# 汉语句法树库标注体系*

周　强

清华大学计算机系

智能技术与系统国家重点实验室

北京 100084

zhouq@s1000e.cs.tsinghua.edu.cn

摘要：语料库的句法标注是语料库语言学研究的前沿课题。本文在研究和总结国内外句法树库标注实践的基础上，提出了一套汉语真实文本的句法树标注体系。它以完整的层次结构树为基础，对句法树上的每个非终结符节点都给出两个标记：成分标记和关系标记，形成双标记集的句法信息描述体系。目前，这两个标记集分别包含了 16 和 27 个标记，对汉语句子的不同句法组合的外部功能分布和内部组合特点进行了详细描述。在此基础上，我们开发完成了 100 万词规模的汉语句法树库 TCT，对其中各种复杂语言现象的标注实践显示了这套标注体系具有很好的信息覆盖率和语料适应性。

关键词：句法树库，标注规范，语料库语言学

# Annotation Scheme for Chinese Treebank

ZHOU Qiang

State Key Laboratory of Intelligent Technology and Systems

Dept. of Computer Science and Technology

Tsinghua University, Beijing 100084

zhouq@s1000e.cs.tsinghua.edu.cn

ABSTRACT: The syntactically annotated corpora, commonly called ‘treebanks’, play an important role in empirical linguistics as well as in machine learning methods in natural language processing. After a brief summarization of several treebank annotation of different language, we proposed a new annotation scheme for Chinese treebank in this paper. Under this scheme, every Chinese sentence will be annotated with a complete parse tree, where each non-terminal constituent is assigned with two tags. One is the syntactic constituent tag, which describes its external functional relation with other constituents in the parse tree. The other is the grammatical relation tag, which describes the internal structural relation of its sub-components. These two tag sets consist of 16 and 27 tags respectively. They form an integrated annotation for the syntactic constituent in a parse tree through top-down and

*本项研究得到国家自然科学基金（项目号：69903007 和 60173008）、国家 973 基金（项目号：G1998030507，G1998030501A-03）、国家高技术研究发展 863 计划（项目号：2001AA114040）资助。

作者：周强，男，1967 年生，博士，副研究员，主要研究方向：计算语言学、语料库语言学、词汇语义学

bottom-up descriptions. Based on this scheme, we built a 1,000,000 words Chinese treebank covering a balanced collection of journalistic, literary, academic, and other documents. The annotating experiments on different kinds of complex linguistic phenomena show the availability and compatibility of this annotation scheme.
KEYWORDS: Tree Bank, Annotation Scheme, Corpus Linguistics

# 1 引言

语料库的句法标注是语料库语言学研究的前沿课题，它的处理目标是对语料文本进行句法分析和标注，形成树库（tree bank）语料。近年来，国内外研究人员在这些方面进行了深入探索，开发完成了许多大规模的树库。在英语方面，有英国的 Lancaster- Leeds 树库[LG91]和美国的 Penn 树库[MSM93]；德语方面，有 NEGRA 树库[SBK98]和 TIGER 树库[BH02]；捷克语方面，有布拉格依存树库（PDT）[Hai99]；汉语方面，有美国宾州大学的 Penn 中文树库[XM00]和台湾中研院的 Sinica 中文树库 [HCC00]。

在这些树库项目的开发过程中，一个特别值得重视的发展趋势是树库构建与语法理论研究的紧密结合。欧洲目前进行的一些树库项目都有很深的语法理论研究背景，如：捷克的 PDT 项目以依存语法为基础；德国的 TIGER 项目以词汇功能语法（LFG）为基础；英国的 LigGO 项目[OFT02]以头驱动短语结构语法（HPSG）为基础等。两者紧密结合的好处是显而易见的。一方面，利用语法理论的最新研究成果，可以很快建立起比较完整的树库标注体系；同时，利用比较成熟的基于不同语法理论的句法分析器作为预处理工具，可以大大降低大规模树库的构建成本。另一方面，通过大规模真实文本的树库构建实验，可以发现许多新的语言现象，为语法理论提供丰富的研究素材，使理论体系得到不断改进和完善。两者相辅相成，互相促进，达到了理论研究和实际应用的完美结合。

