中山大学高校基本科研业务费青年教师培育项目资金资助
中山大学人文社会科学出版基金资助

现代汉语书面语中跨标点句句法关系约束条件的研究

张瑞朋◎著

中国社会科学出版社

基本科研业务费青年教师培育项目资金资助
中山大学人文社会科学出版基金资助

现代汉语书面语中跨标点句句法关系约束条件的研究

张瑞朋◎著

中国社会科学出版社

图书在版编目(CIP)数据

现代汉语书面语中跨标点句句法关系约束条件的研究／张瑞朋著．
北京：中国社会科学出版社，2013.9
ISBN 978-7-5161-3241-8

Ⅰ.①现… Ⅱ.①张… Ⅲ.①现代汉语-书面语-研究
Ⅳ.①H109.4

中国版本图书馆 CIP 数据核字(2013)第 219502 号

出 版 人 赵剑英
责任编辑 任 明
特约编辑 李燕丽
责任校对 方春梅
责任印制 李

出 版
社 址 北京[illegible]158 号（邮编 100720）
网 址 http://www.csspw.cn
中文域名：中国社科网 010-64070619
发 行 部 010-84083685
门 市 部 010-84029450
经 销 新华书店及其他书店

印 刷 北京市兴怀印刷厂
装 订 北京市兴怀印刷厂
版 次 2013 年 9 月第 1 版
印 次 2013 年 9 月第 1 次印刷

开 本 710×1000 1/16
印 张 15.5
插 页 2
字 数 243 千字
定 价 45.00 元

凡购买中国社会科学出版社图书，如有质量问题请与本社联系调换
电话：010-64009791

内容简介

本书是在跨标点句句法关系的理论框架下展开的，主要目的是解决跨标点句共享成分的识别问题，找出这类句法关系在满足栈形规律之外，还应满足哪些可以形式化的约束规则，以便计算机处理。

本书的主要特色表现在以下几个方面：研究范围方面，除了前人已有的研究跨越标点句的主谓关系之外，还研究了跨越标点句的定中关系、状中关系、述宾关系、述补关系、介宾关系等，全面铺开了跨标点句的句法体系的研究。研究角度方面，侧重于约束条件中的形式化特征，研究成果具有较强的可操作性，为计算机自动进行跨标点句句法关系的分析打下了一定的基础。研究方法方面，不满足于举例说明。除了使用传统的自省方法，寻找语言规律的认知理据外，重视真实语料的语言现象统计，以统计数据作为规律可靠性的佐证。

本书的突出特点表现在语言特征的多角度的深入挖掘方面。

作者简介

张瑞朋，女，1979年生，河南灵宝人。2001—2004年就读于华中科技大学语言学及应用语言学专业，获硕士学位。2004—2007年就读于北京语言大学语言学及应用语言学专业，获博士学位。

现为中山大学国际汉语学院讲师，主要从事对外汉语教学、语言信息处理、语言学及应用语言学研究。目前已主持校级以上课题7项，其中国家社科基金青年项目1项。参与国家级项目4项。参与教材编写2部。公开发表论文近20篇。

出版策划：任　明
封面设计：戴东明

序

张瑞朋的这本专著，研究的是现代汉语文本中跨越标点句的句法关系的约束条件。这是一项语言学范畴的工作，目的是为计算机自动分析篇章结构做准备。其中涉及几个概念，这里分别说说我的看法。

首先说标点句。标点句是汉语文本中被逗号、分号、句号、叹号、问号、直接引语的冒号和引号所隔开的词语串。为什么放着语言学中常说的“句子”、“子句”、“小句”等术语不用，非要造一个叫作标点句的新术语？一是因为语言学中这些已有的术语各人定义不一致，二是因为这些术语涉及的有些概念难以在操作层面上界定。作为篇章处理的基本单位而缺乏可操作性，那么后续的工作就无法进行了。标点句基本上是无歧义的。同一篇汉语文本，不同人点的标点可能有很大不同，这没关系。我们说的无歧义的意思是，对于任何正常的点好了标点的文本，抓取标点句就是机械式的了。事实证明，这样得到的标点句，具有明显的规律性。其实，标点句的概念并非凭空而出。赵元任在《汉语口语语法》中把两头被有意的停顿限定的一截话语看作句子，这种有意的停顿在文本中显然表现为标点。陈平在《汉语零形回指的话语分析》中以标点符号为标记，把用逗号、句号、问号等断开的语段算作小句，这明显地就是标点句。称作“标点句”可以明确地与各种已有的相近术语区别开来，并具有顾名思义的效果。

