文章编号：1003-0077(2018)05-0031-11

短语结构树库向句式结构树库的自动转换研究

张引兵[1,2]，宋继华[1]，彭炜明[1]，赵亚伟[1]，宋天宝[1]

(1. 北京师范大学 信息科学与技术学院，北京 100875；
2. 淮北师范大学 数学科学学院，安徽 淮北 235000)

摘　要：该文从短语结构和句式结构的区别与联系入手，设计了一种将短语结构自动转换为句式结构的算法。并以清华短语结构树库(TCT)为测试语料，实现了将大规模短语结构语料向句式结构语料的转换。最后，搭建了一套可扩展的可视化系统，用于不同句法结构语料的可视化查看。这一研究不仅实现了两种结构之间的初步转换，而且极大地丰富了汉语句本位图解树库的语料规模，并为汉语句本位图解树库的后续应用研究奠定了基础。

关键词：转换方法；树库；短语结构；句式结构

中图分类号：TP391　　**文献标识码**：A

Automatic Conversion of Phrase Structure Treebank to Sentence Structure Treebank

ZHANG Yinbing[1,2], SONG Jihua[1], PENG Weiming[1], ZHAO Yawei[1], SONG Tianbao[1]

(1. College of Information Science and Technology, Beijing Normal University, Beijing 100875, China;
2. School of Mathematical Science, Huaibei Normal University, Huaibei, Anhui 235000, China)

Abstract: This paper puts forward a method to convert a phrase structure Treebank to a sentence structure Treebank. Taking Tsinghua Constituent Treebank(TCT) as the test corpus, we realize the conversion of the two structures, together with the visualization display of them using a scalable visualization system. This study enlarges the scale of Chinese sentence structure Treebank, which will promote the follow-up researches of it.

Key words: method of conversion; Treebank; phrase structure; sentence structure

0　引言

树库是标注了句法信息的语料库，是一种深度标注的语言知识资源。一般来说，一个句子虽然表面上呈现词语的线性排列，但其内部的成分组织还是存在一定层次结构的。这种层次结构通常用“树”这种形式工具来表示，大量句子及其对应的树结构的集合就构成了树库[1]。然而，标注树库是一项费时费力的工作，需要完善的标注体系和规范的标注流程以保证标注的质量。另一方面，由于标注规范的复杂性，需要标注者拥有相关的专业背景。即使这样，标注者对句子不同的理解也会产生不同的标注结果，这为树库的建设带来了一定的困难。

基于上述弊端，目前树库的构建主要有两种方法：一是构建自动句法分析器；二是对标注好的另一种体系下的高质量语料进行转换。对于第一种方法，梁欣、臧德滋等人[2]已做了相关的研究；对于第二种方法，党政法[3]、李正华[4]、邱立坤[5-6]以及周惠巍等人[7]的研究也具有十分重要的意义。在树库的转换研究中，Lin[8]较早地进行了将短语结构树库向依存结构树库转换的尝试。Fei Xia[9]在 Lin 的基础上对其算法进行了进一步的完善，完成了从 Penn Treebank 到依存树库的转换，取得了较好的效果。另外，Hiroyasu Yamada[10]、Joakim Nivre[11]和 Tylman Ule[12]等也进行过一些树库转换相关的研究。纵观各种不同结构的树库，之所以能够从一种结构的树库向另一种结构的树库进行转换，是因为这些

收稿日期：2017-03-06　**定稿日期**：2017-04-26

基金项目：国家自然科学基金(61571049)；安徽省高等学校自然科学研究一般项目(KJ2016B002)

不同结构的树库标注方法虽然不同，但它们主要描述的都是句法结构，在更深层次上具有一致性。

目前计算语言学研究者已经为世界上许多语言构造了一定规模的树库，汉语方面也有一定数量的树库。因此如何减少树库建设中的工作量就成为一个重要的研究课题。利用已有的树库向目标树库进行转换，不仅可以减少重复劳动，还能提高工作效率。针对汉语树库，短语结构和依存结构的研究工作已经相当成熟，而句式结构的研究才刚刚起步，其相关研究主要在北京师范大学语言与文字资源研究中心开展。所谓句式结构，即以句本位语法为理论指导的一种图解语法结构。北京师范大学语言与文字资源研究中心在句本位理论的研究基础之上，开发了句式图解标注系统，进行句式结构树库的构建。实现了经典的语法理论与现代信息技术的结合，将复杂的句式结构通过句式图解的方式直观展现，更好地揭示了蕴含在语言内部的层次关系，从而使学习者更容易理清句子各成分间的逻辑关系，把握整个句子的句式结构。无论在中小学语文教学中，还是在国际汉语教学中都有着广泛的应用前景。本文旨在实现短语结构向句式结构的转换，提高句式结构树库的构建效率，扩充现有的句式结构树库的规模。

1 短语结构与句式结构树库及对比分析

目前，世界上成规模的树库主要有短语结构树库和依存结构树库两种类型。在中文领域，成规模的中文树库主要有宾州中文树库、Sinica 中文树库、清华中文树库、国家语委中文树库、北大中文树库、哈工大中文依存树库及北师大句本位句式结构树库。其中，宾州中文树库、清华中文树库、国家语委中文树库、北大中文树库均为短语结构树库[5]。下文给出了本文所采用的实验语料——清华短语结构树库（如无特殊说明，后文短语结构树库均指此库）和北师大句式结构树库的基本情况介绍与比较分析。

1.1 清华短语结构树库

清华短语结构树库由清华大学周强[13]等人构建。语料规模约五万句子、100 万词，涵盖文学、学术、新闻、应用文等多个领域。以“美国 T. A. 爱迪生发明了白炽灯。”为例，其存储形式为“[zj-XX [dj-ZW [np-DZ 美国/nS T. A. 爱迪生/nP] [vp-PO [vp-AD 发明/v 了/u] 白炽灯/n]] 。/。]”，图 1 展示了其短语结构树。

