基于话题链的汉语语篇连贯性描述体系*

周强[1,2]，周骁聪[1,2]

1. 清华信息科学与技术国家实验室（筹）

2. 清华大学信息技术研究院语音和语言技术中心，北京 100084

摘要：汉语简洁灵活的意合型篇章组合结构，对传统的基于关联词的篇章连贯性描述体系提出了新的挑战。本文引入话题链描述形式，设计不同类型的话题评述关系集，构建了以话题链为主，融合关联词语和其他连贯形式描述机制，覆盖话题评述、并列、因果、转折四大类关系的汉语语篇连贯性描述体系。在清华句法树库 TCT 上进行的验证实验，发现话题链和关联词语分别覆盖了约 76%和 50%的汉语复句，并且两者经常同时使用，初步证明了这个体系在句子连贯性描述方面的可行性和有效性。

关键词：话题链；话题评述关系；连贯性描述体系；汉语语篇分析

中图分类号：TP391　　**文献标识码：**A

Topic-Chain-Based Coherence Annotation Scheme for Chinese Text

Qiang Zhou[1,2], Xiaocong Zhou[1,2]

1. Tsinghua National Laboratory for Information Science and Technology

2. CSLT, Research Institute of Information Technology, Tsinghua University, Beijing 100084

Abstract: There are few explicit discourse connectives in Chinese real texts. It brings in new challenge for the traditional connective-grounded coherence annotation scheme. The paper proposed a new idea to deal with the problem. We introduced topic chain (TC) as a main coherence representation and designed several topic-comment relations to describe the complex event relations among TC-linked sentences. Therefore, a new coherence annotation scheme based on TCs and connectives are built accordingly. The tentative confirmatory experiments on the Tsinghua Chinese Treebank (TCT) data set show that more than 76% and 50% Chinese complex sentences have TCs and connectives respectively. They can be co-occurrence in most Chinese sentences. The phenomena verify the feasibility and availability of this scheme.

Keywords: Topic Chain; Topic-Comment Relation; Coherence Annotation Scheme; Chinese Text Analysis

1. 问题的提出

对语篇内容的分析计算是自然语言理解研究的核心课题。经过 50 多年的发展，在英语、德语等西方语言的篇章描述语料库构建和篇章分析技术方面已经取得了长足的进步[1]，开始在文本摘要、信息检索、机器翻译、情感分析和文本质量评估方面显示出重要的应用潜力[2]。而在汉语等东方语言的篇章描述和分析理解方面还很薄弱，有许多研究空白亟待开拓和探索。

在篇章理解方面，修辞结构理论(Rhetorical Structure Theory，RST)[3]描述了篇章整体组织特点，主要强调句子之间各种连贯关系的分析在篇章理解中的重要作用。以 RST 为基础，Marcu & al.(2001)构建了英语篇章标注库 RST-DT，以篇章中的小句为基本单位，自底向上组合形成二叉或多叉的修辞结构树，形成对篇章意义的整体描述[4]。该成果推动了以基于统计和机器学习模型为主的 RST 自动篇章分析器的开发研究[5]。

宾州话语树库 PDTB(Penn Discourse Teebank)[6]则选择篇章连接词作为‘元’谓词，通过分析和标注它所控制的两个句子论元，形成对这两个句子形成的篇章片段连贯性的初步理解。针对真实文本中大量存在的连接词缺失现象，又设计了多层次标记和多重特征相结合的

*收稿日期：　　定稿日期：

基金项目：国家 973 计划项目：互联网环境中文言语信息处理与深度计算的基础理论和方法（2013CB329304）；国家自然科学基金项目：汉语语篇中连贯关系和隐含角色的分析标注研究（61373075）

描述机制。在此标注库上进行的评测实验结果显示[7]，隐式连接词发现和相应关系标记识别[8]是主要分析难点，具有很大的技术挑战性。

针对英语篇章语境中某些事件核心块的省略和缺失现象，Rappenhofer & al. (2010)设计了SemEval-2010评测任务，手工标注了两篇小说体裁篇章中的所有事件缺失块及其在篇章语境中可能的共指成分[9]。相关的进一步研究工作表明[10]，对这个问题的分析研究还有很大的发展潜力。另一项重要评测是在CoNLL-2011中提出的指代消解评测任务[11]，希望能充分利用最新发布的OntoNotes库[12]中同步标注的句法、命题、词义、命名实体和共指信息，提升自动指代消解系统的处理性能。评测结果显示[11]，要达到各层次标注信息互动提高的处理效果，还需要在学习建模技术上有新的创新。

以上研究工作从两个不同途径探索了语篇连贯性(coherence)的描述方法：一类是基于关系的连贯，包括针对篇章整体的RST结构分析和针对篇章片段的PDTB分析。他们强调了对篇章中通过显式和隐式关联标记体现出来的连贯关系的分析和把握。另一类是基于实体的连贯，包括SemEval-2010任务中设计的隐含角色链和OntoNotes中标注的实体提及(mention)链。他们强调了对篇章中通过不同形式的指代引用关系体现出来的连贯片段的分析和把握。虽然两者的研究对象和处理方法各不相同，但从总体上可以归入Halliday & Hasan (1976)提出的广义语篇连贯性描述框架[13]中的连贯和衔接(cohesion)两个不同描述手段上。相对而言，英语语篇研究学者更关注从基于语义的修辞关系角度分析语篇的连贯性。