从 1998 年起，我们开始进行汉语句法树库的开发研究，希望构建完成目前世界上规模最大、信息标注最丰富的汉语句法树库。经过 5 年多的努力，逐步总结形成了一套比较完整的汉语真实文本的句法树标注体系和处理规范。在下面的几节中，我们首先对目前国内外典型树库的句法标注体系进行简单综述（第 2 节）。接着介绍我们的标注体系的主要内容（第 3 节）。然后简要介绍以此为基础进行的汉语句法树库标注实践和目前完成的 100 万词规模的句法树库 TCT 的基本情况，并对有关内容进行总结和展望（第 4 节）。在最后的结语（第 5 节）中，对有关工作进行总结和展望。

# 2 国内外典型树库的标注体系

在英语方面，美国的 Penn 树库的标注体系经历了一个从简单到复杂的不断进化发展过程。最初的 PTB-1[MSM93]采用了骨架分析（Skelton Parsing）思想，形成比较扁平的句法结构树。随后，在扩充版本（PTB-2）[MKM94]中，增加了一些功能标记，用于标注句子中主要句法成分的语法功能，希望能据此自动抽取出句子的谓词-论元（Predicate-Argument）信息。从 2002 年起，他们进一步提出了命题库（PropBank）构建计划[KMM02]，在 PTB-2 上明确标注句子中各个动词的谓词-论元信息，希望借此建立从句法到语义的重要桥梁。

捷克的 PDT 项目[Hai99]则设计了三个层次的标注信息：词法、句法和语义。在词法层

面上，充分利用了捷克语丰富的形态变化信息，总结了 4200 多个词类标记；在此基础上形成的句法依存树，对句子中关键词语的句法依存关系进行了描述；然后，利用动词的详细句法语义描述词典，将表层的句法依存关系转化为深层的语义依存关系。

从描述框架上看，PTB 采用的句法结构树和 PDT 采用的依存树各有优势。句法结构树可以对不同层次的句法成分组合特点进行细致的描述，但缺点是有时层次比较深，操作起来比较麻烦，而且中心词（Head）信息不突出。为此，PTB 项目进行了一些改进，包括采用骨架分析方法减少层次深度，增加功能标记突出中心依存关系等。但从 Collins(1999)在 PTB-2 上进行的中心词依存关系对自动抽取实验结果看，大量人工总结的匹配规则还是必需的。而依存树的优势则在于明确地标注出了中心词之间的句法依存关系，可以方便地转化为语义依存描述，但它对一些没有明确依存关系的成分，标注起来则有些力不从心。因此，较好的处理方法是将两者有机结合起来。在这方面，德语的 TIGER 项目进行了有益的尝试。

在 TIGER 树库中，研究人员采用了一种层次结构和依存关系相结合的标注体系：底层的句法成分主要采用层次结构，可以保留大量丰富的描述信息；高层的语法关系则采用依存结构，描述句子中各主要成分与中心动词之间的各种句法依存关系，形成一种功能强大、处理灵活的描述体系，特别适合于象德语那样语序比较自由的语言。

在汉语方面，目前两个较大的树库是美国宾州大学的汉语树库（CTB）和台湾中研院的 Sinica 汉语树库。在标注体系上，CTB 基本上沿用了英语 PTB-2 的标注体系。目前的总标注规模为 50 万词的新闻语料。另外，他们也在进行汉语的命题库项目[XP03]，计划在现有的汉语树库上标注完整的 PA 关系信息。他们的基本设想是在一个共同的标注框架下，实现英语和汉语的双语信息标注，为进一步进行英汉双向机器翻译和信息抽取研究打下基础。这个研究路线有其合理性和可行性，但把许多汉语独具特色的描述信息硬纳入英语的描述框架，总给以汉语为母语的人许多生硬别扭的感觉。

台湾中研院的树库标注体系则是在他们提出的信息为本的格语法上构建起来的。其标注格式非常类似于 TIGER 的结合描述框架，差别是用 Theta 角色代替了依存关系描述。他们的主要处理特点是按照标点符号对汉语句子分块，对每个小句（块）进行句法分析和标注，形成不同句法树。目前共标有 41100 棵树，约 241008 个词。这种处理方法降低了标注难度和工作量，但也不可避免地丢失了汉语复杂长句中丰富的描述信息。

## 3 我们的句法树标注体系

从 1998 年起，我们开始进行汉语句法树库的开发研究，希望构建完成目前世界上规模最大、信息标注最丰富的汉语句法树库。为此，我们选择了大规模的包含文学、学术、新闻、应用四大体裁的平衡语料文本作为加工对象，以期尽可能多地覆盖汉语的各种语言现象；我们确定了比较自然的书面语文本的断句方法[1]，以期尽可能忠实地反映汉语句子组织信息的本来面目；我们采用了完整的层次结构树描述框架，设计了双标记集的描述体系，对句法树上的每个非终结符节点都给出两个标记：成分标记和关系标记，分别描述其外部功能分布和内部组合特点，以期尽可能详细地描述汉语句子的句法组合信息。

