# 自底向上的汉语句法标注体系设计与实践

杨沐昀 赵铁军 于浩

哈尔滨工业大学计算机科学与技术学院

E-mail: {ymy;tjzhao;yu}@mtlab.hit.edu.cn

摘要：句法标注体系是树库加工基础。在对比分析了目前3种典型的汉语句法标注的基础上，本文提出了一种“自底向上”的汉语句法标注体系的设计方法，完成了一套比较完整的汉语句法标注集。该方法避免了从句法理论角度设计标注体系的难题，在实践中取得了良好的效果，具有比较高的工程可行性。

关键词：句法标注体系，汉语树库，自底向上

# Bottom-UP Design of Chinese Syntax Annotation Scheme

Yang Muyun  Zhao Tiejun  Yu Hao

School of Computer Science and Technology

Harbin Institute of Technology, Harbin, 150001

E-mail: {ymy;tjzhao;yu}@mtlab.hit.edu.cn

ABSTRACT: Syntax annotation scheme is indispensable to any treebank project. To collect a proper scheme for Chinese treebank construction, this paper presents a bottom-up strategy. The proposed method avoids difficulties in choosing Chinese syntax scheme according to linguistic theory and is feasible in treebank annotation. The paper also reports its practice results in bracketing 8000 Chinese syntactic trees.

Keywords: Syntax annotation scheme, Chinese treebank, bottom-up

## 1. 引言

汉语句法标注集是汉语句法分析的知识表示基础，也是任何树库建设中必须首先解决的基本问题之一。在设计一个合理的汉语句法标注体系过程中，目前汉语树库的研究人员遇到一个很大的困难：和英语等西方语言相比，汉语自然语言处理研究尚缺乏一个合适的理论基础和描述体系，也没有一个适合描述汉语句法的理论。

其实，汉语句法分析理论一直是语言学上研究的热点之一。这就出现了一个现象：一

方面汉语句法理论的论述浩如烟海，另一方面计算语言学研究人员却无法根据这些理论拿出一个可靠的标注集。造成这种现象的原因来是多方面的。其一，汉语是一种“意合”语言，信息丰富，歧义众多。从理论上解释这样的语言本身就是一个难题，所造成的结果就是目前的汉语句法理论纷繁复杂，诸多问题尚无定论。也就是说，汉语的自身特点使得西方语言学的一些成果和经验不能完全适用于它[1]。其二，从事计算语言学研究的人员大多不是语言专家，特别是对于微妙的汉语句法理论掌握起来也是勉为其难。而且目前能够将汉语树库标注和汉语句法研究结合起来的研究单位屈指可数。其三，由于汉语的特殊性，简单的句法标注体系不能满足研究需要，而一旦设计出一个比较完善的标注体系后，在标注实践中又会遇到培训代价高、标准不易掌握、标注效果差等问题。（其实，即使是对于英语来说，理论语言学的成果在语料加工过程中与其说是在指导标注实践，不如说是在发现原有理论的缺陷和不符实际[2]）。

为了解决汉语句法标注体系设计过程中理论准备门槛过高的问题，同时能够在一个比较合理可靠的基础上开展汉语树库的加工工作，本文在比较了目前 3 种有代表性的汉语句法分析体系之后，采用了一种“自底向上”汉语句法标注体系设计策略，并介绍了该方法在实践中的初步结果。

## 2. 现有汉语句法标注体系的初步对比

句法标注体系设计的一般首先要从现有的句法标注实践中研究需解决的基本句法现象。为此，我们收集了目前计算语言学研究中实际采用的几种汉语句法分析标准。在这里将简单介绍其中三种典型标准的基本情况(见表一)。

清华大学的句法标注体系包含 16 个短语符号（见表一），除了常见的几种短语符号外，该体系中有 4 个符号专门刻划汉语句子类型。目前，该单位的研究人员正在致力于大规模汉语树库的开发工作。

