

# 图解句式结构体系及其树库构建

彭炜明<sup>1,2,\*</sup>, 赵敏<sup>3</sup>, 宋昱辰<sup>4</sup>, 胡佳佳<sup>1</sup>, 宋天宝<sup>5</sup>, 穗志方<sup>6</sup>, 宋继华<sup>1</sup>

<sup>1</sup> 北京师范大学汉字汉语研究与社会应用实验室, 中国北京, 100875

<sup>2</sup> 宾夕法尼亚大学语言数据联盟 (LDC), 美国费城, 19104-2653

<sup>3</sup> 北京汉雅天诚教育科技有限公司, 中国北京, 100084

<sup>4</sup> 密歇根大学语言学系, 美国安娜堡, 48109-1220

<sup>5</sup> 北京工商大学计算机学院, 中国北京, 100048

<sup>6</sup> 北京大学计算语言学教育部重点实验室, 中国北京, 100871

pengweiming@bnu.edu.cn, zhaom@han-sky.com, songyuch@umich.edu

hjj81@126.com, songtianbao@btbu.edu.cn, szf@pku.edu.cn, songjh@bnu.edu.cn

## 摘要

句式结构是一种基于句本位语法的形式化句法结构, 采用自定义的图解形式呈现句子结构。本文提出了涵盖小句结构、词法结构和句间结构三方面的句式结构体系, 阐明了其设计理念以及句本位的析句原则, 最后概述了基于该体系构建汉语树库的工程进展情况。

关键词: 句式结构; 句本位语法; 图解法; 汉语树库

## Diagrammatic Sentence Pattern Structure System and Its Treebank Construction

Weiming Peng<sup>1,2,\*</sup>, Min Zhao<sup>3</sup>, Yuchen Song<sup>4</sup>, Jiajia Hu<sup>1</sup>, Tianbao Song<sup>5</sup>,  
Zhifang Sui<sup>6</sup>, Jihua Song<sup>1</sup>

<sup>1</sup> Laboratory for Chinese Character and Language Research and Social Application,  
Beijing Normal University, Beijing, 100875, China

<sup>2</sup> Linguistic Data Consortium, University of Pennsylvania, Philadelphia, PA, 19104-2653, USA

<sup>3</sup> Beijing Han-sky Education Technology Co., Ltd, Beijing, 100084, China

<sup>4</sup> Department of Linguistics, University of Michigan, Ann Arbor, MI, 48109-1220, USA

<sup>5</sup> School of Computer Science and Engineering,  
Beijing Technology and Business University, Beijing 100048, China

<sup>6</sup> Key Laboratory of Computational Linguistics,  
Ministry of Education, Peking University, Beijing; 100871, China

pengweiming@bnu.edu.cn, zhaom@han-sky.com, songyuch@umich.edu

hjj81@126.com, songtianbao@btbu.edu.cn, szf@pku.edu.cn, songjh@bnu.edu.cn

## Abstract

The Sentence Pattern Structure (SPS) is a formalized syntactic structure based on sentence-based grammar, which presents the structure of sentences in a customized diagram form. This paper proposes an SPS system covering three aspects: clause structure, lexical structure, and inter-sentence structure, elucidating its design philosophy and parsing principles of sentence-based grammar. Finally, an overview of the engineering progress of building a Chinese treebank based on this system is provided.

**Keywords:** Sentence Pattern Structure; Sentence-based Grammar; Diagrammatic Analyzing Method; Chinese Treebank

\*通信作者: 彭炜明 (pengweiming@bnu.edu.cn)

基金项目: 国家自然科学基金 (62007004); 北京市自然科学基金 (423408); 国际中文教育研究课题 (23YH35C)。

©2024 中国计算语言学大会

根据《Creative Commons Attribution 4.0 International License》许可出版

## 1 引言

句式结构体系是基于黎锦熙《新著国语文法》所创“句本位”语法理论而设计的一种形式化句法结构体系。由于它采用自定义的“图解法”来表现句子结构，故也可称作图解句式结构。图解法的数字化改造工作最早可追溯至（彭炜明，2012），其目的是将曾经流行一时的汉语语法教学工具付诸计算机实现，以推动汉语语法教学和中文信息处理两个领域的信息沟通。其后，我们实现了数个版本的图解方案和可视化工具，构建过规模不等的句式结构树库（也叫句本位语法树库）。今年是第一个以“句式结构”（Sentence Pattern Structure）命名的图解方案公布的10周年（彭炜明等，2014），由于不同时期的设计不尽相同，因此有必要做一个阶段性总结，同时发布最新版的句式结构体系及其树库语料。