我们采用完整的层次结构树描述，而不是目前国际上比较通用的骨架分析或依存关系描

[1] 一般情况下，以句号、问号、感叹号等显性标记作为断句依据。

述，主要基于以下几点考虑：

1）层次结构树可以给出汉语句子最为详细的句法信息描述，覆盖汉语“字/词→块→句→段”各个层次的句法单元。而在具体标注过程中，利用我们提出的分阶段树库构建方法[ZRS02]，可以大大减低人工校对的工作量，从而弥补这种描述体系在具体实现上的弱点。

2）这套描述体系可以与现有研究成果达到完美结合。首先，近年来，国内语言学界在汉语层次分析方面进行了深入研究，积累了许多有价值的研究成果，可以充分吸收到我们的标注体系中；其次，我们在汉语自动层次分析方面进行了大量探索，开发了比较完整的自动句法分析器[ZQ97]和句法知识自动获取工具[ZQ01]，可以为树库构建项目提供有力的支持。

3）在这套体系下多年的研究与教学，已形成了丰富的人才储备库，从中我们可以方便地找到大量高质量的树库校对人员，不需要经过大量培训就可以胜任目前的校对任务。这对降低大规模语言工程的开发费用是至关重要的。

4）基于目前树库的丰富信息容量，我们可以方便地开发自动转换程序，按照不同的应用需求，把现有树库标注格式转换成骨架分析树、依存关系树或两者结合形式。同时，也可以方便地从目前的树库中自动提取基本短语和功能语块标注信息，建立现有的句法树标注体系与汉语部分分析体系[ZQ03]的内在联系，扩大目前树库语料的应用范围。

作为语料库多级加工过程中的一个中间阶段，句法分析和标注应该为进一步进行汉语句子的词语义项和语义关系标注提供有力的支持。理想情况下，在对句子进行正确句法信息标注的前提下，应能依据一个语义知识库和自动标注工具，准确地标注出大部分的语义信息。而要实现这个目标，就必须在这个阶段给出尽可能详细的句法信息描述。我们的基本设想是，对结构信息的完整描述，至少应包含以下内容：

1） 结构的外部功能特征描述：分析它进一步与其他结构相组合的能力；

2） 结构的内部组合关系描述：分析它内部的组成成分之间的语义组合关系；

3） 结构的语义中心词描述：分析它的语义中心词位置；

在汉语的绝大多数结构中，一般可以依据上面 1）和 2）的信息唯一地确定 3）的位置。因此，在我们的标注体系中，主要对 1）和 2）两部分信息进行显性描述和标注。为此，我们设计了以下两个标记集：成分标记集和关系标记集。下面分别进行简要说明，有关的详细内容可参阅[Th02]。

## 3.1 成分标记集设计

我们目前设计了 16 个成分标记（见表 1）。它们沿用了我们最初提出的树库标记集内容[ZZY97]，基本上覆盖了汉语“字/词→块(短语)→句→段”各个层次的句法单元，具有较强的适应性，可以方便地加工处理大规模的真实文本语料。

首先，我们设计的 10 个短语和准短语标记，通过与下面的句法关系标记相配合，可以对汉语“字/词→块（短语）”之间的一些连续变化单元，包括语法词、复合词、短语等给出详细描述。

而标记组 {dj,fj,zj,jq} 则较好地体现了汉语短语到句子的实现关系和句子之间的组合关系：dj 和 fj 作为特殊的短语--句型标记，一方面可以灵活地充当句子成分，体现了汉语独特的成分套叠现象[LJM93]，另一方面又可以通过在句尾加上语调标点（句号、叹号、问号）实现为一个完整的句子（zj）。相反，zj 则一般不充当句子中的句法成分，这反映了 zj 与

dj 和 fj 在语法层次和具体使用上的差别，但多个 zj 仍可以进一步组合为更大的语法单位——句群(jq)[2]。

另外，我们还设计了两个标记（yj 和 dlc），对汉语句子中的直接引语和独立成分进行了描述，并对一些常见的独立成分进行了分类标注，包括插入语、称呼语、补充说明、复指成分、强调成分等，有关详细内容可参阅[Th02]。