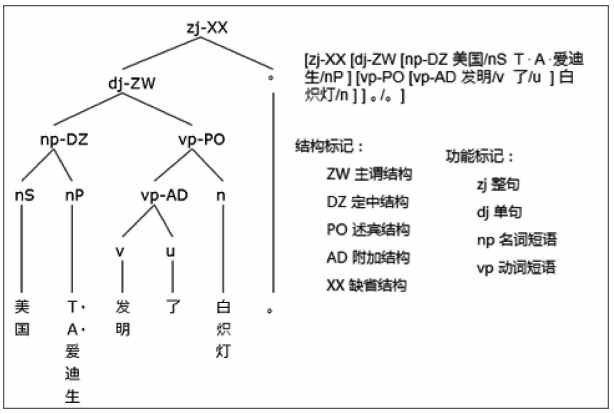

图 1 清华短语结构树示例

从图 1 中可见，除了词性节点（若不计词形节点，可视为叶节点）外，非叶节点均按“功能标记—结构标记”格式标记短语信息，例如，其中的“dj-ZW”节点，“dj”是其外部功能标记，表明这是一个单句；“ZW”是内部结构标记，表明其子节点是主谓关系。由于可以借助“短语”节点的层级嵌套，它可以刻画较为细致的层次结构（数据存储时通过括号的嵌套层级表示）。

在清华短语结构树库的标注体系中，采用了 16 个短语功能标记和 27 个句法关系标记，详细信息可以参考文献[13]。

1.2 北师大句式结构树库

1.2.1 句本位语法理论

在所有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[7]。句本位语法是黎锦熙先生在《新著国语文法》中建立的语法理论体系。其主张是在以句子为研究对象的基础上来研究语法，指出：“句本位的文法，退而分析，便是词类底细目；进而综合，便成段落篇章之大观。”[14]句本位语法以句子作为观察点和立足点，以句子成分和句法格局为主要特征，着力研究各类句式的结构规律。这种语法体系是在借鉴西方传统语法理论和体系，考虑汉语这种分析性语言的特殊性的基础上建立起来的，注重句法与语义的统一[15]。

关于句本位语法的学术思想和理论价值，黄昌宁先生曾经指出：“黎锦熙先生在《新著国语文法》中倡导的句本位语法体系和中心词分析法具有鲜活的生命力。”[16]所谓句本位语法，主要体现在两个方面：1）词类上“依句辨品、离句无品”；2）句法上采用中心词分析法，并以图解法作为析句工具。

1.2.2 句式结构图解标注平台

北京师范大学语言与文字资源研究中心的研究

着眼于现代汉语语法教学的实际需求，对黎氏体系中一些术语和图形展现方式进行了调整和完善。何静[17]、彭炜明[18-21]、杨天心[22]等在图解法的形式化及其树库的构建方面进行了深入的系统研究。在坚持黎氏语法句本位析句思想的同时，改进了部分陈旧的术语和分析方式，力求实现与经过多年发展的主流教学语法体系的兼容。按照“基本句式→扩展句式→复杂句式”的思路设计出一套形式化的句式图解系统方案，并实现了一个基于句本位语法的图解标注系统（见图 2），为汉语句式结构图解树库的构建奠定了基础。

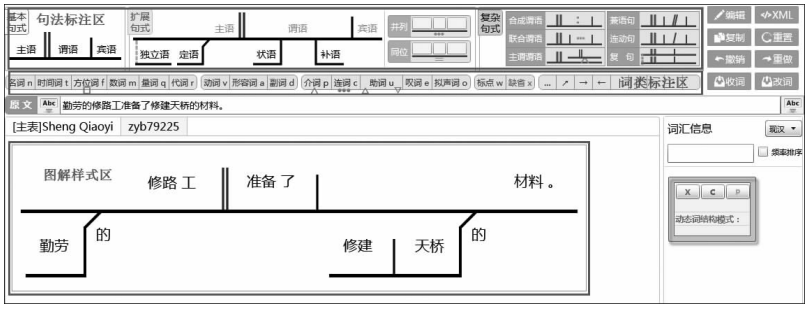

图 2　基于句本位的句式图解标注平台

图解析句法形式化设计方案，包括可视化的图解样式与 XML 格式的数据存储结构两个方面，二者是一种编码与解码的关系。一个具体例句的图解及其 XML 结构分别如图 3 和图 4 所示，标记集如表 1 和表 2 所示。

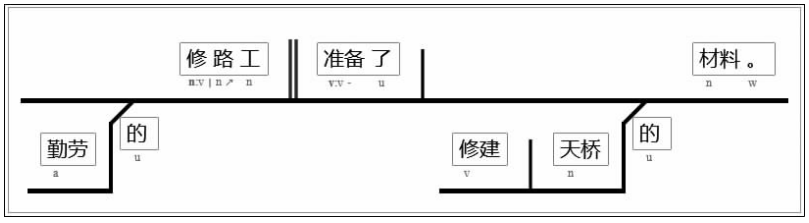

图 3　句式图解结构示例

```
<ju>
  <xj>
    <sbj>
      <att>
        <a >勤劳</a>
        <uu>
          <u >的</u>
        </uu>
      </att>
      <n mod="v↗n↗n">
        <v>修</v>
        <n>路</n>
        <n>工</n>
      </n>                    动态词
    </sbj>
    <prd scp="VO">
      <v mod="v2-u">
        <v>准备</v>
        <u>了</u>
      </v>                    动态词
    </prd>
    <obj>
      <att>
        <prd scp="VO">
          <v>修建</v>
        </prd>
        <obj>
          <n>天桥</n>
        </obj>
        <uu>
          <u >的</u>
        </uu>                 虚词位
      </att>
      <n >材料</n>
      <w>。 </w>
    </obj>
  </xj>
</ju>
```