这里涉及一个更大的问题，就是我们的研究对象是什么。语言学当然是研究语言的，语言的实例是言语，言语有语音和文本这两种外在形式。对于拼音文字的语言来说，这两种形式具有天然的联系，文本的符号形式同语音密切相关，文本的语法直接受制于语音。如词的分隔、词的归类（part of speech）、词的句法角色和语义角色、句子的句法结构，

都在不同程度上具有语音方面的依据，而且这种依据还以文字符号的形式在文本中显现。汉语就不同了。汉语文本的文字符号和语音没有必然联系。但是，长期以来，汉语的语法学家是在语音的基础上研究文本语法的。比如讲词长的时候讲的是单音节词、多音节词，研究句法、话题时往往要使用停顿、音调等语音特征。

计算机处理汉语文本，面对的是字符序列。最基本的单位是字，再往上的具有形式标记的单位就是标点句。如果要区分词长就得说单字词、多字词，并且无法直接使用停顿、音调等语音特征，因为这些特征在文本中并无直接表现。于是，作句法分析以字为基础，作篇章分析以标点句为基础，是最自然不过的事情。这样建立的语法研究的体系和方法也许可以叫作文本语法学。文本语法学的特点是完全基于文本中可见的符号，不能直接使用语音特征。这里，并不是说在字的上面不再需要词语这样的语法单位，也不是说从标点句可以直接分析段落和篇章，并且也不排除通过知识库在文字符号和语音之间建立关联进而利用语音特征作语法研究。但是，研究句法的原始出发点只能是字，研究篇章语法的原始出发点只能是标点句，这是我们不得不面对的客观现实。

其次说跨越标点句的句法关系。标点句在语法和语义上往往不能独立，会有一些句法结构是跨越标点句的。下面的几个例子都是《围城》中的实例，取自于书中第 10 章关于把字句的内容。

刘小姐把她拉进去了，

自信没给客人瞧见脸色。

第 2 标点句共享第 1 标点句的主语“刘小姐”。

火铺里晚上不点灯，

把一长片木柴烧着了一头，

插在泥堆上。

第 2 标点句共享第 1 标点句的处所和时间主语“火铺里晚上”，第 3 标点句共享第 2 标点句的介宾短语形式的状语“把一长片木柴”。

这把**五人**吓坏了，

跟办事员讲了许多好话，

第2标点句以第1标点句的介词宾语“五人”为主语。

切不可锦上添花，
让学生**把**分数看得太贱，
功课看得太容易—

第2标点句共享第1标点句的动词短语“切不可”，第3标点句共享第2标点句的介词“把”。

山里的寒气把旅客们的睡眠**冻得**收缩，
不够包裹整个身心。

第2标点句共享第1标点句中组合式述补短语的述语与助词组合“冻得”。

这些例子每个只有两三个标点句，实际上整篇的文本都可以依据跨越标点句的句法关系，用这种换行缩进图式展示出一种跨越句子的语法结构。张瑞朋的书中引用了一些比较长的篇章片段。

语言的结构复杂性之一在于递归。要让人能使用语言交际，就必须得让听者/读者识别出这种递归结构。这就需要形式标记。英语中的关系代词、关系副词、介词、动词的不定形式和分词形式、名词短语前的冠词等都属于这种形式标记。汉语没有这类形式标记，怎么能让说者/写者组织起长篇文本，而听者/读者又能即时地理解呢？这本书的工作揭示，跨越标点句的句法关系是支持这一认知过程的重要机制。

最后说约束条件。当我们看出标点句之间存在句法关系以后，紧跟着的问题就是“人是如何感知到这种关系的，进而机器如何能自动分析出这种关系”。拿前面的例子来说，如何判定“自信”的主语是“刘小姐”，“插在泥堆上”的状语是“把一长片木柴”等。一个典型的例子是：

他娶了一个漂亮姑娘，很骄傲，虚荣心很强。

这个句子应当按照下面哪一种图式去理解：

他娶了一个漂亮姑娘，　　**他娶了一个漂亮姑娘**，
很骄傲，　　很骄傲，
虚荣心很强。　　虚荣心很强。

他娶了一个漂亮姑娘，　　他娶了**一个漂亮姑娘**，
很骄傲，　　很骄傲，
虚荣心很强。　　虚荣心很强。

谁很骄傲，是“他”还是“一个漂亮姑娘”？虚荣心很强的又是谁？进一步，“骄傲”可以表示“高傲”，也可以表示“自豪”，这里是哪一种意思？所有这些判断的依据又是什么？

还有关于标点句独立性的问题，就是如何判断一个标点句首部是否共享另一标点句中的成分。可以想象，一个标点句如果以谓词性成分起首，它往往需要到另一个标点句中去找谓词的主体；如果以属性名、部件名起首，则往往需要到另一个标点句中去找表示属性主体、部件整体的成分。反过来，如果一个标点句以专名或人称代词起首，它在句法上往往是自足的。这样说在多数情况下并不错，但还是有反例：

