129
图 6－3　用 SVM^{multiclass}对目标特征进行封闭和开放测试的结果与随机概率对比 …… 130
图 7－1　汉语依存句法树示例(刘海涛,2009) …… 137
图 7－2　篇章多级依存原文结构和文摘结构示意图(刘挺、王开铸,1999) …… 138
图 7－3　转折关系的语义图 …… 139
图 7－4　篇章多义关系示例及标注选择(CJPL2091506.3) …… 140
图 7－5　数据库知识发现步骤图(Fayyad et al.,1996) …… 150
图 7－6　数据库知识发现一般模式(Andrássyová and Paralič,1999) …… 151
图 9－1　addition－n 和 preparation－s 的图示(CJPL2240288) …… 159
图 9－2　antithesis－s、concession－s 和 evaluation－n 的图示(CJPL2236207.4－5) …… 161
图 9－3　attribution－s、antithesis－s 和 NVC 的图示(CJPL2423034.2－3) …… 163
图 9－4　circumstance－s 及 evluation－n 的图示(CJPL2031739) …… 165
图 9－5　concession－n、antithesis－s 和 purpose－s 的图示(CJPL2247936.7) …… 167
图 9－6　condition－s 和 justify－s 的图示(CJPL2269991.7) …… 168
图 9－7　contrast－m 和 elaboration－s 的图示(CJPL3148638.3) …… 170
图 9－8　disjunction－m 和 justify－s 的图示(CJPL3171559.7) …… 172
图 9－9　篇头组织元素(DOE)和 evaluation－m 及 justify－s 的图示(CJPL2066838) …… 174
图 9－10　评价－n、论证－s 关系的图示(CJPL2250706) …… 175
图 9－11　evidence－s 和 VC 的图示(CJPL2081402.7) …… 177
图 9－12　interpretation－n 和 NVC 的图示(CJPL2423034.3) …… 178