图 4　XML 结构存储示例

表 1　元素标记集

序号	结构成分	标记	序号	词类	标记
1	整句	ju	1	名词	n
2	小句	xj	2	时间词	t
3	主语	sbj	3	方位词	f
4	谓语	prd	4	数词	m
5	宾语	obj	5	量词	q
6	定语	att	6	代词	r
7	状语	adv	7	动词	v
8	补语	cmp	8	形容词	a
9	独立语	ind	9	副词	d
10	介词位	pp	10	介词	p
11	连词位	cc	11	连词	c
12	助词位 1	uu	12	助词	u
13	助词位 2	un	13	叹词	e
14	助词位 3	uv	14	拟声词	o
15	方位词位	ff	15	标点	w

表 2 属性标记集

元素/属性	取值(说明)
prd/@scp (prd 的辖域)	V(单动) VO(动+宾) VC(动+补) VOO(动+宾+宾) VCO(动+补+宾) VOC(动+宾+补)
cc/@fun (cc 的接续关系)	COO(并列,连 NP) APP(同位,连 NP) SYN(合成式,连 VP) UNI(联合式,连 VP) PVT(兼语式,连 VP) SER(连动式,连 VP)
词类元素/@sen	词语的义项码
词类元素/@mod	词法结构模式

从图、表中可见句式结构特色主要有如下两点,更多的详细信息可参考文献[21]。

(1) 动态词。对词库中未收录,而又不适合进行句子成分切分的一些结构做词法标注,即设置动态词单位。如图 3 中的"修路工""准备了"。

(2) 虚词位。对于不单独充当句子成分的虚词(主要有介词、连词、助词等),在结构中设置"虚词位"结点,如图 4 中"的"标记为"<uu><u>的</u></uu>"。

1.3 对比分析

树库的构建都是在特定的语法理论的框架下制定标注规范的,不同结构类型的树库之间最本质的区别不在于采用了何种标注体系,而在于依照何种语法体系制定的该标注体系。从这个角度上讲,短语结构树库最本质的特征在于其标注体系的制定是站在"短语"的角度,采用了"短语中心"的语法理论,这种语法理论是通过直接描写句子"直接成分"(如主谓、定中、述宾、附加等)的方式分析句子的结构,进而制定标注体系。而句式结构树库是站在"句子"的角度,采用了"句子中心"的语法理论,是通过传统语法中的主语、谓语、宾语等句子成分分析句子的结构,进而制定标注体系的,进一步的论述可以参考文献[23]。

2 转换方法

句式结构树库构建过程中所采用的标注体系为"句子成分分析法",以"句子成分"作为节点;而短语结构体系采用"直接成分分析法",句子成分信息蕴含在"直接成分"节点的"结构标记"中。从短语结构到句式结构主要依据"结构标记"进行转换,而忽略"NP、VP"等"功能标记"。基本思路是,逐层地将句子成分从节点"结构标记"信息中提取出来,用于对应转换规则的确定。而对于一般的叶子节点则直接进行转换,即将"词/词性"直接转换为"<词性>词</词性>"。对于标点符号,可以看作这里的"词",而词性统一使用"w"。

在具体转换规则的制定过程中,将要转换的对象分成两类。一类是只涉及两种体系下所采用的标注体系不同,而不涉及体系本质的不同。对于这一类,只需按照两种体系的对应关系,制定对应的转换规则,直接按照对应转换规则进行转换即可。另一类是由于两种不同的标注体系之间某些部分具有完全不同的本质区别,因而需要进行特殊结构的单独处理,进行必要的人工干预。

2.1 转换规则

在短语结构的标注体系中,大部分节点是二分结构,也有少量是多分结构,对于不同的情况应区别对待。此处是否"二分"的判断,仅从实义节点考虑,暂不计标点符号、连词、助词等形式节点的影响。

2.1.1 二分结构

短语结构标注体系中的二分结构主要有缺省(XX)、主谓(ZW)、述宾(PO)、述补(SB)、定中(DZ)、状中(ZZ)、连谓(LW)、介宾(JB)、方位(FW)等九种结构形式,各节点的左右子树分别以【LP】、【RP】表示。其转换举例如表 3 所示。

表 3 二分结构的转换规则

序号	标记	结构关系	短语结构形式	转 换 规 则
1	XX	缺省	[zj-XX【LP】【RP】]	<ju><xj>【LP】【RP】</xj></ju>
2	ZW	主谓	[xp-ZW【LP】【RP】]	<sbj>【LP】</sbj>【RP】
3	PO	述宾	[vp-PO【LP】【RP】]	【LP】<obj>【RP】</obj>

续表

序号	标记	结构关系	短语结构形式	转 换 规 则
4	SB	述补	[vp-SB【LP】【RP】]	【LP】<cmp>【RP】</cmp>
5	DZ	定中	[np-DZ【LP】【RP】]	<att>【LP】</att>【RP】
6	ZZ	状中	[vp-ZZ【LP】【RP】]	<adv>【LP】</adv>【RP】
7	LW	连谓	[vp-LW【LP】【RP】]	【LP】<cc fun="SER" />【RP】
8	JB	介宾	[pp-JB【LP】【RP】]	<pp>【LP】</pp>【RP】
9	FW	方位	[xp-FW【LP】【RP】]	【LP】<ff>【RP】</ff>