她哥哥也到香港做事，
上海家里只剩她母亲、嫂子和她。

后一个标点句起首的是专用地名“上海”，但仍要共享前一句的主语“她”。

这些都是从分析的角度看。如果从生成的角度看，也有类似的问题。比如，在许多情况下，前句的宾语如果用作后句的主语，后句就可以不出现这个主语：

出房碰见**孙小姐**，
脸上有些红点，
扑鼻的花露水香味，
也说痒了一夜。

但是，下一例后句中指代“孙小姐”的“她”就必须出现：

这次买船票咱们已经带累了孙小姐，

她是脸皮嫩得很的女孩子，

还有，多数情况下，连续两个标点句如果是同一个主语，后句往往共享前句的主语，但并不尽然。下例中连续两个标点句首部都是“我”，后句的“我”必须出现：

我不稀罕你家的钱，
我会写信给我爸爸。

所有这些，都涉及标点句之间发生句法关系有什么约束条件。张瑞朋工作的主要部分就在这里，包括如何判断标点句首是否缺成分，如果缺成分，这个成分能否在上句中找到，如果能在上句中找到，它是上句中的哪个成分。张瑞朋分门别类地进行了研究，还特别研究了否定成分的辖域，把字句、被字句、兼语句等几种特殊句型与后面标点句的句法关系，以及并列型成分被后面标点句共享的情形。为了寻找约束条件，她标注了《围城》全篇20多万字2万多标点句的跨标点句的句法关系，粗标了标点句的浅层句法结构，还在数千万字的小说语料中作了专项调查。她观察这些现象，进行统计分析，探究现象的本质原因，提出了一些很有价值的新颖的语言特征，如句子的动态性和静态性的区别，动词对施事和对受事的不同影响，宾语的信息量，词语的多种属性等，进而梳理出了一批跨标点句句法关系的约束条件和规则。这一工作既有庞大的工作量，又有相当的深度。

我推荐张瑞朋的工作，支持这本书的出版，并不是因为完全认可她归纳出的各种规则。书中所归纳的规则并不一定都正确。所有的判断都是人工进行的，难免发生错判。而且，正如书中所言，各规则的成熟度是不同的。我更看重书中经作者筛选后给出的大量实例，以及为了将实例归纳分类而找出的种种特征。语法学应当属于自然科学，自然科学的目标是发现，发现的基础是大量客观现象的调查分析和归纳。类似于生物学家或地质学家历经千辛万苦采集来的标本，这些材料对于有兴趣于这一领域的学者，是不可多得的宝贵样本。

1990年我在伊利诺伊大学郑锦全先生处做访问学者，郑先生建议我做汉语篇章结构分析。我就整天泡在亚洲图书馆读中国小说，那时第一次注意到汉语的这种语言现象。但是不知道应该归入语言学的什么范

畴，于是就称作辖属关系，即被共享的成分管辖两个或多个共享者。后来精力放在计算语言学应用技术研究方面，这个课题被放下了。10 年后，我到香港城市大学邹嘉彦先生处访问，有时间继续推进这方面的工作。那时发现这种管辖关系可以从句法角度认识，于是有了跨标点句的句法关系分析的目标。2004 年，张瑞朋来北京语言大学跟我读博士，她有很好的语言学训练，我就把这个题目交给她，希望她能进行深入的语言现象调查，在此基础上理出一些头绪。本书是她博士论文的成果。此后，这一课题两度得到国家自然科学基金的资助，于是又有新的学生参加进来。目前，我们正在从话题结构的角度研究相关现象，将跨越标点句的句法关系分析作为其中的一个核心子课题。我们的愿景是第一步形式化，第二步算法化，第三步应用于篇章分析。当然，要实现这个愿景，还得走很长很长的路。

科学研究的魅力不在于结果，而在于过程。观察现象，探求现象的本质特征，归纳特征间的因果关系，这是一种不断地自我挑战而又战胜挑战的过程，其乐无穷。我建议读者认真地品味书中的大量实例，琢磨作者提出的特征，检验特征间的关联，提出自己的见解。我相信，这个过程是引人入胜的。

