汉语特殊结构的句法标注策略*

高松　赵怿怡　刘海涛
中国传媒大学应用语言学研究所 北京 100024
E-mail: gaos_0808@sina.com

摘　要：句法树库建设是当今国内外计算语言学研究的热点之一。本文探讨了在依存树库中，分析和处理汉语特殊结构的一些问题，如："X 是"和"X 说"结构、离合词+趋向补语结构、特殊的重叠结构。通过对比短语结构树库和依存树库对这些结构的处理方法，给出它们在我们的依存树库中的处理规则，做出具有可操作性的依存句法分析，确定统一的标注原则，为后续的相关研究积累经验。这说明不断地完善树库标注的规则对树库建设具有重要的作用。

关键词：依存树库，短语结构树库，"X 是"，"X 说"，离合词，重叠

Syntactic Tagging Strategy for Chinese Special Structures

Gao Song　　Zhao Yiyi　　Liu Haitao
Institute of Applied Linguistics, Communication University of China, Beijing 100024
E-mail: gaos_0808@sina.com

Abstract: Treebanking is one of hot issues of computational linguistics in domestic and abroad. This paper discusses how to analyze and process some problems of Chinese special structures, such as "X shi" and "X shuo" structures, clutch words with directional complement, and special overlapping structure. Comparing the processing approaches of phrase structure treebank with those of dependence treebank, we provide the related rules for our dependence treebank, make operable analysis under dependence syntax, and establish a uniform tagging principle which will be useful for the future researches. It shows that improvement of treebank tagging rules plays an important role in treebanking.

Keyword: Dependence Treebank, Phrase Structure Treebank, "X shi", "X shuo", clutch words, overlap.

1 引言

语料库的句法标注是语料库语言学和计算语言学研究的一个前沿课题。它的目标是对语料文本进行句法分析和标注，形成树库。树库作为获得句法结构的知识源和评价句法分析结果的工具，越来越受到研究者们的重视。（周强 1997，Liu 2007）

目前，国内外研究者们已开发了一些大规模的树库。这些树库有两种占主流的句法标注体系：一为基于短语结构语法 PSG(phrase structure grammar)的句法标注，另一个为基于依存语法 DG(dependency grammar)的句法标注。汉语方面，基于 PSG 建立起的短语结构树库，如：美国宾夕法尼亚大学的汉语树库 Penn Chinese Treebank(PCT)、清华大学的汉语句法树库 Tsinghua

* 本文为中国传媒大学'211 工程'三期重点学科建设项目阶段成果

Chinese Treebank(TCT)、台湾中央研究院的汉语（繁体）树库 sinica；基于 DG 建立起的依存树库，如：哈工大信息检索室的汉语依存树库[1]和中国传媒大学的汉语依存树库（Liu 2007）等。

短语结构树库，采用部分与整体的方式来描述句法的结构。（党政法等 2005）如图 1 为汉语句子“中国是多民族国家”的短语结构图。依存树库则是通过建立词语之间的联系来描述句法的结构，它以依存关系为基础。图 2 为汉语句子“中国是多民族国家”的依存句法结构图示。

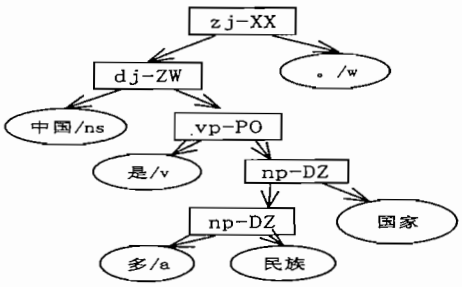

图 1“中国是多民族国家”的短语结构图

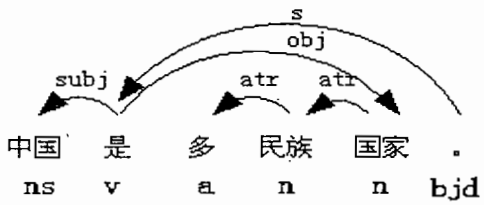

图 2 “中国是多民族国家”的依存句法结构图

由图 2 可见，依存关系是两个词之间一种有向的、非对称的关系。它具有三个组成部分：支配词、从属词、依存关系标记。句子中的每个词都有自己的支配词，即它是受哪个词支配的，它依存于哪个词。把这种依存关系用符号标记出来，这些符号就是依存关系标记。