2.1.2　多分结构

短语结构标注体系中的多分结构主要有联合(LH)、兼语(JY)、框式(KS)等三种结构形式。转换举例如表 4 所示。

表 4　非二分结构的转换规则

序号	标记	结构关系	短语结构形式	转换规则
1	LH	联合	[np-LH【NP1】【NP2】连词/c【NPn】]	【NP1】<cc fun="COO" />【NP2】<cc fun="COO"><c>连词</c></cc>【NPn】
			[vp-LH【VP1】【VP2】连词/c【VPn】]	【VP1】<cc fun="COO" />【VP2】<cc fun="UNI"><c>连词</c></cc>【VPn】
2	JY	兼语	[vp-JY【VP1】【NP】【VP2】]	【VP1】<obj>【NP】</obj><cc fun="PVT" />【VP2】
3	KS	框式（主要有 3 类）	[pp-KS 介词/p【NP】助词/u] 如：对我来说	<pp> 介词 </pp>【NP】<un><u></u></un>
			[pp-KS 介词/p【NP】方位词/f] 如：除此之外	<pp> </pp>【NP】<ff><f>方位词</f></ff>
			[vp-KS 是/v【VP】的/u] 如：是分不开的	<prd><v>是</v></prd><cc fun="SYN"/>【VP】<uv><u>的</u></uv>

2.1.3　词法结构转换规则

正如朱德熙先生所说："句法研究的是句子的内部构造，以词为基本单位；词法研究的是词的内部构造，以语素为基本单位。可见句法和词法是两个平面的东西。"[24] 句式结构树库中的动态词结构来源有二：一是汉语中的构形，二是句法构词。

根据葛本仪先生的研究[25]，汉语中构形分为附加式和重叠式两类。附加式构形主要是：名词加词尾"们"表示多数，动词加词尾"着""了""过"表示进行态、完成态和经历态。重叠式构形主要有："VV""V 了 V""V 一 V""V 不 V"等，分别对应短语结构中的附加结构(AD)和重叠结构(CD)，转换规则如表 5 所示。

句式结构中定义的句法构词种类很多，常见的如"数词－量词"构成的数量词结构、"单音名词＋方位词"构成的处所名词、动结式动词、动趋式动词，以及图 3 中的"修路工"等，并且句法构词与短语结构之间的对应关系相对复杂，转换时具有一定的歧义性，详见 2.2 节。

表 5　词法结构转换规则

序号	标记	结构关系	短语结构形式	转换规则
1	AD	附加	[np-AD 名词/n 们/k] 如：同志们	<nmod="n-u"><n>名词</n><u>们</u></n>
			[vp-AD 动词/n 了/u] 如：吃了	<v mod="v-u"><v>动词</v><u>了</u></v>

续表

序号	标记	结构关系	短语结构形式	转换规则
2	CD	重叠	[vp-CD 动词/v 动词/v] 如：研究研究	<v mod="v·v"><v>动词</v><v>动词</v></v>
			[vp-CD 动词/v 了/u 动词/v] 如：看了看	<v mod="v·u·v"><v>动词</v><u>了</u><v>动词</v></v>
			[vp-CD 动词/v 一/m 动词/v] 如：看一看	<v mod="v·m·v"><v>动词</v><m>一</m><v>动词</v></v>
			[vp-CD 动词/v 不/d 动词/v] 如：看不看	<v mod="v·d·v"><v>动词</v><d>不</d><v>动词</v></v>

2.2 特殊结构转换处理

在由短语结构向句式结构进行转换的过程中，除了按照如上所述的对应转换规则进行转换之外，由于两种体系结构之间的差异及汉语语法及句式的复杂性、灵活性，在实际的转换过程中，会出现转换的歧义现象以及某些特定情形的不可预期性。

在短语结构体系中关系标记区分了各种复句类型，而句本位语法体系着重于对小句的分析，所以对于一般的复句结构，简单地转换为若干小句即可。例如，“财政是一个历史范畴，它随着国家的产生而产生。”其短语结构字符串为：“[zj-XX [fj-LS [dj-ZW 财政/n [vp-PO 是/vC [np-DZ [mp-DZ 一/m 个/qN][np-DZ 历史/n 范畴/n]]]],/,[dj-ZW 它/rN [vp-ZZ[pp-JB 随着/p [np-DZ 国家/n 的/u 产生/vN]][vp-XX 而/c 产生/v]]]]。/。]”而在句式结构语法体系中是将其分为“财政是一个历史范畴，”“它随着国家的产生而产生。”两个单句来进行处理的。故从这个角度而言，从短语结构向句式结构的转换无法做到转换的完全对应。

2.2.1 紧缩复句

需要注意的是，短语结构体系中的“紧缩复句”在句式结构体系中分析为“联合谓语”句。紧缩复句一般也为二分结构，其转换规则为：

[fj-JS【LP】【RP】] → 【LP】<cc fun="UNI"/>【RP】

2.2.2 含能愿动词的状中结构

在短语结构中，“能愿动词＋VP”的组合归为状中结构，例如，[vp-ZZ 能够/vM 演化/v]。而句式结构语法中能愿动词称为“助动词”，其与 VP 的组合按“合成谓语”分析。因此，修正 ZZ 结构的转换规则为：

当【LP】为：“助动词/vM”时，[vp-ZZ【LP】【RP】] → <prd><v>助动词</v></prd><cc fun="SYN"/>【RP】

2.2.3 连谓结构

一般而言，短语结构中的“连谓结构”（LW）主要对应句式结构中的“连动句”结构。但句式结构的“连动句”定义更为严格，要求前后 VP 之间：

① 无关联词语；

② 为序列关系。

因此，表 3 中的[vp-LW【LP】【RP】]的转换规则需考虑以上两种例外情形，例如：

① 在形式上多**顺应**中国戏曲及文明戏**以适应**观众的欣赏趣味。

② 在雷达发明之前，**利用**脉冲无线电装置**测量**电离层高度的工作已进行多年。

按照句本位语法，①应转为联合谓语句，可以通过判断【RP】中连词或关联副词的存在来识别；②应转为状中结构，其转换规则需要判断两个 VP 之间的语义关系，而并无具体的形式标记可资利用。这可视为转换中需要消解的结构歧义之一。