与英语相比，汉语篇章中的各种意义衔接手段的使用更为丰富灵活。汉语句子之间的意义连接一般很少或不使用关联词语，各个相邻小句之间的核心角色承前和蒙后省略现象非常普遍，各个小句和句子之间完全通过其中的事件转承变化关系连接起来。考虑下面几个真实文本汉语句子描述实例：

(1) 供大于求，价格未能上扬。

(2) 你们年纪还小，(s1)还要成家立业，(s2)不要虚度年华，(s3)更不要成为社会讨厌的人。

(3) 她穿上那件旧花袄，(s1)走出窑来，(s2)解下门扇上的铁链子，(s3)拨开了门闩。

(4) 我无意中碰到了身边的一个什么东西，(s1)伸手一摸(o1)，(s2)是他给我开的饭，两个干硬的馒头。

其中例句(1)描述了无标记的因果关系，两个小句描述的事件之间的因果联系需要通过使用经济学常识推理得到。例句(2)隐含了一个劝诫性因果关系：因为（你们年纪还小，…立业），所以(不要虚度年华，…的人)；同时，后面三个小句的核心主语也承前省略，形成隐含角色共指链：你们—s1—s2—s3。例句(3)描述了汉语中典型的时序连贯关系，顺序发生的多个动作的主体承前省略，形成类似上句的隐含角色共指链。例句(4)的情况更为复杂，各个小句存在着多个核心角色承前蒙后省略隐含情况，但是通过考虑各个小句之间的不同事件关系：小句1-2之间为连贯，小句3-4之间为解注，小句1-2和3-4组合之间为解注，还是可以计算推导出可能的隐含角色共指链：我—s1,东西—o1—s2—馒头。汉语的这种简洁灵活的意合型篇章组合结构，对于人们日常理解交流没有任何困难，但对汉语篇章连贯性描述计算模型则提出了很大的挑战。

本文希望通过挖掘汉语语篇中的各种有效连贯描述形式，分析它们与不同连贯描述内容之间的内在联系，在建立形式和内容相结合的汉语连贯性描述体系方面进行初步探索。在下面几节中，我们首先对国内外的相关研究工作进行分析梳理(第2节)，然后提出我们的解决方案(第3节)，并进行初步的实验验证(第4节)，最后给出相关结论分析(第5节)。

2. 相关研究进展

近年来，通过引进和吸收英语方面的篇章分析理论，在基于关系的汉语语篇的分析标注方面进行了许多有益的探索。乐明(2008)基于RST开发了一个针对汉语财经评论文章的标注库，加工规模为97个篇章[14]。Zhou(2012)针对汉语句子连接词语缺乏的描述现状，对PDTB

体系中的显性和隐性连接词语区分标注方法进行了大幅度调整，直接在相邻句子片段中标注PDTB中定义的各种连贯关系，取得了较好的实验效果[15]。张牧宇等(2013)在PDTB体系上进行适当改良，提出了面向中文的层次化篇章关系体系，对大规模的汉语新闻语料进行了语篇关系标注实验[16]。这些工作初步证实了基于关系的连贯描述在汉语语篇分析标注中的可行性。

从上世纪八十年代开始，许多语言学家也开始从不同角度关注汉语语篇的分析研究。廖秋忠(1992)对汉语语篇中的时空表示、指代成分、指同表达、连接成分、管界问题、论证结构等进行了许多开拓性的研究[17]。邢福义(2001)对汉语复句问题进行了深入研究，提出了因果、并列、转折三分的复句描述体系,并对每个类别下的常用关联词语的描述特点进行了深入分析[19]。另外，吴为章,田小琳(2000)对汉语句群内部组合结构的分析[18]，徐赳赳(2012)对汉语语篇中的零形回指、代词回指、名词回指和联想回指等多种指代描述形式内部关系的深入分析[20]，也可以为我们提供许多有益的借鉴。

与篇章理解相关的另一项重要工作是语言学家对汉语话题和话题链的深入探索。在汉语研究方面，赵元任(Chao, 1968)最先将话题(Topic)引入汉语结构分析研究中[21]，他使用了“话题”和“说明”(Comment)这对概念来解释汉语的主语和谓语结构。Li & Thompson(1976)进一步总结了汉语的话题凸显语言描述特点[22]。曹逢甫(Tsao, 1979)则强调了话题的篇章本性[23]。在汉语篇章中，话题的语义范围可以延伸到小句之外，控制相关话题的代词化和省略形式。汉语话题的这种篇章衔接作用在话题链结构中得到了很好的体现。曹逢甫(Tsao, 1990)最早提出了汉语话题链(Topic Chain)的概念[24]，细致地分析了话题在控制小句连接方面的作用。话题链的形成主要依赖各种指代回指(anaphor)形式,即零形回指(Zero Anaphor, ZA)、代词回指(Pronoun Anaphora, PA)和名词回指(Nominal Anaphor, NA)的选择方法。曲承熹(1998/2005) 总结了前人的研究成果，提出了以下操作性较强的话题链定义“一组以零回指ZA 形式的话题连接起来的小句”([25],p259)。

话题链分析中另一个需要关注的是观察者视域(perspective)问题。复杂语篇中往往会出现多个视域交叉的情况。其中有的视域只与个别的段落、句子发生关系，有的视域却影响到了整个篇章的结构。刘大为(2004)初步总结了两类进行视域描述的动词：言说动词（说、告诉、讲解、讲述、宣称等）和意向动词（认为、相信、知道、希望、喜欢、害怕等）[26]。杨彬(2009)进一步总结形成了汉语中常用的言说动词和意向动词表[27]。通过对真实文本句子中这些动词的管界内容[28]的深入分析，可以对语篇中描述的不同视域中的不同事件内容进行有序组织。