我们标注的树库是面向有声媒体语言的。语料选自中央人民广播电台的“新闻和报纸摘要”、“今日论坛”和北京人民广播电台的“603 访谈时间”节目。我们从中各选出一期节目，共三期，来进行树库标注。这种语料选择是考虑到它们既包含新闻播报类又包含访谈对话类，涉及的范围和内容比较广泛。语料中共有 480 个句子，12393 个词（含标点符号）/10870 个词（不含标点符号）。标注体系，采用的汉语依存关系句法标注体系为 Liu/Huang(2006)。

从之前构建的汉语依存树库中（周明等 1994，刘伟权等 1996，Liu 2007，关润池等 2007），我们吸收了很多经验，提高了句法分析和标注的效率。但同时也在标注过程中遇到了一些特殊的结构，值得深入考虑。如“X 是”和“X 说”结构、“离合词+趋向补语”结构、特殊的重叠结构。这些结构都是汉语中特殊的语言现象，出现的频率不低。如果对其标注规则不进行规范的话，将影响到树库标注的一致性问题，进而也会影响到树库标注的质量。本文以这几种特殊结构为研究对象，对比现有的短语结构树库和依存树库对其的处理方法，给出这些结构在我们依存树库中的处理规则，做出具有可操作性的依存句法分析。同时，通过不同标注体系的树库的对比，有助于我们对所处理的结构有更清楚的认识。

2 几种特殊结构的依存句法分析

2.1 “X 是”和“X 说”结构

标注的树库中，一些连词和副词（双音节居多，也有单音节和多音节）后面经常出现一个“是”或“说”字，形成“X 是”、“X 说”组合。其中的“是”不再表示判断或强调，“说”不再表示具体的言说之义，它们的意义变得很难分析，产生了虚化。“是”、“说”与其前成分“X”的关系更为紧密，没有了语音上的停延。这种组合的出现频率不低。以下是树库中“X 是”、“X 说”组合的例子：

“商品标注不明确，特别是不少内容让消费者很难去理解和操作……”

[1] 哈工大信息检索研究室依存树库网址:http://ir.hit.edu.cn/demo/ltp/Sharing_plan.htm (2009-3-12)

“比如说我们有天然气化工，但是我们没有石油化工。”

宾州汉语树库 PCT，对这种结构的处理，如下（Fei 2000）：

For simplicity, we treat 特别是[particularly]/AD as one word.

可见，PCT 把“特别是”视为一个词来处理，词性为副词。

清华大学树库 TCT 中，党政法、周强（2005）提出对这种结构的句法标注如下：

[dlc-XX [tp-ZZ 特别 是/d 现代/t]]

其中，dlc 表示独立成分，XX 缺省结构，tp 为时间短语，ZZ 为状中结构。可见，TCT 将“X 是”结构同它后边出现的成分捆绑在一起，作为一个整体，结构标记为时间短语，功能标记为独立成分，而并非将“X 是”结构单独视为独立成分。“X 是”结构中，“X”和“是”拆开，当作两个词来处理。TCT 并未标记词间的关系，仅标记出单词的词性以及词与词组块形成的短语的内部结构和外部功能。

中研院的汉语树库 sinica 中[1]，对这种结构处理的线性结构如下：

S(agent: NP(Head: Nhaa: 他)|Head: VG1: 作為|range: NP(quantifier: DM: 一個|Head: Nab: 作家)|range: VP(manner: Dh: 按理|Head: VE2: 說|goal: VP(deontics: Dbab: 可以|deiXis: Dbab: 去|Head: VC33: 寫|theme: NP(Head: Nac: 小說)))) 。由该结构可知，sinica 将“X 说”结构切分成了两个词，“X”修饰述词“说”，语意角色为 manner，即方式，“说”是“X 说”结构的中心语。“说”后面的成分作它的述词论旨角色，标记为 goal。“X 说”和它后面的成分一起作动词中心“作为”的述词论旨角色，标记为 range。