表 1 汉语成分标记集

| 序号 | 标记代码 | 标记名称 | 序号 | 标记代码 | 标记名称 |
|---|---|---|---|---|---|
| 1 | np | 名词短语 | 9 | mbar | 数词准短语 |
| 2 | tp | 时间短语 | 10 | mp | 数量短语 |
| 3 | sp | 空间短语 | 11 | dj | 单句句型 |
| 4 | vp | 动词短语 | 12 | fj | 复句句型 |
| 5 | ap | 形容词短语 | 13 | zj | 整句 |
| 6 | bp | 区别词短语 | 14 | jq | 句群 |
| 7 | dp | 副词短语 | 15 | dlc | 独立成分 |
| 8 | pp | 介词短语 | 16 | yj | 直接引语 |

## 3.2 关系标记集设计

我们目前设计了 27 个关系标记（见表 2），希望能尽可能全面地覆盖汉语的各种句法语义关系描述。其中，在“词→块→小句”层面上，主要描述了小句中核心谓词（主要是动词和动词短语）与周边成分的支配关系，包括：主谓、述宾、述补等；和各个描述成分与中心词之间的修饰关系，包括：定中、状中等；以及各种实体概念与功能词之间的句法组合关系，包括：附加、方位、介宾、框式、标号等。它们形成了汉语基本事件和实体内容描述的基本框架，从中可以直接推导出汉语句子的常见语法范畴，包括主语、谓语、宾语、定语、状语、补语、附加语、中心语等。从而为进一步进行目前的句法结构树向依存树的转换打下很好的基础。在“句→段”层面上，则侧重描述了汉语复杂事件的各种逻辑关系组合，包括各个事件之间的顺序连接关系，如：连谓、兼语、连贯、递进、流水、解注等；以及各个事件之间的条件蕴涵关系，如：条件、假设、转折、因果、目的等。为进一步进行汉语不同事件描述关系的分析提供了研究基础。而并列联和关系描述则覆盖“词→块→小句→句→段”等各个层面，形成成分重叠、短语联合及小句并列或选择关系。有关的详细内容可参阅[Th02]。

以上不同关系标记设计基本上沿用了汉语语法研究的相关术语，其中比较特别的是我们提出的顺序(SX)关系标记。它最初是为了描述真实文本中大量出现的“起点→历程→终点”的时空变化顺序而提出来的，主要包括以下情况[Th02]：

- 通过多个介词结构描述，如：[pp-SX [pp-JB 从北京] [pp-JB 经天津 ] [pp-JB 到上海 ] ]
- 通过“X p(至|到) Y”描述“起点→终点”的时空顺序，如：[tp-DZ 12 月 [tp-SX 5

[2] 在我们目前的标注实例中，这种情况主要出现在复杂引语中。

日 [pp-JB 至 7 日 ] ] ]

- 通过“ X –|— Y ”描述“ 起点→终点 ”的时空顺序，如：[sp-SX 北京 – 上海 ]

但随着研究工作的不断深入，我们发现这种顺序关系在汉语中大量出现[Dai02]，包括动作之间的时序关系、事件之间的连贯关系等，是否在更高层次上对这些关系进行进一步抽象，使用统一的关系标记进行描述，还需要进行进一步的探索。

表 2 句法关系标记集

| 序号 | 标记代码 | 标记名称 | 序号 | 标记代码 | 标记名称 |
|---|---|---|---|---|---|
| 1 | ZW | 主谓结构 | 15 | SX | 顺序结构 |
| 2 | PO | 述宾结构 | 16 | BL | 并列关系 |
| 3 | SB | 述补结构 | 17 | LG | 连贯关系 |
| 4 | DZ | 定中结构 | 18 | DJ | 递进关系 |
| 5 | ZZ | 状中结构 | 19 | XZ | 选择关系 |
| 6 | LH | 联合结构 | 20 | YG | 因果关系 |
| 7 | LW | 连谓结构 | 21 | MD | 目的关系 |
| 8 | AD | 附加结构 | 22 | JS | 假设关系 |
| 9 | CD | 重叠结构 | 23 | TJ | 条件关系 |
| 10 | JY | 兼语结构 | 24 | ZE | 转折关系 |
| 11 | JB | 介宾结构 | 25 | JZ | 解注关系 |
| 12 | FW | 方位结构 | 26 | LS | 流水关系 |
| 13 | KS | 框式结构 | 27 | XX | 缺省关系 |
| 14 | BH | 标号结构 | | | |