台湾中央研究院中文句结构树库（Sinica Treebank）目前的 1.0 版包含 38,725 棵中文句法分析树，239,532 个词（http://godel.iis.sinica.edu.tw/CKIP/treebank/）。该树库的理论基础是“基于信息的格文法（Information-based Case grammar，简称 ICG）”[3]，不仅标注了语法结构，还标注了丰富的语义信息。表一简要地列出该体系中的短语结构符号。

美国宾夕法尼亚大学著名的语言数据协会( Linguistic Data Consortium，简称 LDC)在 2000 年发布了其汉语树库语料(Chinese Penn Treebank)。该项目始于 1998 年夏，已完成 100,000 词，共 4185 句。全部 325 篇文章来自 1984~1998 年的新华社文章，采用 SGML 标注（http://www.ldc.upenn.edu/ctb/）。根据其标注手册[4]，表一列出了其中的短语符号。除了这些符号以外，该树库中还使用了 24 个附着在这些短语符号后面的功能标记。

通过对这三个句法分析体系的对比分析，我们可以发现：

1）汉语句法分析体系反映了所采用的词性标注体系。清华大学体系中所采用的时间词短语、处所词短语和区别词短语都反映了《北大汉语信息词典》中对汉语词类划分的认识。

2）名词短语、动词短语、形容词短语、副词短语、介词短语、小句（或称从句）等是公认的短语类型，可以称之为汉语句法核心短语。

3）虽然每种体系都力求通用性（general purpose），但是每个体系彼此间的对应转换还是很困难的。也就是说，即使不考虑不同的语言学理论基础，一个标注集仍然无法完全代替另一个标注集的标注工作。

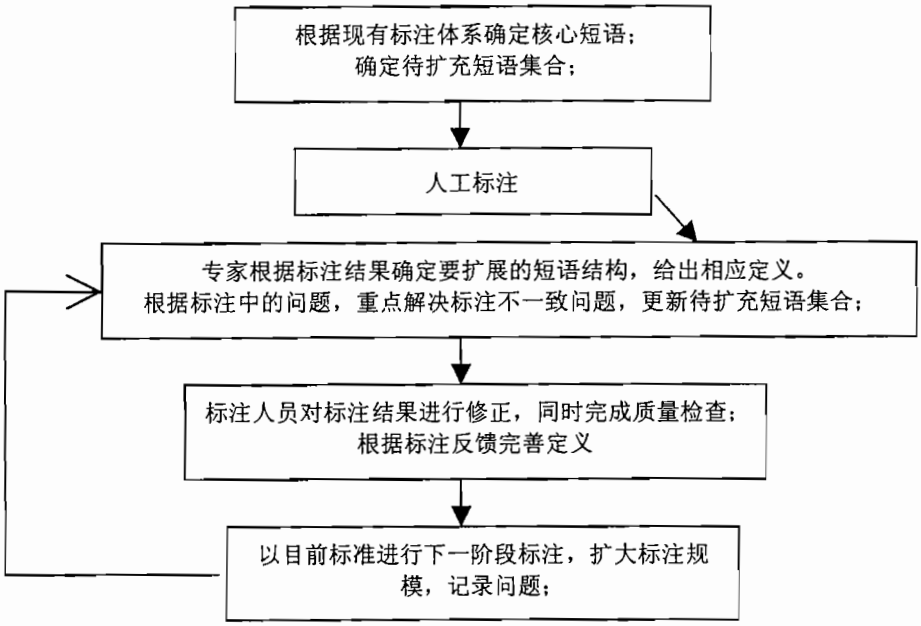
| 序号 | 名称 | 清华大学 | 台湾 | LDC 汉语树库 |
|---|---|---|---|---|
| 1 | 的字结构 | | | DNP |
| 2 | 介词-方位结构 | | GP | LCP |
| 3 | 名词短语 | np | NP | NP |
| 4 | 时间名词短语 | tp | | |
| 5 | 地点名词短语 | sp | | |
| 6 | 介词短语 | pp | PP | PP |
| 7 | 形容词短语 | ap | A?的 | ADJP |
| 8 | 区别词性短语 | bp | | |
| 9 | 副词短语 | dp | ADV?的、ADV?地 | ADVP、DVP |
| 10 | 数量短语 | mp | DM | QP |
| 11 | 指量短语 | | | DP |
| 12 | 数词准短语 | mbar | | |
| 13 | 动词短语 | vp | VP | VP |
| 14 | 从句 | S | S | IP、CP |
| 15 | 并列结构 | | | UCP |
| 16 | 独立成分 | dlc | | PRN |
| 17 | 片语 | | | FRAG |
| 18 | 感叹语 | | | INTJ |
| 19 | 列表 | | | LST |
| 20 | 分类短语 | | | CLP |
| 21 | 单句 | dj | S/句子 | |
| 22 | 复句 | fj | | |
| 23 | 整句 | zj | | |
| 24 | 句群 | jq | | |
| 25 | 直接引语 | yj | | |
| 26 | 未知成分 | | | X |