尽管许多语言学家都强调了话题链对汉语语篇描述的独特作用,近年来的深入研究也发现英语中实际上也存在类似汉语话题链的篇章组织结构。孙坤(2013)对英汉篇章组织模式进行了对比研究[29]。王建国(2013)把话题链的描述作用从句子拓展到超句(句群)和篇章，重新定义话题链为“由同一话题引导的系列语句”，并深入分析了话题链在汉英语篇中的不同描述特点[30]。刘礼进(2011)使用人工标注的小规模汉英篇章对比语料库，深入分析了话题链在汉英篇章的宏观语义结构描述功能上的差异情况[31]。

在汉语语篇结构的计算分析研究方面,舒江波(2011)以邢福义(2001)提出的复句理论为指导，对汉语复句关联词的自动识别方法进行了研究和探索[33]。宋柔(2012)提出了汉语广义话题结构模型，从标点句入手分析了汉语句子相邻小句片段中的话题隐现情况，总结出了若干有效的基于堆栈结构的回指话题恢复策略[32]。张明尧(2013)提出了基于事件链的篇章语义表示模型，通过对篇章中共指实体链的分析标注，自动识别这些共指实体相关的事件链，初步构建了基于事件链的篇章连贯性计算模型[34]。

通过对以上汉语语篇研究工作的简单综述，我们发现：1)以修辞关系描述为主体的 RST 可以很好地分析和标注汉英语篇的语义结构和交际功能,其开放关系标记集的设计理念使它更适合于篇章结构生成的研究,而形式描述手段的缺乏则制约了它在语篇连贯性分析计算方面的应用潜力；2) PDTB 选择的关联词语描述切入点很好地解决了连贯形式和内容的结合问

题，语言学家在汉语复句和句群研究中积累的丰富关联词语描述信息可以与 PDTB 模型形成内容衔接，但汉语真实文本中关联词语使用范围狭窄的现状限制了这种描述体系在汉语语篇分析中的应用效果；3) 理论语言学家对汉语话题链的深入研究，已初步形成了一套可操作的汉语语篇连贯性描述框架。计算语言学家在汉语广义话题结构分析和实体链、事件链上的计算探索又初步证明了其可计算性。把它引入汉语语篇连贯性分析计算模型中，应该可以为相关模型的改进和完善提供新的活力。

基于以上几点考虑，我们希望能把话题链引入汉语语篇连贯性描述体系中，形成一套以话题链为主，融合关联词语和其他连贯形式描述机制，重构现有连贯关系描述集。

3. 连贯性描述体系

3.1 研究对象分析

传统的语篇研究对象包括书面文本和口语对话两大部分。我们的研究对象则主要集中在其中的书面文本部分，重点探索对新闻、学术、文学和应用等体裁的叙事、说明、描写等类型文本中的事件情景连贯特征的分析计算问题。

在汉语语篇的“字/词→小句→句子→段落→篇章”的各个分析描述层次中，根据汉语语篇目前的研究现状，把研究范围限制在中间的“小句→句子→段落”部分，重点分析书面文本段落内部以各个小句描述的基本事件为基础构建形成的复杂事件关系网络。初步定义事件描述小句(Event Description Clause, EDC)为汉语语篇中以逗号、分号、句号等点号分隔的词语序列，其结构组合主要包括：1) 包含谓语成分的小句；2) 句首的主语或话题成分；3) 句首或句中的状语或状语从句。

这里定义的 EDC，基本上与宋柔(2012)中定义的标点句相当，主要差别在于我们的 EDC 包含了由逗号分隔的体词性并列成分，以避免相应的不完整标点句对后续的篇章结构分析的影响。从描述内容上看，EDC 大部分又都可以归入沈家煊(2012)定义的“零句”形式[35]，其中通过标点划分出的话题和状语从句部分，可以很好地融入后续的连贯性分析计算框架中。

为便于后续的计算处理，进一步引入下面两个中间处理层次：1)事件句式(Event Construction, EC)，把它作为 EDC 中描述基本事件内容的句法语义链接(Syntax-Semantics Linking, SSL)结合体。其中融合了浅层的主状谓宾补等句法功能结构和深层的谓词论元结构(Predicate-Argument Structure, PAS)[36]。虽然大多数简单 EDC 中只包含一个 EC，但汉语真实文本小句中也存在许多复杂的 EDC 组合，其中的多个 EC 会形成并列、连谓、兼语、述结、定语从句嵌套等复杂结构关系，与“小句→句子”的组合关系有很强的相似性。因此，我们把它们作为语篇连贯性分析的基本单元；2) 句群(Sentence Group, SG)，是汉语段落中多个句子组合形成的针对同一话题展开的、前后衔接、语义连贯、具有一定交际目的和功能的篇章描述单元。它们基本上与汉语语言学家定义的“句群”概念相当，只是更强调了句群片段描述意义的内部完整性和外部功能性。它们可以作为句子到段落分析的中间计算单元。

这样，就可以把本文关注的汉语语篇连贯性描述体系分为以下处理阶段：1)“事件句式→小句→句子”；2) “句子→句群→段落”。每个阶段的连贯性分析描述又有不同侧重点：