哈工大信息检索研究室的汉语依存树库对“X 是”组合的处理如下：首先把它切分为两个词，即“X”和“是”。具体标注又分为两种情况：有些组合（如：特别是），将“是”作为“X”的附加成分，“X”支配“是”，依存关系为后附加关系，标记为 RAD。整个组合“X 是”作后面名词的状语，形成状中关系，依存关系标记为 ADV；有些组合（如：甚至是），将“是”作为谓词中心，“X”修饰“是”，形成状中关系，标为 ADV。对“X 说”组合的处理，该依存树库没有把所有情况都切分为两个词。切分为两个词的（如：不是说），形成“X”和“说”，“说”作“X”的补语，标为 CMP，“X”作后面谓词中心的状语，标为 ADV。没切分为两个词的（如：按理说），“X 说”是一个整体作后面谓词中心的状语，标为 ADV。

董秀芳（2003、2004）认为，这种组合形式已从非词单位变为了词，发生了词汇化。因为在句法上它们可以作为一个单位来使用，中间不能插入其他成分，“是”和“说”与其前成分都不能单独被修饰，这种组合的整体意义不是两个构成成分意义的简单相加。文中提到的“X 是”和“X 说”组合有：或者是、特别是、尤其是、可以说、比如说、难道说等。有些例子，可能有的人认为是词，有的人认为不是词。存在这种认识分歧是正常的，因为有些例子的词汇化程度还不高，其内部结构比较透明，还有可能被判断为短语。即使它们目前还不是词，但无疑也具有一定的词汇性质，有向词发展的强烈倾向。

因此，从定性分析的角度，我们有理由将这种组合按一个词来处理。在树库中，我们把这种组合当作一个整体，视为插入语。在结构上，它们和句子的其它成分没有联系，但在语义表达上还是有用的。我们将其定为一种附加语，其支配者为小句谓词，依存关系标记为 ina。用直观的依存句法结构图来表示，见图 3。

[1] 中央研究院汉语树库 sinica 网址：http://turing.iis.sinica.edu.tw/treesearch/ （2009-3-13）

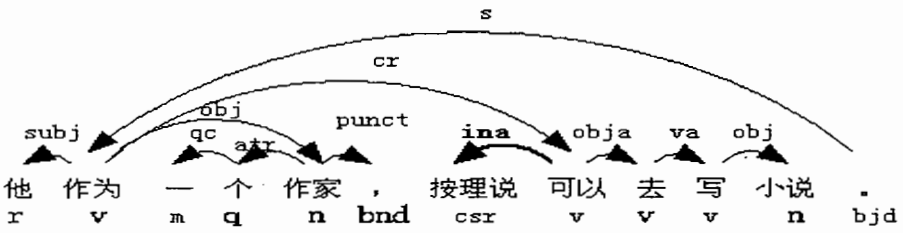

图3“他作为一个作家，按理说可以去写小说。”的依存句法结构图

2.2 “离合词+趋向补语”结构

离合词是现代汉语中一种特殊的语言现象。因为它能离能合，把词法和句法交织在一起，与其它词语很不相同。意义上，它具有整体性和单一性；结构上，它的两个语素之间结合得不太紧密，中间可以扩展，可以加入其它成分，具有短语的特点。单音节趋向动词要放在离合词中的动词之后,双音节趋向动词一般要把离合词中的宾语放在趋向动词的之间。如树库中的例子：

“南京楼市刮起精装修之风来，……”

“他觉得这不是一本时尚的书，而是一本让人静下心来慢慢读的书。”

通常，离合词“刮风”、“静心”在遇到双音节趋向动词“起来”、“下来”作补语时，离合词中动词后面的宾语放在了趋向动词之间。该结构在树库中，我们要考虑如何对其进行依存句法分析。

宾州中文树库的标注手册中，对动词后有趋向动词的情况，没有给出动宾式动词（即离合词）带趋向动词的处理方法。只给出了单音节动词+双音节趋向动词、双音节动词+单音节趋向动词、双音节动词+双音节趋向动词的标注方法。具体处理如下（Fei 2000）：

```
(VP (VRD (VV 降 )
         (VV 下来)))
(VP (VRD (VV 提取 )
         (VV 出)))
(VP (VRD (VV 联合 )
         (VV 起来 )))
```

清华大学树库 TCT 处理这种结构，会将整个动词短语切分出来，进行标注。

如：[vp-SB 发展/v 起来/vB]。vp 表示动词短语，SB 表示述语结构。对于这种内部结构复杂的动词短语（如：静下心来），TCT 没有处理其内部各成分间的关系。