2.2.4 动态词情形

分析短语结构树库和句式结构树库的语料标注情况，可以看出在词语层面分歧较多，主要表现在对“动态词”切分粒度的不同。句式结构树库中“动态词”是指一般词库中没有收录，而在句法分析时又不适宜做进一步句子成分切分的造句单位。动态词范围非常广泛，除专有名词、惯用语外还包括大量的临时性句法构词，如全校、桌椅、张老师、家里、看清、举起、每天、五六年等[26]。

句式结构树库中定义的句法构词种类很多，常见的如“数词＋量词”构成的数量词结构、“单音名词＋方位词”构成的处所名词、动结式动词、动趋式动词等。为此，文献[19]专门构建了动态词结构模式知识库，以辅助句式结构析句时的动态词识别。部分常用的动态词结构模式如图 5 所示。

id	mod	example	pos	xml_mod	repath	char_num
1	a↗a↗n	小白兔	n	a↗a↗n	(Null)	3
2	n↗a↗n2	邓副主席	n	n↗a↗n2	(Null)	4
3	v·了—·v	看了一看	v	v·u·m·v	(.)了一\1	4
4	v←来·v←去	飞来飞去	v	v←v·v←v	(.)来\1去	4
5	r2-q	多少个，若干次	r	r2-q	(多少|若干).	3
6	a↗n↗n	黄眉怪，大字报	n	a↗n↗n	(Null)	3
7	m-a-q	一小块，一整箱	m	m-a-q	.+[大小整满长].	3
8	n2↗v2↗n	群众接待站	n	n2↗v2↗n	(Null)	5
9	v | n↗n	安家费，拔秧机	n	v | n↗n	(Null)	3
10	r||v...r||v	你争我夺	v	r||v...r||v	(Null)	4
11	v·不·v	看不看	v	v·d·v	(.)不\1	3
12	n2-之-n	华山之巅	n	n2-u-n	..之.	4
13	v-诸-v2	付诸实践	v	v-p-v2	.诸..	4
14	v-而-d→v	失而复得，笑而不语	v	v-c-d→v	.而..	4
15	v←上·v←下	跑上跑下	v	v←v·v←v	(.)上\1下	4

图 5　动态词结构模式知识库示例

动态词在短语结构树库中又分为两种情形：

① 是直接作为单词标记为叶子节点；

② 是按短语结构分析。

在句式结构中，动态词则直接进行词法分析(见图 3 和图 4)。在转换过程中，①的情形因为没有对应的内部结构信息，故直接转换，留待后续人工分析；②则需要根据短语结构类型及其内部成分的音节数、语素是否自由及语义整合程度等约束条件进行综合判断。具体可参考文献[27]中所构建的动态词结构模式知识库的应用。

3　工程实现

从短语结构树库向句式结构树库的转换包括两个方面：一是在两种不同的语法结构体系下词性标记集的转换；二是两种不同句法结构体系下对应结构层次的转换。

3.1　转换算法

3.1.1　数据预处理

句式结构体系中词性标记粒度比短语结构更粗，只设置了 15 个大词类，转换时一般取短语结构体系中词性标记的第一个字母即可。如短语结构中的词性 vN、rN、qC、nS、dN、aD 等，在句式结构体系下对应的词性分别为 v、r、q、n、d、a 等。特殊情况做相应映射即可。短语结构树库中的标点符号是用其自身标记的，句式结构树库中标点符号统一转为“w”。采用这种转换映射处理方式，使得词性信息粒度变粗了，但并不会丢失词性的大类信息，而句本位语法体系对词类的划分不要求太细，故可以满足后继应用的需要。

3.1.2　算法

结合清华短语结构树库存储结构信息，算法 1 给出了短语结构向句式结构的转换方法的算法描述。

算法 1：短语结构向句式结构的转换方法

输入：短语结构字符串

输出：句式结构的 XML 数据

算法流程：

(1) 针对输入的短语结构形式的字符串，进行数据的预处理操作。将短语结构字符串中的词性标记符号和标点标记符号转换为句式结构对应的词性标记符号和标点标记符号。

(2) 小句获取。如果是复句，则先将其切分成小句；如果是单句，则可以直接对所输入的数据进行解析，构造短语结构树。

(3) 从短语结构树的根节点出发，逐层扫描短语结构树。针对扫描到的当前节点，判断其是否为叶子节点(不计词语节点，视词性节点为叶子节点)。

① 如果当前节点不是叶子节点，首先判断其是否满足句法处理的要求，若满足，则将节点的结构标记与句法结构转换规则中的结构标记进行匹配并进行对应转换；若不满足，则结合动态词模式库按照词法转换规则进行对应转换。

② 如果当前节点是叶子节点，则继续判别其父节点的功能标记是否为 VP。若其父节点的功能标记为 VP，则直接将该叶节点转换为谓语成分，其转换得到的句式结构的形式为：“＜prd＞＜v＞词语＜/v＞＜/prd＞”。若其父节点的功能标记不是