## 4 句法树标注实践

基于上面介绍的句法树标注体系，我们总结制定了一部比较完整的汉语句法树标注规范[Th02]，对大规模汉语真实文本进行了句法树标注实践。其加工对象选自清华大学和北京语言文化大学联合开发的 200 万汉字的平衡语料库。它的主要语料来源是 90 年代的现代汉语书面语以及准口语（包括剧本、谈话录、演讲录等）的真实文本，按文体分为文学、新闻、学术、应用四类。经过自动切词、词性标注和人工校对，已经形成了准确度很高的切分和词性标注精加工文本，为进一步进行句法树库构建打下了很好的基础。

大规模的树库构建是一项庞大的语言工程。在目前的条件下，完全由机器自动完成是不可能的，一定的人工投入是必需的。关键问题是如何寻找一个合适的人工介入点，以最少的人工投入，获得最佳的整体处理效果。为此，我们提出了分阶段的树库构建设想[ZRS02]：

- 第一阶段：在经过正确切分和词性标注处理的汉语语料文本上，人工标注正确的功能语块信息，形成语块库[ZRZ01]。
- 第二阶段：在汉语句子的语块标注结果上，利用自动句法分析器，分析并标注句子的句法结构树，并进行人工校对，形成完整正确的树库语料。

这种“ 逐步求精 ”的树库构建设想，可以大大提高整体的工作效率，以最小的人力物力投入，取得最佳的树库构建效果。

下面是对一个具体汉语句子实例的分阶段标注结果：

- 输入句子：我/rN 哥哥/n 送/v 给/v 我/rN 一/m 本/qN 很/d 漂亮/a 的/u 书/n 。/w[3]
- 功能语块标注结果：[S 我/rN 哥哥/n ] [P 送/v 给/v ] [O 我/rN ] [O 一/m 本/qN 很/d 漂亮/a 的/u 书/n ] 。/w
- 句法树分析和校对结果：[zj-XX [dj-ZW [np-DZ 我/rN 哥哥/n ] [vp-PO [vp-PO [vp-SB 送/v 给/v ] 我/rN ] [np-DZ [mp-DZ 一/m 本/qN ] [np-DZ [ap-ZZ 很/d 漂亮/a ] 的/u 书/n ] ] ] ] 。/w ]

具体的加工过程则是标注规范、校对人员和自动分析工具之间的互动调整过程。经过 5 年多的努力，我们加工完成了 100 万词的汉语句法树库 TCT（Tsinghua Chinese Treebank, v1.0）。其中不同文体语料所占比例（按词项数计算）分别为：文学 47.3%、学术 26.3%、新闻 20.0%和应用 6.4%。另外，对 4 万多个整句内部组成结构进行分析，发现由复句形成的占 56.8%、由单句形成的占 32.6%、由动词短语形成的占 5.7%。这种分布格局在 4 个文体的语料中基本相同，表明在真实文本的汉语句子描述中，复杂句子占了绝大多数。这种现象对目前以单句为中心的汉语句法理论研究和自动分析方法探索提出了新的问题和挑战。

## 5 结语

综上所述，我们在大规模汉语树库构建方面进行了以下探索性研究：

1）选择了大规模的包含文学、学术、新闻、应用四大体裁的平衡语料文本作为加工对象，这在国内外的大型树库项目中还没有看到。相比而言，PTB, PDT 和 TIGER 和 CTB 主要采用了新闻语料，台湾的 Sinica 树库语料虽然取自他们的 500 万字的平衡语料库，但规模较小。

2）采用了完整的层次结构树描述框架，设计了双标记集的描述体系，对句法树上的每个非终结符节点都给出了尽可能丰富的汉语句法描述信息。

目前，我们已经开发完成了目前世界上规模最大、信息标注最丰富的汉语句法树库 TCT（Tsinghua Chinese Treebank, v1.0），并且开始在 TCT 上进一步进行更深层次的句法分析和词汇语义标注研究。

## 参 考 文 献


[BH02] Brants, S., & Hansen, S. (2002). Developments in the TIGER annotation scheme and their realization in the corpus[A]. In *Proceedings of the Third Conference on Language Resources and Evaluation (LREC-02)*[C]. *Las Palmas de Gran Canaria, Spain*. p.1643-1649.

[Col99] Collins, M. (1999) Head-Driven Statistical Models for Natural Language Parsing[D]. Ph.D. Thesis. Dept. of Computer Science and Information, The University of Pennsylvania.

[Dai02] 戴浩一（2002）概念结构与非自主性语法：汉语语法概念系统初探[J]，《当代语言学》，4(1), 1-12.