表一、3 种句法标注体系对照表

## 3. 自底向上的句法标注体系的设计与实践

在对比分析之后，传统的标注体系设计往往采用“兼容并蓄”的策略，在所采用的理论框架允许的范围内尽可能地吸收各家之长。同时，也必然根据实际情况，去掉其中一些不符合特定需求的类型划分。从整体上看，这是个先求并集再删减的过程，而且在标注之前要形成一个比较详细的标注手册。

但是语料标注尤其是树库标注使一个主观性很强的工作，即使是对于英语也不例外。一个著名的例子是在计算语言学协会（ACL）1991 年的年会上，来自 9 个研究机构的自然语言处理研究者对同一个英语句子进行了分析，得到的结果如下[5]：He said this constituted a (very serious) misuse (of the (Criminal Court)process)，其中只有括号部分是大家都同意的划分，对于其余部分的划分大家都是各持己见。可以想见，对于语法语义十分灵活的汉语来说，采用传统方法，从理论上设计一个较理想的汉语树库标注体系是件十分困难的是事请。




图一、自底向上策略的工作流程

为了解决这一问题，本文采取了一种与传统方法不同的自底向上的设计策略。如图一所示，我们首先根据上文的对比分析可肯定地得出结论：名词短语、动词短语、形容词短语、副词短语、介词短语、数量短语、小句是汉语句法核心短语，是任何句法标注工作都不可或缺的语法结构。这些短语一般为各种语言理论所公认。这样我们确定了汉语句法标注最初的 7 个短语符号，而且这些短语即使对于普通标注人员（非语言专业大学生）也不必进行过多的解释。树库标注的第一阶段就可以在经过分词、词性标注的语料中迅速展开，标注过程此时可以大部分依赖标注人员的语言直觉进行，还没有翔实的标注手册。

对于其他短语类型，我们并不是完全忽略，而是将剩下的各种短语归纳为某个核心短语的子类（如地点短语、时间短语、的字结构可以认为是名词短语的子类）或者是某种特殊语言现象（如非对称并列结构、片语）。此外，还要进一步根据树库的设计目标、用途以及以往的研究经验，扩展出目前标注体系中所忽略的短语类型（如汉语动词特有的并且对机器翻译影响较大的把字句、被字句、V 得如何如何等等）。将所有这些短语编入待扩展集合，晴标注人员观察并进行适当的纪录（此时，还不必对这些短语进行详细的定义）。

当标注结果达到一定规模的时候(比如上千句时，具体数量和所加工的语料性质有关)，

第一期标注工作结束。此时，标注人员既完成入门培训，熟悉了工作流程，掌握了辅助工具，又完成了一定规模的树库标注（当然还比较粗糙，不能满足需求）。而语言工程师（不一定是语言学权威人士）和计算语言学研究人员此时就可以对这些基本短语分类研究，主要解决以下问题：