中研院汉语树库 sinica,对离合词后带趋向补语处理的线性结构如下：S(agent: NP(Head: Nhaa: 他)|Head: VC31: 装出|theme: NP(quantifier: DM: 一副|Head: Nab: 笑臉)|duration: Dbab: 來)。sinica 把“装出”处理为一个词，它是句子的中心语。“一副笑脸”在语意上是“装出”的述词论旨角色，标记为 theme。“来”是修饰述词“装出”的语意角色，标记为 duration。数量短语“一副”中数词和量词的关系没有细分。

哈工大的依存树库中，对这种结构的处理如下：

“这件事，教育部、各省市都要下气力抓，要尽快地抓出成果来。”

例句中，“抓”支配“出”和“来”，依存关系标记为 CMP，即动补结构。“抓”支配“成果”，依存关系标记为 VOB，即动宾关系。

哈工大树库对这种结构标注的方法值得提倡。因为只有这样，才不致产生交叉弧。在依存句法分析中，交叉弧往往是衡量一个句子语法是否合格的重要条件。我们的处理方法同哈工大树库

一致，即：离合词中动词支配后面拆分开的趋向动词，依存关系为补语，标记为 comp。同时，它还支配离合词宾语，依存关系为宾语,标记为 obj。直观的表示见下图 4。

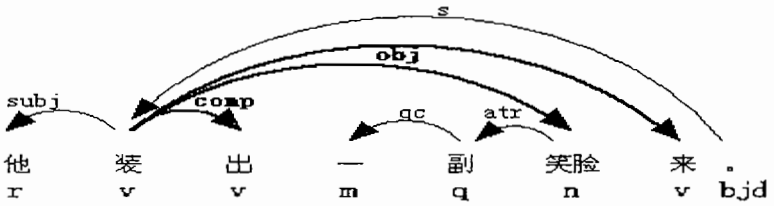

图 4 “他装出一副笑脸来”的依存句法分析图

2.3 特殊的重叠结构

树库中，我们将词的重叠形式视为词的特殊结构，对其进行了切分。通过以往的标注，我们总结了重叠结构的标注经验，提出了标注的原则，即“顺序原则”。该原则处理一般的重叠结构具有普适性，但对于某些特殊的重叠结构，我们仍需进一步探讨分析方法。树库中有：

单音节动词重叠 A 一 A、A 了 A 格式，如：试一试、想了想

双音节形容词重叠 ABAB 格式，如：冰凉冰凉、雪白雪白

副词重叠 AA 格式，如：特别特别、非常非常

宾州中文树库 PCT，Fei(2000)对这种特殊的重叠结构的处理如下：

A-one-A: (A/V one/CD A/V)/V

EX: (想[think]/VV -[one]/CD 想[think]/VV)/V

A-le-A: (A/V le/AS A/V)/V

EX: (想[think]/VV 了/AS 想[think]/VV)/V

ABAB: AB is a verb; ABAB/V

EX: 研究研究[research]/VV, 雪白雪白[snow white]/VA

清华大学树库 TCT 将 A 一 A, A 了 A 视为动词性准短语，标为 vbar；副词重叠，如：非常非常，为副词性短语，标为 dp。短语结构内部词间的关系不作分析。

中研院树库 sinica,对以上几种格式进行分析的线性结构如下：

VP(Head: VF1: 試|goal: VP(time: Dd: 一|Head: VF1: 試))

VP(Head: VE2: 想|aspect: Di: 了|goal: VE2: 想)

VP(Head: VH11: 冰涼|Head: VH11: 冰涼); VP(Head: VH11: 特別|Head: VH11: 特別)

sinica 将动词重叠“A 一 A”中的“一”来修饰后边的“A”，语意角色标记为 time。前面的“A”为中心动词，“一 A”共同来修饰前面的动词“A”，语意角色标记为 goal。动词重叠“A 了 A”中的“了”和后面的“A”都修饰前面动词“A”，“了”的语意角色为 aspect, 后面的“A”的语意角色为 goal。对形容词重叠 ABAB 式和副词重叠 AA 格式，都处理为并列结构。

哈工大依存树库对单音节动词重叠 A 一 A 格式的处理如下：

“梁连起从王学强那里了解到下叔村贫困的情况，他决定到下叔村看一看。”