在“事件句式→小句”层面，研究重点是对一些特殊句式结构中的多个事件之间复杂关系的分析和把握，包括以时序关系连接的连谓结构、以“动作-结果”为基础的述补结构、以事件内容同义平行形成的并列结构、以前后景配置[25]形成的定语从句关系化结构以及表示广义使役关系的使字句式、得字句式、把字句式、重动句式等[37]。为了深入挖掘其中的复杂事件关系，可能还要涉及基于词汇语义学的谓词意义分解和基于事件语义学的逻辑表达式重构等研究。这部分内容不是本文的描述重点。

在“小句→句子→句群”层面，研究重点是挖掘汉语真实文本中有效的连贯描述形式，选择可以尽可能多地覆盖全部句子和句群的连贯内容的连贯关系描述集合，分析不同的连贯

形式和关系内容描述之间的内在联系，逐步形成适合汉语的形式内容相结合的连贯性描述体系。这是本文的研究重点，相关内容将在后面进一步展开。

在“句群→段落”层面，需要研究段落中不同句子或句群描述的话题之间的内在关系，构建段落内部不同话题的演化网络，探索不同句子和句群中的微观话题及其描述内容组合形成整个段落的宏观主题(theme)和中心思想的内在机制。这部分内容也不是本文的描述重点。

3.2 描述体系介绍

在“小句→句子→句群”层面，我们的设计目标是：对句子或句群中的每个事件描述片段给出合理的连贯性解释。这种解释可以通过对这些片段寻找有效的连贯描述形式，确定合适的修辞关系标记，分析内部事件组合层次，构建事件关系网络等步骤来实现。这里的基本假设是：汉语真实文本中的连贯性描述片段都是有其相应的连贯描述形式支撑的。我们的研究重点就是如何挖掘出这些连贯描述形式和手段，建立它们与不同的连贯关系内容描述之间的内在联系，为进一步构建可计算的汉语语篇连贯性分析模型打下基础。

为此，我们从前人的研究成果中，提炼出了如下几种汉语连贯描述形式：1) 话题链；2) 关联词语；3) 其他连贯形式。并以此为基础，构建了我们的连贯关系描述体系。下面对相关内容进行简要说明：

1) **话题链**：主要作用是连接各个小句或句子。综合曲承熹(1998/2005)和王建国(2013)的研究成果，我们提出了论文中使用的话题链概念的操作性定义：一组以 ZA、PA 或 NA 形式的话题连接起来的小句或句子。在句子内部的各个小句之间形成的话题链，主要以 ZA 形式表示。而在句群内部的各个句子之间，则更多地会采用 PA 或 NA 形式。由于链首话题的不同导入方式，句子或句群内部的话题链会形成不同的内部结构，它们可以为不同的修辞关系内容解释提供真实理据支撑[25]。

2) **关联词语**：主要作用是连接各个小句或句子，同时显性标识其可能的修辞关系。因此在许多汉语真实文本句子中，关联词语会与话题链同时出现，用于凸显话题链中描述的各个小句片段之间需要强调的修辞关系，特别是在话题链中描述的信息违反常规的后景到前景的变化流程时。

3) **其他连贯形式**：主要作用是提供话题链和关联词语之外的其他连贯性判据，包括：

- 实体链：将汉语小句或句子中话题位置之外的其他具有共指关系的实体成分连接起来的共指实例链，类似张明尧(2013)中定义的实体链，显示小句和句子描述内容之间的实体衔接关系；
- 平行结构：多个内部结构相似的小句或句子并置在一起，体现其描述内容之间的对等或对比关系，一般使用顿号、逗号或分号等点号来分隔；
- 谓词组合：通过谓词所带的“了、着、过”等体标记的不同反映相应谓词小句之间的前后景关系[25]。如：持续体标记‘-着’一般表示后景事件，而完成体标记‘-了’则大多标识前景事件。

在以上几种连贯形式中，我们认为话题链和关联词语是汉语语篇中使用的主要连贯形式，它们是建构汉语连贯性描述体系的基础。而其他连贯形式则是辅助性的，它们通过与话题链和关联词形式配合使用或单独使用，凸显某些特殊的连贯表示结构。

以此为基础，我们重新建构了新的连贯内容描述体系。它包括四大修辞关系描述：1) 话题评述关系；2) 广义并列关系；3) 广义因果关系；4) 广义转折关系。其中话题评述关系主要对应于话题链形式，通过设置不同内部子关系层次对不同话题链体现的事件前后景分布特点进行详细描述。有关内容将在下面进一步展开。广义并列、因果和转折关系主要对应关联词形式，基本上沿用了邢福义(2001)提出的复句三分体系，并基于我们的理解进行了适当调整，如：将递进关系从原来的广义并列关系集移到广义转折关系集，共同与原有的转折关系

形成顺转和逆转的对比描述集合。

在话题评述关系集中，根据不同话题链描述特点，又区分出以下几种子关系描述：

1) 时空顺序关系

针对同一话题描述的多个事件在时间轴和空间体上形成的事实理据顺序关系。其话题链大多是由首句（小句）主语为基准话题形成的单一 ZA 链。这是汉语话题链的主要描述形式。

2) 解释注解关系

对新导出的话题的描述内容进行进一步的解释说明。其话题链主要是由话题导出句（小句）宾语为基准话题形成的单一 ZA 链。典型使用场景是在更大的主话题链中作为一条子话题链，对主话题链描述的前景主线中的某个特殊实体的相关背景进行介绍，形成后景描述。

3) 视域变换关系

通过视域动词的使用，将句子（句群）描述的内容分成两个不同视域，其中分别形成不同的话题链描述相应事件内容，两者通过视域动词建立起内在联系。典型实例是由‘说’、‘宣布’等言说动词引导的转述结构。