- 根据语料情况，总结上阶段标注工作中的问题，尤其是不一致现象以及各种错误，有针对性地写出这些核心短语的初步标注规范；
- 根据分类考察结果以及标注人员的纪录和统计结果，判断是否需要扩展某个核心短语，决定标注体系中增加哪些新的短语符号；
- 根据语料，给出这些新增短语的标注说明；
- 考虑是否有新的待扩展短语分类；这些新的扩展可能来自于实际语料中的现象，也可能来自系统开发的需要；

| 序号 | 符号 | 短语/结构名称 | 注释或实例说明 |
|---|---|---|---|
| 1 | AP | 形容词短语 | |
| 2 | ASIDE | 似的结构 | 象 X 似的/象 X 一样/X 似的/X 也似的… |
| 3 | CO | 并列结构 | 性质不相同的短语的并列形式 |
| 4 | DP | 副词短语 | |
| 5 | INP | 插入语 | 括号或者破折号之间的短语，包括括号或破折号 |
| 6 | MP | 数量短语 | |
| 7 | NP | 名词短语 | |
| 8 | NDE | 的字结构 | X+的 |
| 9 | NS | 处所名词短语 | 超出了基本短语识别范围的处所名词序列 |
| 10 | NT | 时间名词短语 | 超出了基本短语识别范围的时间名词序列 |
| 11 | PP | 介词短语 | |
| 12 | PFP | (介词)方位结构 | p+X+f 序列形式，p 可以省略 |
| 13 | SS | 小句(或子句) | |
| 14 | VP | 动词短语 | |
| 15 | VBA | 把字结构 | “把”字及其宾语均作为整个短语的成分 |
| 16 | VBEI | 被字结构 | “被”字及其后面的施动者均作为整个短语的直接成分 |
| 17 | VC | 动补结构 | |
| 18 | VJ | 兼语结构 | |
| 19 | VO | 动宾结构 | |
| 20 | VOO | 双宾结构 | |
| 21 | VSUO | 所字结构 | 带“所”字的动词短语，“所”字与后面的动词构成该短语 |
| 22 | VV | 连动结构 | |
| 23 | XP | 搭配结构 | 一些不易归类的体现汉语特点的搭配结构 |

表二、哈工大汉语句法标注集

随后，标注人员将根据新产生的标注规范，对第一期的标注结果进行改进，同时可以进行交叉校验。标注过程中仍然要纪录标注定义中的问题以及特殊的语言现象，特别是要找出定义中模糊不清或定义不完善的情况。在完成第一期标注任务后，如果定义基本可行，则可以以此为依据，进行下一期的标注工作。“标注—检验—扩充定义—再标注”，通过大

约几次这样的循环，标注体系基本就可以稳定下来，以后的工作就是不断通过实践完善标注手册的过程，到此所有的语料标注工作情况就大体相似了。

应用这一方法，我们标注完成了 8000 句汉语树库，共计 66,038 个词，平均句长 8.25 汉字。标注人员主要来自工科专业大学四年级学生，工程先后进行了大约 3 个月。在标注过程中，经过大约 2000 句的标注实践，含有 23 个短语符号标注体系就基本确定下来（见表二），但是标注手册的修订一直持续到标注过程的结束。而且，随着语料的进一步扩大，相信仍然有需要完善的地方。

自底向上的句法标注设计策略实际上是从语料库标注实践出发的一种描写主义的方法。其主要优点是具有比较好的工程可行性，能够迅速的完成一定规模的树库，满足实际系统开发的需要。同时，能够为今后树库标注积累充分的实践经验。但是，这种做法仍然有很多地方值得进一步说明。

其一，自底向上的策略中的一个明显的问题是改变了语言理论对标注实践的先导地位，有悖于“语言理论为标注体系提供了答案”的理念，是否有“南辕北辙”之嫌。好在语言学研究中还有一种意见认为：语料标注应该独立于任何理论，否则就是本末倒置，因为语言标注就是要收集并纪录原始数据，为语言学理论准备素材（也为应用语言学准备数据）。如果后一种意见成立，即语言学就是要对客观事实作出解释，那么在决定树库中收录什么短语类型之前，句法理论是无法最终完成的。实质上，第一种意见是把语料标注看作是用理论解释现象的过程，而后一种意见是把语料标注当作搜集素材归纳理论的过程。本文的做法明显符合后者的认识，即把语料标注本身当作是逐步完善定义的过程。