在这个句子中，前一个动词“看”支配后一个动词“看”，依存关系标记为 VV，连谓结构。后一个动词“看”支配前面的数词“一”，依存关系标记为 ADV，状中结构。

参照他们的处理方法，我们的处理原则为：对双音节形容词重叠 ABAB 式和副词重叠 AA 式，采取前一个词支配后一个词，依存关系为并列，标记为 coor。单音节动词重叠 A 一 A 式，前一个动词支配后一个动词，依存关系为连动，标记为 va, 后一个动词支配数词“一”，关系为

adva。单音节动词重叠A了A式，前一个动词支配后一个动词，依存关系为连动，标记为va，“了”受前一个动词支配，依存关系为时态附加语，标记为ta。

3 结语

本文通过不同树库对以上几种特殊结构的处理和学者们的研究（周强2004，刘海涛2007），看到了短语结构树PST (phrase structure tree)和依存树DT (dependency tree)的不同之处： PST采用部分与整体的方式来描述句法的结构。将句子分“块”来描述，着重对“块”的外部功能进行描述。DT通过建立句子中词与词之间的关系来描述其结构，分析出句子中各词之间的支配关系、从属关系；PST注重的是句子的表层句法结构，而不是深层意义关系，它没有表示出词间的关系，也就无法描述句子的语义。DT既可以描述语法，又可以把标注出的词间依存关系转化为语义描述；PST不能给出句子的中心词信息。DT能标示出句子中的中心词，给出中心词信息；PST可以对不同层次的句子成分的组合进行细致地描述，但其节点较多，层次较深，存储的空间大，操作不便。DT节点较少，没有短语节点，每个节点都与句子中的词相对应。同一个节点上标注出单词的词类信息和句法功能信息，是一种“多标记树模型”（multi-labeled tree model）。它需要的存储空间少，而且便于操作。对于汉语这种通过语序和虚词来表示语法意义、词与词之间存在不对称性的语言来说，依存语法模型似乎是一种适宜的形式化描述工具。

对比短语结构树库和依存树库对汉语中几种特殊结构的处理，做出具有可操作性的依存分析，确定统一的标注规则，能提高树库标注的质量和树库标注者的标注效率，也为进一步的树库研究工作提供参考。这说明不断地完善树库标注规则在树库建设中有重要作用。对这几种特殊结构的句法标注策略的探讨使我们对所处理的结构有了更清楚的认识。我们进一步要做的是：在树库中，挖掘更多汉语中的特殊现象，探讨其特殊性和相应分析策略，如，其它词汇化格式在依存树库中的处理以及其句法功能、语义功能等问题，进而完善汉语依存树库标注体系。

参 考 文 献

[1] Liu H, Huang W. (2006) A Chinese Dependency Syntax for Treebanking. *Proceedings of The 20th Pacific Asia Conference on Language, Information and Computation*. 2006. Beijing: Tsinghua University Press, 2006:126-133.

[2] Liu, H. (2007) Building and using a Chinese dependency Treebank. *grkg/Humankybernetik*, 2007, 48(1): 3-14.

[3] Fei Xia. (2000).The Segmentation Guidelines for the Penn Chinese Treebank (3.0)[R].Technical Report IRCS, University of Pennsylvania.

[4] 刘海涛. 泰尼埃的结构句法理论[J]. 《北华大学学报（社会科学版）》，2007（5）：68-77.

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

[6] 周强. 汉语句法树库标注体系[J].《中文信息学报》，2004,18(4):1-8.

[7] 党政法，周强. 短语树到依存树的自动转换研究[J].《中文信息学报》，2005（3）：21-27.

[8] 周明，黄昌宁.面向语料库标注的汉语依存体系的探讨[J].《中文信息学报》，1994(3):35-51.

[9] 刘伟权，王明会，钟义信.建立现代汉语依存关系的层次体系[J],《中文信息学报》，1996（2）：32-46.

[10] 关润池，赵怿怡.口语依存树库中特殊结构处理[A],2007年计算语言学会议论文集[C].大连，2007.

[11] 董秀芳.“X说”的词汇化[J].语言科学，2003（2）.

[12] 董秀芳.“是”的进一步语法化：由虚词到词内成分[J].当代语言学，2004（1）.