至此，我们初步形成了一个形式和内容相结合的汉语连贯性描述模型：在连贯形式方面，提取了话题链、关联词和实体链、平行结构、谓词组合等其他形式；在连贯内容方面，构建了话题评述、广义并列、广义因果和广义转折等四大修辞关系描述集。下面通过第 1 节中列出的几个实例的具体分析，对这个体系的形式内容结合描述特点进行简要说明：

(1) 例句：她穿上那件旧花袄，走出窑来，解下门扇上的铁链子，拨开了门闩。

- 连贯形式：话题链“她-ZA-ZA-ZA”，完成体标记‘-了’；
- 连贯内容：话题评述—时空顺序，话题‘她’顺序完成的多个动作；

(2) 例句：你们年纪还小，还要成家立业，不要虚度年华，更不要成为社会讨厌的人。

- 连贯形式：话题链“你们-ZA-ZA-ZA”，关联词语“还、更”，平行结构“不要…，不要…”；
- 连贯内容：并列+并列→因果

(3) 例句：我无意中碰到了身边的一个什么东西，伸手一摸，是他给我开的饭，两个干硬的馒头。

- 连贯形式：主话题链“我-ZA”，完成体标记‘-了’，次话题链“Φ-ZA”，实体链“东西-Φ-饭-馒头”；（Φ 表示不在此句子中出现的隐含话题）
- 连贯内容：主话题链描述时空顺序关系，次话题链描述解释注解关系

4. 实验验证与分析

4.1 实验设计

我们以清华句法树库 TCT Ver 1.0[38]的全部标注句子作为实验数据来验证相关体系描述的可行性。TCT 选择了新闻、学术、文学和应用等四种体裁的汉语平衡语料文本进行了句法结构树的分析和标注。总标注规模为 100 万词，约 4.7 万句。TCT 除了标注小句层面的名词短语(np), 动词短语(vp)等句法结构信息外，还设计了包含 11 种关系标记的复句描述体系[38]，对汉语复句内部的各种事件逻辑关系进行了详细描述，并对一些特殊引述句中的复杂句群组合关系进行了初步描述，为我们进行汉语“小句→句子→句群”层面的连贯性分析描述打下了很好的基础。

我们提取了 TCT 中所有标注了以下 11 种事件关系的复句(fj)成分：并列(BL)、选择(XZ)、连贯(LG)、递进(DJ)、因果(YG)、目的(MD)、条件(TJ)、假设(JS)、转折(ZE)、解注(JZ)、流水(LS)等。为了更有效获取这些复句内部的连贯性表示形式，我们对它们进行了以下预处理：

首先，自顶向下提取复句控制的所有子成分，包括内部事件小句 EDC 和嵌套复句(i-fj)，形成复句内部小句块序列：EDC* + i-fj*；

对每个内部 EDC，进一步提取其中的主状谓宾块等形成的事件句式 SDPO[1]；对内部嵌套复句，只提取其控制的第一个内部 EDC 的相应事件句式作为代表；

对每个复句内部子成分块（EDC 或 i-fj），设计了如下简单的内部连贯性判据：

- 如果该小句事件句式没有主语块，则判定为存在零形回指话题(ZAT)形式；
- 如果该小句句首和状语块中包含关联词语(CW)，包括：连词(c)、连接语(l)和关联副词(d)[2]等，则判定为存在关联词(CW)形式；

考虑到嵌套复句内部多个关联词语使用的歧义性，规定该复句句首的连词和连接语只在嵌套复句层面起作用，在其内部 EDC 序列的连贯性状态分析时不起作用。

据此，按照复句内部各个小句块的连贯性判据值，可以把所有复句分成以下 4 类：

1：只通过话题链连接：内部小句包含一个以上 ZAT，并且不包含任何 CW；

2：只通过关联词连接：内部小句包含一个以上 CW，并且不包含任何 ZAT；

3：同时通过话题链和关联词连接：内部小句同时包含一个以上 ZAT 和 CW；

4：通过其他方式连接：内部小句不包含任何 ZAT 和 CW；

4.2 实验数据

表 1 列出了目前获得的完整统计数据，从中可以看出目前论文关注的三种连贯形式在汉语真实文本句子中的大致分布特点：

1) ZA 形式话题链是汉语复句的主要连贯形式，覆盖 75.92%的汉语句子（1 类—36.10%，3 类—39.82%）；
2) 关联词语也是汉语复句的重要连贯形式，覆盖 49.67%的汉语句子（2 类—9.85%，3 类—39.82%）；其中关联副词的贡献达到了 22.78%（2 类—5.85%，3 类—16.93%），显示了它们在汉语句子连贯性描述方面的重要作用；
3) 汉语句子中关联词语与 ZA 话题链同时使用是其应用常态，占其覆盖句子的 80%左右，初步证实了关联词语在凸显话题链描述的不同事件关系中的重要作用[25]；
4) 使用其他连贯形式的复句约占 14.23%，主要分布在流水和并列复句中，其中的不同连贯形式描述特点需要在后续工作中进一步深入分析。

表 1 包含不同连贯标记的 TCT 不同复句关系统计

	LG	LS	BL	XZ	DJ	YG	MD	TJ	JS	ZE	JZ	Total
1	6670	3034	1663	35	320	236	63	47	222	148	409	12847
2	285	1217	605	9	123	289	3	81	156	641	97	3506
3	3720	3131	1960	79	1196	1038	176	338	598	1431	504	14171
4	669	2696	1213	2	75	104	3	20	49	78	157	5066
合计	11344	10078	5441	125	1714	1667	245	486	1025	2298	1167	35590