其二，没有理论指导的树库标注必然缺乏完备性和通用性。诚然，这是目前该策略中的一个主要问题。但是，通用性和完备性是一个相对的概念，是一个实践中几乎无法完全实现的标准。在分析上述几个标注集的过程中，我们发现每个标注集其实都只适用于他所针对的那个领域的语料。当语料变换（甚至是语料扩大）时，就会暴露出其局限性。所以，就目前的骨架句法分析（skeleton parsing）研究来说，本文的做法是可以满足需要的：语料中所标注的短语都是最基本、最必要的短语划分，虽然不能保证进一步研究的需要，但是可以最大限度利用有限的投入。

其三，本文的方法似乎忽略了标注手册的作用，这一点在以往的语料加工中是不可想象的：一般做法都要先书写细致的标注规范，其详尽程度只是考虑到培训时间的限制才有所裁减。的确，一本好的标注手册可以有效地保证语料标注的质量。但是本文提出的方法并不是忽视标注手册的作用，而是要在实践中逐步完善形成一部操作性更强的标注规范，而且这个规范由语言工程师和标注人员在工作中共同形成，在书面和标注人员的脑海里同时形成。这样做，优势在于避免了冗长枯燥的初期培训，最大限度的节省了人力和时间。

从另一个侧面来看，标注时最大问题就是在语言现象（实际语料）的模糊性，许多在实际语料加工过程中遇到的特殊语言现象恰恰是汉语语法研究著作所没有涉及的。某些语

法区别对于语言学家甚至是标注者来看都是基本的语言区别，但是某些特殊情况下非常难以标记。标注手册很难在制定之初预料到这些问题，解决方法也只能是随着标注工作的进行，通过个例说明加以完善。大量的英语语料加工实践也表明：哪怕是语言专家们有时也很难一致地区分真实语料中的现象，因此本文提出的方法在标注树实践之初弱化标注手册的作用也是一个现实的选择。

最后值得指出的是，不管人们承认与否，目前树库体系的设计实际上更多的是依赖于人的直觉和常识（或者说语言素养），而不是什么逻辑思维的结果。这个问题带来的直接后果就是：研究人员充满希望要建立正确合理的句法分类，但却在实践中被大量的语言特例把已有的理念打得支离破碎，甚至工作越深入越发现得到的数据错误百出，一无是处。也许这一现象的根源来自于于语言研究中的悖论：凡规则必有例外，但是无论如何这一问题必须在句法标注中予以解决。

## 4. 结束语

本文在对比分析了我国清华大学、台湾中央研究院、美国宾夕法尼亚大学语言数据小组的三种汉语句法分析体系的基础上，采用一种自底向上的汉语句法标注体系设计策略，完成了一套包含 23 个短语符号的标注集合，开发了包含 8000 句的汉语树库。该方法避免了汉语树库标注过程中的语言学理论门槛，而且避免了语料库加工过程中的诸如标准设计、人员培训等难题，可以作为计算语言学研究中解决问题的一个思路。

致谢：感谢清华大学孙茂松老师、周强老师为本工作提供的资料；感谢参加标注集讨论的机器翻译研究室成员：吕雅娟、刘芳、刘晓军、康振国、李曼丽、但汉松、常薇等。同时还要感谢那些和我们一起完成了这次汉语树库建设工作的全体同学。

## 参考文献

[1] 许嘉璐．现状和设想——试论中文信息处理与现代汉语研究．中文信息学报，2001，15(2)

[2] Geoffrey Sampson. Where Should Annotation Stop. Proc. of the Workshop on Linguistically Interpreted Corpora(LINC-2000). June, 2000. PP28-34

[3] KEH-JIANN CHEN, et al . The CKIP Chinese Treebank: Guidelines for Annotation. http://godel.iis.sinica.edu.tw/CKIP/treebank/

[4] Wenxue Nian, Fei Xia et al, ed. The Bracketing Guidelines for the Penn Chinese Treebank (Draft II) . http://www.ldc.upenn.edu/ctb/

[5] R. Garside, G. Leech and T. McEnery. Corpus Annotation: Linguistic Information from Computer Text Corpora. Longman, 1997.