宋　柔

北京语言大学

2013 年 7 月 1 日

摘　要

目前，汉语的句法分析研究基本上以单句为对象，但在真实语料中，汉语单句边界的自动确定是很困难的。在句子层面上，主要的形式标记是标点。计算机处理汉语的前提是汉语的形式化，因此标点句自然而然就成了计算机处理汉语句子的基本单位。标点句的边界是清楚的，但很多标点句的句法成分不完整，需要到上下文语境中去寻找。但跨标点句的句法分析问题尚无系统性方法，这就使得汉语长句分析和长句生成效果很差，并已经成为汉外机器翻译和汉语理解等深层次汉语处理应用系统的瓶颈。为了解决这个问题，首先要对汉语跨标点句的句法关系作一番仔细的调查分析，总结出一些规律和约束条件。

本项工作是在跨标点句句法关系的理论框架下展开的，主要目的是解决跨标点句共享成分的识别问题，找出这类句法关系在满足栈形规律之外，还应满足哪些可以形式化的约束规则，以便计算机处理。

本书的工作包括两方面：

（1）语料库标注、调查和统计。

标注了钱钟书《围城》全文，共计 226641 字，24115 个标点句。标注内容包括跨标点句间的句法关系类型、共享成分、标点句内部的浅层句法结构，从中得到了标记语料中各种跨标点句句法关系的统计数据。笔者还借助文本检索工具对数千万字的中国现代小说、当代小说进行了多项专门调查和统计。

（2）约束条件挖掘。

在标注语料和专项调查的基础上，分列大小 100 多个方面总结出跨标点句句法关系发生的各种约束条件。重点研究原配句和续配句同源并且是正序关系的情况。涉及的内容包括：

- 名词或代词开始的标点句主语是否缺失。
- 主动宾结构的标点句，续配句主语是原配句主语还是宾语。其中讨论了原配句为感知动词句、“有”字句、句宾动词句、连动结构、“像”字句、“V着”句、“V完”句的情况以及一些关联词、副词、形容词、名词对于共享成分的影响。
- 续配句共享原配句状语的认定，涉及多种形式的状语，专章讨论了否定词的跨标点句管辖的判断。
- 续配句共享原配句定语的认定，涉及量词、形容词、代词、名词和名词短语的情况。
- 原配句是把字句、被字句时，句内成分被共享的情况。
- “跟”与“和”连接的名词短语被续配句整体共享或部分共享的区分。
- 原配句是兼语句时，句内成分被共享的情况。

本书的工作在如下方面是有特色的：

（1）研究范围方面，除了前人已有的研究跨越标点句的主谓关系之外，还研究了跨越标点句的定中关系、状中关系、述宾关系、述补关系、介宾关系等，全面铺开了跨标点句句法体系的研究。

（2）研究角度方面，侧重于约束条件中的形式化特征，研究成果具有较强的可操作性，为计算机自动进行跨标点句句法关系的分析打下了一定的基础。

（3）研究方法方面，不满足于举例说明。除了使用传统的自省方法，寻找语言规律的认知理据外，重视真实语料的语言现象统计，以统计数据作为规律可靠性的佐证。

本书的创新性主要表现在语言特征的多角度的深入挖掘方面。择要列举如下：

原配句是主动宾结构的情况下，关于缺主语的续配句共享原配句主语还是宾语，本书指出了几种重要的区别特征：

- 指出区别主语话题与宾语话题的主要标志之一是静态句、动态句，从形式上界定了这两种标点句，指出了这两种标点句同主语话题和宾语话题的关系。
- 根据动词对施事、受事的影响，把动词划分为只对施事产生影

响的动词和对施事、受事都产生影响的动词，用以区别主语是否转换。

- 提出信息量的概念，指出原配句是“有”字句以及续配句是中间态形容词谓语句时，续配句的主语确定同原配句宾语的信息量有关，宾语信息量越小，宾语作为续配句主语的可能性越大。

把标点句分为独立标点句和不独立标点句，用于解决标点句之间是否发生共享关系。

把名词从总体上分为独立名词和不独立名词，用于判断标点句的完整与否。

对于一些主—副型的连动谓语句，本书采用句型变换的方法归结为主动宾型的单谓语句，再决定续配句的主语认定问题。

把动词和形容词作谓语的情况总体划分为方向性谓语和非方向性谓语，用于解决并列名词短语被整体共享还是部分共享的问题。

把副词、时间词等状语总体划分为句子状语和词语状语，用于解决状语成分是否被共享的问题。

对于各种词性的词语从语义角度进行了细致的分类，用于解决跨标点句共享成分的确定问题。这些词类多数曾散见于多种语言学文献中，但界定方法和使用目标不同，有些是本书首次提出的。本书将这些词类综合使用，有些进行了重新界定，并在高频词范围内给出了这些词类的词表。其中包括：