为了更好地显示不同连贯形式与连贯内容之间的对应关系，我们按照 TCT 标注规范中给出的 11 类复句关系的描述特点，将它们初步映射到上节定义的 4 种主要修辞关系类，形成以下 4 大类事件关系描述集合：

1) 话题评述关系：映射连贯(LG)、流水(LS)和解注(JZ)3 种关系，分别对应时空顺序、视域变换、解释注解等关系小类；
2) 广义因果关系：映射因果(YG)、目的(MD)、条件(TJ)和假设(JS)4 种关系，分别对应相应的描述小类；

---

[1] S—主语块，D--状语块，P—谓语块，O—宾语块。

[2] 目前主要考虑了以下关联副词："便","才","倒","都","非","就","马上","却","也","一","又","越","凡是","不论","尽管","即使","就是","虽然","早在","刚","仍然"。

3) 广义转折关系：映射递进(DJ)和转折(ZE)2 种关系，分别对应顺转和逆转两个小类；

4) 广义并列关系：映射并列(BL)和选择(XZ) 2 种关系，分别对应相应的描述小类。

这样，我们可以把表 1 内容归并形成表 2 数据。从中可以看出，在 TCT 数据集上，我们目前提出的三种连贯形式和四种修辞关系之间存在很强的对应联系：

- 话题链是话题评述关系的凸显描述形式，覆盖相应句子实例的 77%以上；而该类句子在真实文本中的分布比例也达到了 63.47%，因此研究话题链和话题评述关系的互动作用效果对理解真实文本中大部分句子的连贯性描述特点具有重要意义；
- 关联词是广义因果和转折关系的凸显描述形式，覆盖相应句子实例的 82%左右；但该类句子在真实文本中的分布比例只有 20.89%，这就使其发挥作用的范围受到了很大限制；
- 相对而言，广义并列关系句子中各种连贯手段的应用相对平均，话题链、并列连词、平行结构等多种连贯形式都会在广义并列关系的识别理解中发挥作用。而且其在真实文本中的分布比例也达到了 15.64%，需要对其内部连贯特点进行进一步分析。

表 2 包含不同连贯标记的 4 大类映射连贯关系复句统计

	话题评述	广义因果	广义转折	广义并列
1	10113	568	468	1698
2	1599	529	764	614
3	7355	2150	2627	2039
4	3522	176	153	1215
合计	22589	3423	4012	5566

为了进一步分析话题链和关联词两种连贯形式在不同体裁的汉语真实文本中的使用特点，我们分别统计了它们在不同体裁的文本句子中描述 4 类不同连贯关系时的分布比率，得到了图 1 和图 2 的数据结果。从中可以看出：

话题链在不同体裁的不同连贯关系复句中应用很均衡，在话题评述、广义因果和广义转折复句中的应用比例都达到了 75%以上，在广义并列复句中的应用比例也达到了 65%以上，显示了其在汉语句子的连贯性描述计算中的重要作用；

关联词在不同体裁的不同连贯关系复句中的应用则不太均衡：在不同体裁文本中，学术类句子使用相对较多，以适合学术类内容描述的严谨性要求；而侧重事务描述的应用类句子中则使用较少。在不同连贯关系复句中，关联词在广义因果、转折和并列复句中使用较多，而在话题评述类复句中则使用较少，因为其中的话题链已经可以提供很好的连贯性描述支持。