VP,则直接根据词性转换规则转换为:“<词性>词</词性>”。

(4) 生成句式结构的 XML 文件。

算法 1 对应的流程如图 6 所示。

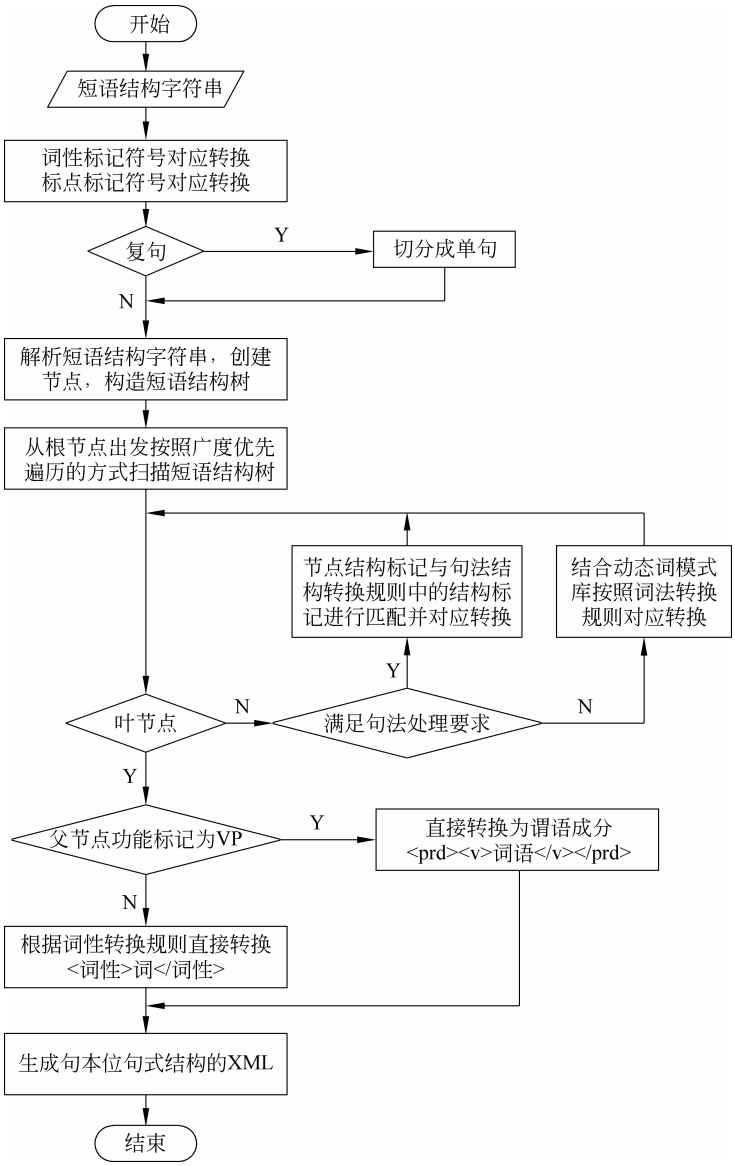

图 6 转换算法流程图

以“美国 T. A. 爱迪生发明了白炽灯。”一句为例,其短语结构字符串为“[zj-XX [dj-ZW [np-DZ 美国/nS T. A. 爱迪生/nP] [vp-PO [vp-AD 发明/v 了/u] 白炽灯/n]] 。/。]”。由于该句是单句,所以无需再切分。接着由预处理过的短语结构字符串构造类似于图 1 的短语结构树,结果如图 7 所示。经过算法 1 各步执行之后,最后生成句式结构的 XML 文件,如图 8 所示。

3.2 可视化展示平台

为了更加形象地对语料转换前后的结构进行对比,搭建了一套可扩展的可视化平台,用于不同句法结构语料的可视化查看。图 9 为两种结构下的可视化展示界面。在图 9 所示的系统中,不仅能够可视化查看不同的句法结构,而且可以对转换后的语料是否正确进行校对,后期将陆续完善相关功能,将短语结构向句式结构、依存结构向句式结构的转换集成其中。