- 动词词类：存现动词、准存现动词、感官动词、关系动词、认知动词、心理动词、行为动词、使令动词、身体行为动词；
- 名词词类：器官名词（部件名词）、属性名词、亲属名词、心理名词；
- 形容词词类：动态形容词、静态形容词、中间态形容词；
- 副词词类：短暂动作副词、心理副词、情态副词、时间副词、关联副词、评注性副词、范围副词、程度副词等；
- 提出了心理词的概念，包括心理名词、心理动词、心理形容词、心理副词。

其中本书首次提出的词类有：中间态形容词、短暂动作副词、心理

词、心理副词、心理形容词。

语言学文献中出现过，但界定方法和范畴不同的有：准存现动词、动态形容词、静态形容词、情态副词。

使用平行结构的方法判断成分共享。

在跨标点句句法关系领域，本书的工作是相当初步的。由于时间的关系，许多问题还未涉及，许多问题只是开了一个头。研究成果还比较零乱，系统性不够，更未涉及算法化、程序化的工作。这些工作将在今后逐步展开。

关键词：标点句；共享成分；句法关系；约束条件

Abstract

Currently, the Chinese syntactic analysis is basically targeted at single sentence. However, the border of Chinese single sentence is very difficult to assure automatically in real corpus. The main form tag is punctuation sentence levels. The prerequisite of Chinese language processing is to formalize. So punctuation sentence become the basic units that computer processes Chinese sentence automatically. The border of Punctuation sentence is clear, but the syntactic elements of many of punctuation sentences is incomplete, and we need to find them in context. But the problem of syntax analysis of inter-punctuation sentence is not systemic . This makes the parsing of Chinese Long Sentences and the generating of long sentences a poor result, and has become the most difficulty of foreign and Chinese machine translation and the deep-rooted understanding of Chinese Processing. To solve this problem, first, we must investigate the syntactic relations of Chinese-punctuation-sentences carefully and summed up some rules and constraints.

This work is based on the theory framework of the punctuation sentence. The main purpose is to identify the common element in punctuation sentence, and in order to computer process punctuation sentences expediently, we need find the formal binding rules besides the stack-type rules in the syntax relation.

This work consists of two aspects:

(1) mark the corps and make a survey and statistics

We Marked the total of Qian Zhongshu's "WeiCheng", 226641 words and 24115 punctuation sentence. The tags include the syntactic relations between punctuation sentence, the common ingredients, the shallow syntactic

structure within the punctuation sentence and we gain the statistical data about each kind of punctuation sentence in marked Corpus. I also use text retrieval tools to do some specialized investigations and statistics on modern and contemporary Chinese novel of tens of millions of characters.

(2) Mining the constraints

On the basis of marked corpus and special investigations, we summed up various of constraints of punctuation sentence from about a hundred of big or small aspects . We focus on the punctuation sentences that yuanpei sentence and xupei sentence is homologous and ordinal . The contents include :

- whether the punctuation sentences whose beginning element is noun or pronoun miss subject.
- If structure of yuanpei-sentence is subject - verb-object, the subject of xupei-sentence is subject or object in yuanpei-sentence . We discusses these punctuation sentence whose yuanpei-sentence' predicate is sense verb, "有", sentence-object verb, two-verb structure, "像", "V着", "V完" as well as the affect to common elements of relevance words, adverb, adjective and noun.
- How to identify the adverbial modifier of xupei-sentence, involving various forms of adverbial. We discuss the domain of negative word in punctuation sentence in a special chapter.
- How to identify the attribute of xupei-sentence, involving quantifiers, adjectives, pronouns, nouns and noun phrase.
- If yuanpei-sentence is 把 sentence and 被 sentence, how to identify the common components in sentence.
- How to identify the overall or part of the noun phrase connected with "跟" in yuanpei-sentence is shared by xupei-sentence.
- If Yuanpei-sentence is jianyu-sentence, how to identify the common components in sentence.

This work is characteristics in the following aspects:

(1) About the scope of the study, in addition to previous studies about the subject-predicate punctuation sentence, We also studied the attribute-head