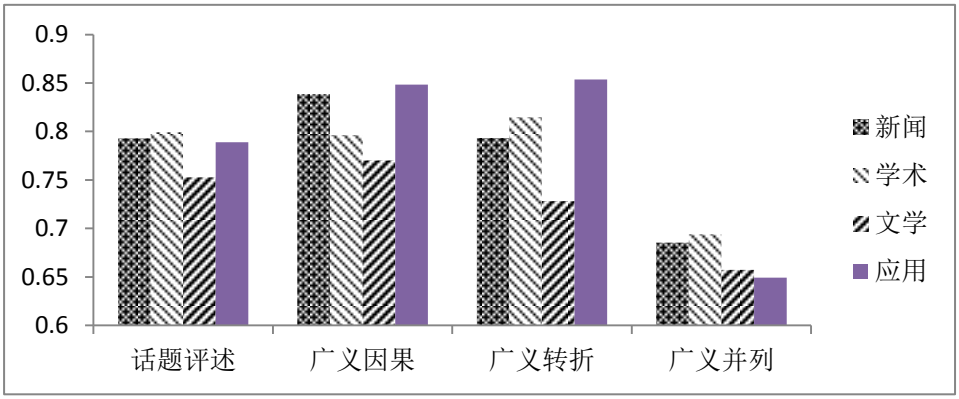

图 1 话题链在不同体裁的 4 大类连贯复句中的使用分布率

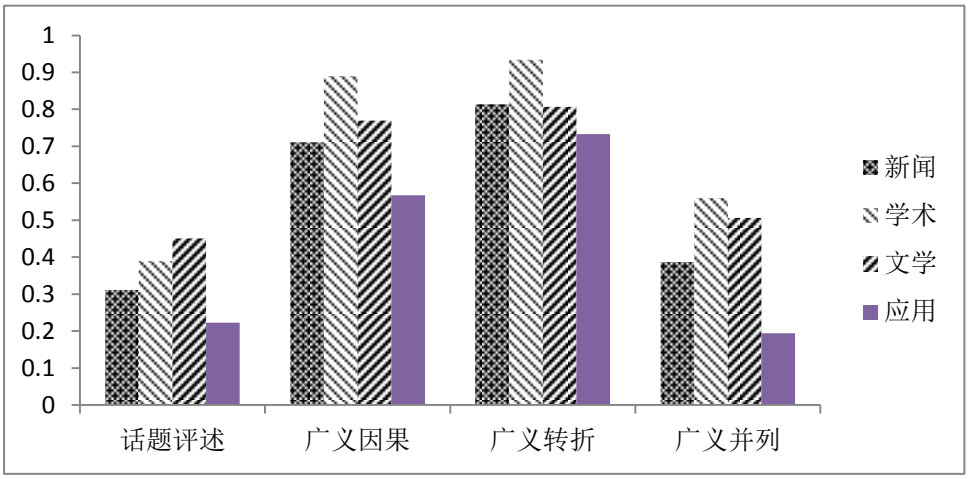

图 2 关联词在不同体裁的 4 大类连贯复句中的使用分布率

4.3 分析与讨论

表 3 列出了从目前的自动分析结果中随机抽出的 8 个复句片段的自动分类数据。从中可以看出，尽管我们目前只使用了简单的连贯形式判据，但获取到的分析数据基本上还是符合我们预期的。其中，例句 3 反映的广义条件关系，需要通过对话题链描述内容推断得到，这将是后续工作的一个研究重点。而例句 7 则是由于对嵌套复句句首关联词语的简单排歧规则处理而导致的类别 3 漏判断。据此，我们初步判断目前得到的相关结论还是比较可靠的。

表 3 各类体裁文本随机选取的 2 个复句分析实例（关系=TCT 复句关系；类别=自动识别类）

序号	体裁	关系	类别	复句片段
1	学术	LG	1	为了维持它的存在并实现其职能，通过产品分配和再分配命名其掌握必要数量的社会产品，用于保证国家机器运转的需要
2	学术	YG	2	这一派认为，人之生气以阳为主，治病则应重用温药和补药
3	新闻	TJ	1	要使这项工作抓出成果，首先需要这些部门的主要领导以身作则、身先士卒
4	新闻	MD	3	巴解组织要求以色列政府收缴犹太移民的武器，以制止被占领土暴力事件的不断发生
5	小说	LS	2	程玲好不容易才看明白，过振福奇迹般地为她办妥了转到河北平原插队落户的手续
6	小说	BL	4	京城豪门，山珍海味不新鲜，新鲜的反倒是地方风味小吃
7	应用	JS	1	如果每天安排 1 — 1 ． 5 小时，不到一年即可完成
8	应用	LS	4	您可将报款寄给北京青年报小红帽报刊发行服务公司（不另收邮寄费），我们将给您寄递

目前在汉语真实文本上的话题链和关联词使用统计数据还比较少。宋柔(2012)在 40 万字左右的广义话题结构标注库上的统计结果显示，汉语篇章中 40%左右的标点句首部缺少话题[32]。Zhou(2012)对 CTB 标注库中随机抽取的 20 个语篇文件进行了分析，发现 82%的复句使用了隐性关联词，与英语 PDTB 标注库得到的 54.5%的数据有很大差距[15]。这些数据从不同侧面验证了汉语文本中 ZA 话题使用频繁、关联词语使用较少的分布特点，与我们的实验结果可以互为验证。

5. 结论

本文针对汉语篇章结构简洁灵活、很少使用关联词语的描述特点，提出引入话题链描述形式，设计不同类型的话题评述关系集，构建了以话题链为主，融合关联词语和其他连贯形

式描述机制，覆盖话题评述、并列、因果、转折四大类关系的汉语语篇连贯性描述体系。在清华句法树库 TCT 上进行的初步验证实验表明，话题链在不同体裁的汉语真实文本数据上都有很好的适用性，可以很好地解决显性关联词不足导致的连贯性判据缺失问题。

在后续研究中，我们将在这个描述体系指导下，重构 TCT 标注库中“小句→句子”层面的标注信息，发现并标注句子中的不同话题链，据此确定合适的句子连贯关系标记。构建新的融合话题链、关联词和其他连贯形式的汉语复句连贯性标注库，为进一步探索高效的汉语句子连贯性计算模型打下基础。

参考文献

[1] B.Webber, A. Joshi (2012) Discourse Structures and Computations: Past, Present and Future [A]. In Proc. of ACL-2012 Special Workshop on Rediscovering 50 Years of Discoveries [C], pp. 42-54.

[2] B. Webber, M.Egg and V. Kordoni (2012) Discourse structure and language technology [J]. Natural Language Engineering. 18(4): 437-439. Cambridge University Press.

[3] Mann W C, Thompson S A. (1998) Rhetorical Structure Theory: Toward a functional theory of text organization [J]. Text, 8(3):243-281.

[4] Carlson L, Marcu D, Okurowski M E. (2001) Building a Discourse-Tagged Corpus in the Framework of Rhetorical Structure Theory [A]. Proceedings of the Annual Sigdial Meeting on Discourse and Dialogue [C], Morristown: Association for Computational Linguiscs, p30-39.

[5] DuVerle D A, Prendinger H. (2009) A Novel Discourse Parser Based on Support Vector Machine Classification [A]. *Proc. of ACL-IJCNLP 2009* [C]. Morristown: ACL, pp. 665-673.