4 实验与分析

在测试过程中,经过对转换结果的初步分析,我们发现句子的长度对转换正确率有着较大的影响。在对文献[28]中关于“清华汉语树库”语料句子长度

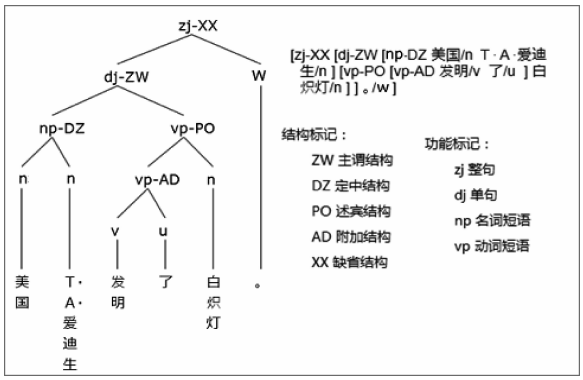

图 7　数据预处理后生成的短语结构树

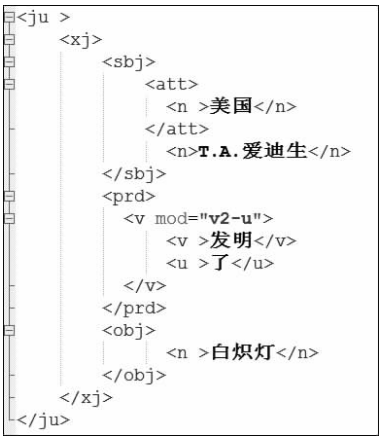

图 8　转换后的 XML 数据

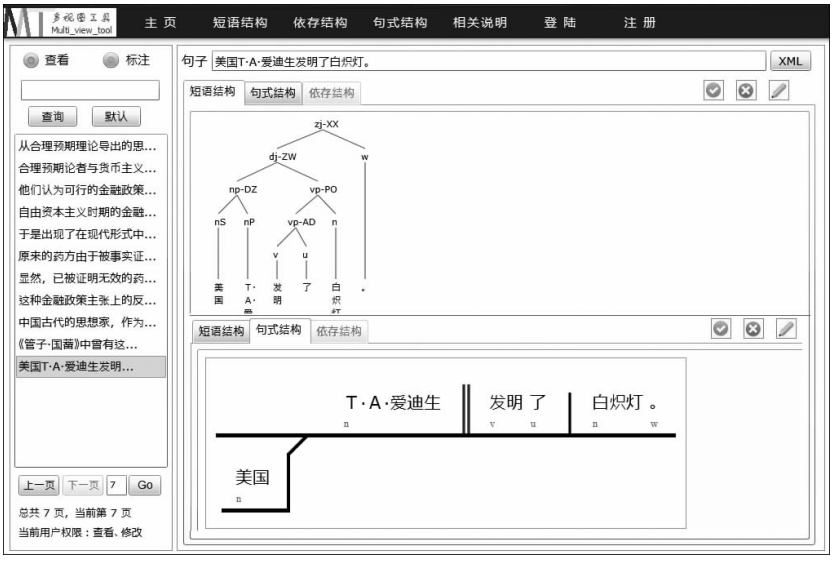

图 9　多视图可视化界面

分析的基础上，综合考虑了语料库中全部句子的平均句长、简单句的平均句长及复杂句的平均句长等因素，我们选取了句子长度为 20 个汉字和句子长度为 30 个汉字作为断点，对不同区间的转换正确率进行了统计。对清华短语结构树库中的 1 000 条文本进行了转换正确率的测试，通过对转换结果的校对统计，发现其中转换正确的句子有 929 句，总体正确率为 92.9%。表 6 给出了不同区间的句子长度对应的转换统计情况。

表 6　转换正确率

序号	句子长度	句子数目	转换正确的数目	正确率/%
1	[0,20)	542	538	99.261
2	[20,30)	287	262	91.289
3	[30,—)	171	129	75.438
总计	—	1 000	929	92.900

由表 6 可以看出，当句子长度在 30 个汉字以下时，转换的正确率都在 90%以上。而当句子长度在 30 个汉字以上时，转换的正确率则明显地降低。在 1 000 句被测试的句子中，30 个汉字以下的句子有 829 句，所以整体的转换正确率还是比较理想的。实验结果表明：①所设计的从短语结构向句式结构转换的算法是切实可行的；②转换结果的总体正确率为 92.9%，对于不是太长的句子而言，转换结果的正确率都在 90%以上；③最终转换结果的正确率与句子的长度密切相关。如果有些句子比较复杂，大多句子的长度达到了 30 字以上，其正确率会有明显的下降。

通过对转换不准确的语料进行分析，可以看出转换不准确的原因主要有两个方面，一是原始标注语料不一致。例如，在短语结构语料中“专家学者”有的标注为“[np-DZ 专家/n　学者/n]”，有的标

注为“[np-LH 专家/n 学者/n]”,“小白菜”有的标注为“小白菜/n”,有的标注为“[np-DZ 小/a 白菜/n]” 等。二是动态词模式库中所收集的结构模式有限。由于动态词的结构模式繁多,动态词知识库中动态词的结构模式不可能穷尽收集所有的结构模式。因此,在由短语结构向句式结构进行转换的过程中,该部分内容仅仅靠程序自动地进行转换则无法达到完全的一致,这更多地依赖于动态词知识库应用过程中的不断完善。

5 结语

本文从短语结构和句式结构的区别与联系入手,设计了一种将短语结构自动转换为句式结构的算法,实现了从短语结构向句式结构的自动转换。为句式树库的构建提供了一种由已有的短语结构树库通过自动转换的方式快速构建树库的方法。并以清华短语结构树库(TCT)为测试语料,实现了将大规模短语结构语料向句式结构语料的转换。

另外,本文在设计了从短语结构向句式结构自动转换算法的基础之上,还搭建了一套可扩展的可视化系统,用于不同句法结构语料的可视化查看。通过句法结构体系分析的可视化系统,我们可以方便地比较从短语结构向句式结构转换的正确情况,研究两种体系下的语料规律。下一步的工作主要是从转换不准确的语料入手,特别是对于较长的句子,找出影响转换正确性的因素,进一步提高转换算法的精度,同时向系统中添加依存结构向句式结构的转换模块,实现一套多视图的汉语树库自动转换系统。

参考文献

[1] 王跃龙,姬东鸿. 汉语树库综述[J]. 当代语言学,2009,(01):47-55,94.

[2] 梁欣,臧德滋. 自然语言句法分析器自动构造系统[C]. 全国计算机语言系联合学术会议,1993.

[3] 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,2005,19(03):21-27.

[4] 李正华,车万翔,刘挺,等. 短语结构树库向依存结构树库转化研究[J]. 中文信息学报,2008,22(6):14-19.

[5] 邱立坤. 多视图汉语树库构建的理论研究与实践[R]. 北京:北京大学博士后研究工作报告,2012.

[6] 邱立坤,金澎,王厚峰. 基于依存语法构建多视图汉语树库[J]. 中文信息学报,2015,29:9-15.

[7] 周惠巍,黄德根. 短语结构到依存结构树库转换研究[J]. 大连理工大学学报,2010(04):609-613.

[8] Lin D. A dependency-based method for evaluating broad-coverage parsers[C]//Proceedings of IJCAI. Montreal, Quebec, Canada, 1995: 97-114.

[9] Fei Xia, Martha Palmer. Converting dependency structures to phrase structures[C]//Proceedings of the Human Language Technology Conference(HLT). San Diego, CA, 2001: 1-5.

[10] Hiroyasu Yamada, Yuji Matsumoto. Statistical dependency analysis with support vector machines[C]//Proceedings of 8th International Workshop on Parsing Technologies, 2003: 195-206.

[11] Joakim Nivre, Mario Scholz. Deterministic dependency parsing of English text[C]//Proceedings of COLING,2004.

[12] Tylman Ule, Sandra Kübler: From phrase structure to dependencies, and Back[C]//Proceedings of the International Conference on Linguistic Evidence, Tübingen, Germany, January , 2004.