[Hai99] Hajic, J. (1999). Building a syntactically annotated corpus: The Prague Dependency Treebank[A]. In *E. Hajicova (Ed.), Issues of valency and meaning. Studies in honour of Jarmila Panevova. Prague, Czech Republic:*


[3] 有关的标记符号简要说明如下：rN—名代词，n—名词，v—动词，m—数词，qN—名量词，d—副词， a—形容词， u—助词， w—标点符号；S—主语块，P—述语块，O—宾语块。

*Charles University Press*

[HCC00] Chu-Ren Huang, Feng-Yi Chen, Keh-Jiann Chen, & al.(2000). Sinica Treebank: Design Criteria, Annotation Guidelines, and On-line Interface[A], *Proceedings of the Second Chinese Language Processing Workshop*[C], *HongKong*. 29-37.

[KMM02] Kingsbury, P.; Martha Palmer, and Marcus, M. (2002). Adding Semantic Annotation to the Penn TreeBank[A]. In *Proceedings of the Human Language Technology Conference*[C], *San Diego, California.*

[LG91] Leech, G.; and Garside, R. (1991). Running a grammar factory: The production of syntactically analysed corpora or 'treebanks' [A]. In *Stig Johansson and Anna-Brita Stenstrom (eds.) English Computer Corpora : Selected papers and Research Guide*. 1991. 15-32

[LJM93] 陆俭明 (1993). 汉语句法成分特有的套叠现象[A]，《陆俭明自选集》，河南教育出版社，174-192.

[MKM94] Marcus, M., Kim, G., Marcinkiewicz, M.,& al. (1994). The Penn Treebank: Annotating predicate argument structure [A]. In *Proc. of the ARPA Human Language Technology Workshop*[C]. San Francisco, CA.

[MSM93] Mitchell P.Marcus, Mary Ann Marcinkiewicz, and Beatrice Santorini (1993). Building a Large Annotated Corpus of English: The Penn Treebank[J], *Computational Linguistics*, 19(2), 313-330.

[OFT02] Stephan Oepen, Dan Flickinger, Kristina Toutanova, et. al. (2002). LinGO Redwoods --- A Rich and Dynamic Treebank for HPSG [A], In *Proc. of First Workshop on Treebanks and Linguistic Theories (TLT2002)* [C] , 139-149.

[SBK98] Skut,W., Brants, T., Krenn, B., & Uszkoreit, H. (1998). A linguistically interpreted corpus of German newspaper text [A]. In *Proceedings of the Conference on Language Resources and Evaluation LREC-98*[C]. Granade, Spain. pp. 705–711

[Th02] 汉语句子的句法树标注规范 V2.0 [R]，清华大学计算机系智能技术与系统国家重点实验室，技术资料，2002 年 6 月。

[XCM02] Xue N.W., Chiou F. and Martha P. (2002). Building a Large-Scale Annotated Chinese Corpus [A]. In *Proc. of 19th International Conference on Computational Linguistics (COLING-02*) [C], Taiwan.

[XM00] Xia, Fei, Martha Palmer, & al. (2000) Developing Guidelines and Ensuring Consistency for Chinese Text Annotation [A]. In *Proceedings of the second International Conference on Language Resources and Evaluation (LREC-2000*[C]*), Athens, Greece.*

[XP03] Nianwen Xue and Martha Palmer. 2003. Annotating Propositions in the Penn Chinese Treebank [A], In *Proceedings of the 2nd SIGHAN Workshop on Chinese Language Processing*, in conjunction with ACL'03 [C]. Sapporo, Japan

[ZQ01] 周强（2001） 汉语句法知识的自动获取研究 [A]，中文信息学会二十周年学术会议论文集[C]，北京，156-165

[ZQ97] Qiang Zhou (1997) A Statistics-Based Chinese Parser[A], In *Proc. of the Fifth Workshop on Very Large Corpora*[C], *p.*4-15.

[ZRS02] 周强,任海波, 孙茂松(2002) 分阶段构建汉语树库[A], In Proc. of The Second China-Japan Natural Language Processing Joint Research Promotion Conference[C], Beijing, China. p189-197.

[ZQ03] 周强 (2003) 汉语部分分析研究[A], 孙茂松、陈群秀主编《语言计算与基于内容的文本处理》[C]，清华大学出版社，p116-121.

[ZZY97] 周强，张伟，俞士汶 (1997). 汉语树库的构建[J]，《中文信息学报》，11(4)，1-11