[6] Prasad R, Miltsakaki E,Dinesh N,Lee A,Joshi A,Robaldo L,Webber B. (2008) The Penn Discourse Treebank 2.0 Annotation Manual[R]. USA: University of Pennsylvania

[7] Lin ZH, Ng H T, Kan M Y. (2010) A PDTB-styled end-to-end discourse parser [D]. Singapore: National University of Singapore.

[8] Zhou ZM, Xu Y, Niu ZY, Lan M, Su J, Tan C L. (2010) Predicting Discourse Connectives for Implicit Discourse Relation Recognition[A]. Proceedings of the 23rd International Conference on Computational Linguistics [C]. Morristown: Association for Computational Linguistics, pp1507-1514.

[9] J. Ruppenhofer, C. Sporleder, R. Morante, C. Baker, M. Palmer. (2010). SemEval-2010 Task 10: Linking Events and Their Participants in Discourse [A]. In Proc. of SemEval-2010 [C], 45-50.

[10] Josef Ruppenhofer, Philip Gorinski, and Caroline Sporleder. (2011) In Search of Missing Arguments: A Linguistics Approach [A]. In Proc. of RANLP-2011 [C], pp. 331-338.

[11] Sameer Pradhan, Lance Ramshaw, Mitch Marcus, Martha Palmer , Ralph Weischedel and Nianwen Xue (2011). Modeling Unrestricted Coreference in OntoNotes [A]. In Proc. of CoNLL-2011 [C]. p1-27.

[12] Ralph Weischedel, Eduard Hovy, Martha Palmer, Mitch Marcus, Robert Belvin, Sameer Pradhan, Lance Ramshaw, and Nianwen Xue. (2011). OntoNotes: A Large Training Corpus for Enhanced Processing [A]. In Joseph Olive, Caitlin Christianson, and John McCary, editors, Handbook of Natural Language Processing and Machine Translation [C]. Springer.

[13] Halliday, M.A.K. and Hasan, R. (1976) Cohesion in English [M]. London: Longman.

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

[15] Yuping Zhou, Nianwen Xue. (2012) PTDB-style Discourse Annotation of Chinese Text[A]. In Proc. of ACL-2012 [C], p. 69-77.

[16] 张牧宇、秦兵、刘挺 (2013) 中文篇章关系体系与标注实践[J]. 中文信息学报. 2013

[17] 廖秋忠 (1992) 廖秋忠文集 [M]. 北京：北京语言学院出版社

[18] 吴为章,田小琳 (2000) 汉语句群[M].北京：商务印书馆

[19] 邢福义 (2001) 汉语复句研究[M]. 北京：商务印书馆
[20] 徐赳赳 (2012) 现代汉语篇章语言学[M]. 北京：商务印书馆
[21] Chao, Yuan Ren (赵元任) (1968) A Grammar of Spoken Chinese [M]. Berkeley and Los Angeles: University of California Press.
[22] Li, Charles N. and Sandra A. Thompson (1976). Subject and Topic [M]. New York: Academic Press.
[23] Tsao, Feng-fu (曹逢甫) (1979) A Functional Study of Topic in Chinese: the First Step toward Discourse Analysis [M]. Taipei: Student Book Co.
[24] Tsao, Feng-fu (曹逢甫) (1990) Clause and Sentence Structure in Chinese: A Functional Perspective [M]. Taipei: Student Book Co.
[25] 曲承熹 (1998/2005) 汉语篇章语法[M]. 北京：北京语言大学出版社 （潘文国等译）
[26] 刘大为 (2004) 意向动词、言说动词与篇章的视域[J], 修辞学习 2004 年第 6 期(总 126 期)
[27] 杨彬 (2009) 话题链语篇构建机制的多角度研究[D]，上海：复旦大学博士论文
[28] 廖秋忠 (1992) 篇章中的管界问题 [A]. 北京语言学院出版社：《廖秋忠文集》 [C]，pp92-115.
[29] 孙坤 (2013) 话题链视角下的汉英篇章组织模式对比研究[J]，解放军外国语学院学报，第36 卷第3 期，2013 年5 月
[30] 王建国 (2013) 论话题的延续：基于话题链的汉英篇章研究[M]. 上海：上海交通大学出版社
[31] 刘礼进 (2011) 英汉篇章结构模式对比研究[A]. 刘礼进著《英汉语篇和语法问题研究》[C]，中山大学出版社，p166-178.
[32] 宋柔 (2012) 汉语篇章广义话题结构研究[R]，北京语言文化大学内部资料
[33] 舒江波(2011) 面向中文信息处理的复句关联词自动识别[D]，武汉：华中师范大学博士学位论文
[34] 张明尧(2013) 基于事件链的语篇连贯研究[D]，武汉：武汉大学博士学位论文
[35] 沈家煊 (2012) “零句”和“流水句”[J]. 中国语文 2012 年第 5 期(总第 350 期)
[36] 邱晗 (2014) 汉语谓词论元结构的分析标注研究[D]. 北京：清华大学硕士论文
[37] 吴平 (2011) 汉语特殊句式的事件语义分析与计算[M]. 北京：中国社会科学出版社
[38] 周强 (2004) 汉语句法树库标注体系[J]. 中文信息学报 18(4), 1-8

作者简介：

作者一周强（1967——），男，博士，研究员，主要研究领域为自然语言理解、语料库语言学、词汇语义学、机器学习。Email: zq-lxd@mail.tsinghua.edu.cn；

作者二周骁聪（1991——），男，本科生，主要研究领域为自然语言处理、机器学习。