[13] 周强. 汉语句法树库标注体系[J]. 中文信息学报,2004,18(04):1-8.

[14] 黎锦熙. 新著国语文法[M]. 北京:商务印书馆,1992.

[15] 廖序东. 论句本位语法[J]. 北京师范大学学报,1990,(02):7-14.

[16] 黄昌宁,李玉梅. 从树库的实践看句本位和中心词分析法的生命力[J]. 北京师范大学学报(社会科学版),2010,(5):53-58.

[17] Jing He, Weiming Peng, Jihua Song, et al. Annotation schema for contemporary Chinese based on JinXi Li's grammar system[C]//Proceedings of the 14th Chinese Lexical Semantics Workshop(CLSW2013), LNAI, Volume 8229, Springer, 2013: 668-681.

[18] 彭炜明,宋继华,王宁. 基于句式结构的汉语图解析句法设计[J]. 计算机工程与应用,2014,06:11-18.

[19] 彭炜明,宋继华,俞士汶. 中文信息处理的词法问题——以句本位语法图解树库构建为背景[J]. 中文信息学报,2014,28(02):1-7.

[20] 彭炜明. 析句图解法及其信息化[J]. 暨南学报(哲学社会科学版),2014,36(7):106-112.

[21] 彭炜明. 句本位语法树库构建及其在对外汉语教学中的应用[R]. 北京:北京大学博士后研究工作报告,2014.

[22] 杨天心,彭炜明,宋继华. 基于句式结构的高效语法图解标注系统[J]. 中文信息学报,2014,28(04):43-49,67.

[23] 何静,彭炜明,宋继华. 汉语句式结构的数字化——句本位语法与“图解法”改造[J]. 北京师范大学学报(自然科学版),2016,(04):413-419.

[24] 朱德熙. 语法讲义[M]. 北京:商务印书馆,1999.

[25] 葛本仪. 汉语词汇研究[M]. 北京:外语教学与研究

出版社，2006.
[26] 郭冬冬. 句本位树库构建中的动态词及其结构模式分析 [D]. 北京：北京师范大学硕士学位论文，2016.
[27] Dongdong Guo，Shuqin Zhu，etc. Construction of the dynamic word structural mode knowledge base for the international Chinese teaching [C]//Proceedings of the 16th Chinese Lexical Semantics Workshop (CLSW2016)，2016：251-260.
[28] 王东波，谢靖. 基于清华汉语树库的有标记联合结构统计分析[J]. 现代图书情报技术，2010(04)：12-17.

张引兵(1979—)，博士研究生，主要研究领域为中文信息处理。
E-mail：zhangyinbing@mail. bnu. edu. cn

彭炜明(1985—)，通信作者，博士，讲师，主要研究领域为中文信息处理、词汇语义学。
E-mail：pengweiming@bnu. edu. cn

宋继华(1963—)，教授，博士生导师，主要研究领域为语言信息处理、计算机教育应用。
E-mail：songjh@bnu. edu. cn

第十五届全国自然语言处理青年学者研讨会在南京成功举行

2018 年 5 月 4 日至 5 日，第十五届全国自然语言处理青年学者研讨会(YSSNLP 2018 会议)在南京召开。本次研讨会由中国中文信息学会主办，计算机软件新技术国家重点实验室(南京大学)承办。

本次研讨会的主题为“关注学科交叉，增进产学交流”，旨在促进自然语言处理领域国内外学者间的学术互动，加强学术研究和产业发展的交流对话，共同促进整个自然语言处理领域的进步。本次研讨会由南京大学戴新宇、黄书剑担任组织主席，清华大学刘知远担任程序主席。

开幕式上，南京大学计算机科学与技术系副主任高阳教授、中国中文信息学会副理事长孙乐研究员、南京大学自然语言处理教研室主任陈家骏教授分别致辞。开幕式由南京大学自然语言处理教研室戴新宇教授主持。

研讨会邀请了京东 AI 研究院常务副院长何晓冬博士和中国科学院自动化所研究员余山博士做特邀报告。何晓冬探讨了深度学习对自然语言理解的驱动作用，介绍了如何让 AI 通过 NLP 技术理解人类以及如何让 AI 的结果能被人类理解接受两方面的最新研究进展。余山报告了语言的脑内表征及有效交流的神经机制，介绍了近年来脑科学在理解人类信息交流的神经机制方面的进展。

研讨会设置了前沿论坛、NSFC 基金项目论坛、产业论坛、新会员论坛四场专题论坛。在这些论坛中，与会的国内自然语言处理领域的青年学者，对自然语言处理领域的前沿技术、不同场景下的应用、以及目前所面临的机遇与挑战等方面展开热烈深入的讨论。

本次研讨会上还召开了学会青工委全委会，第二届执委会主任、清华大学刘洋老师做本届工作报告：在第二届执委任内青工委从 76 人扩充为 130 人，委员组成兼顾了年龄、性别、校企等方面的平衡，较好地代表了国内自然语言处理青年学者群体；青工委持续开展了全国自然语言处理青年学者研讨会、顶会会议论文预讲会、学术沙龙、学术交流、学术专栏、论文访谈间等特色学术活动，在国内 NLP 学者具有较大影响力；青工委还致力于完善组织条例，增进内部文化建设，积极参与学会和各专委服务工作，获得广泛好评。本次全委会选举产生了第三届青工委执委会委员，他们是(按姓名拼音排序)车万翔、黄书剑、贾珈、兰艳艳、廖祥文、刘康、刘知远、邱锡鹏、汤步洲、肖桐、张家俊。执委会委员推选中国科学院自动化所刘康担任主任，哈尔滨工业大学车万翔、清华大学刘知远、东北大学肖桐担任副主任。