## 2 相关工作

汉语的树库构建起步于20世纪90年代，数十年间建成了一批影响较大的中文树库。先是以宾州中文树库（CTB）（Xue N., 2005）、清华汉语树库（TCT）（周强，2004）为代表的短语结构树库，其后是以哈尔滨工业大学中文依存树库（HIT-CDT）（Che W., 2012）、苏州大学汉语依存树库（CODT）（郭丽娟，2019）为代表的依存结构树库。这两种结构各有侧重，前者突出短语功能和结构层次，后者强调中心词和结构关系，它们是目前中文信息处理领域的主流句法结构。北京大学的多视图中文树库（PMT）（邱立坤，2015）尝试融合二者优点，实现了两种结构的自动切换。

由于主流树库大多是以计算机自动句法分析为背景，其结构体系是在英语成熟的句法结构体系基础上改造而成的，在面对汉语文本时表现出一些不适应性。近年来，汉语树库研究出现了如下一些新的动向：

- **句法标注向浅层分析转向。**即避免繁琐标注过程，不追求以词为单位的分析粒度，而改以组块为单位，以“呈现句子骨干结构”为目标，例如北京语言大学的汉语组块依存树库（钱青青，2022）。
- **突破句法约束直接标注语义结构。**应对汉语中大量存在的非投影结构，采用更为自由的图（而非树）来表示句子的语义结构。例如哈尔滨工业大学的语义依存图库（Wang Y., 2016）、南京师范大学的中文抽象语义表示语料库（CAMR）（李斌，2017）。
- **分析单位从句子延伸至段落乃至篇章。**针对汉语多流水句和“意合”的语法特点，树库建设开始关注语句衔接和篇章结构问题。例如苏州大学的微观和宏观篇章结构语料库（奚雪峰，2017；褚晓敏，2020）和北京语言大学的小句复合体语料库（宋柔，2017；2022）。
- **面向语言教学回归传统教学语法。**为适应面向人的语言教学需求，采用传统教学语法描写句子结构。例如北京师范大学的黎氏语法树库（He J., 2013）、句本位图解树库（Song T., 2016），以及北京语言大学自动构建的句式结构树库（谢晨晖，2023）。

以上有些是相互关联的，比如，北语组块依存树库进一步发展为基于篇章的句法结构树库（卢露，2022）、意合图表示（荀恩东，2023），就体现了前三个方面。而（4）其实就是前三个动向的综合，传统教学语法在突显句子结构骨架，处理句法语义的结构冲突，以及促进段落篇章理解方面有着同样的诉求。

本文所谓的传统教学语法即以“句子成分”分析（而非“直接成分”分析或层次分析）为基础，以归纳句型句式为主要目标的汉语语法体系。《暂拟汉语教学语法系统》（张志公，1956）及此前的多数汉语语法体系大都可归入这一类，最典型的即黎锦熙“句本位”语法。由于黎氏语法的句子成分划分与今天通行的教学语法存在较大差异，因而后来的句式结构体系设计（Peng W., 2015）便以《暂拟系统》作为主要参照。

## 3 句式结构体系

句式结构分析首先由程序根据截句标点<sup>0</sup>（。|？|！）自动切分出“整句”，整句之下再分出若干小句。小句内部的成分结构以二维平面上自定义的图解样式呈现，各小句的图解样式再以不同虚线相连，表示各种句间结构。词语散布在图解样式上，标注词类、义项和词法结构信息。

<sup>0</sup>截句标点的范围也可视文本情况扩充到冒号、分号，以及一部分的省略号、破折号。

总之，一个完整的句式结构分成小句结构、词法结构和句间结构三个层面。图解操作在图形界面上进行，由程序完成图解样式和数据格式（XML）间的编、解码转换。

### 3.1 小句结构

句式结构共设主、谓、宾、定、状、补和独立语七类句子成分，虚词不单独充当句子成分而占据特定“虚词位”。如图 1 所示，句子成分沿一条长横线平铺排列：主、谓、宾为主干成分，置于横线上方；定、状、补、独立语为附加成分，附于横线下方。主干维持“主-谓-宾”单谓核核心（简称单谓核）格局的句式为“基础句式”；打破这一格局的为“复杂句式”，其中联合谓语<sup>1</sup>、连动谓语、兼语谓语和合成谓语<sup>2</sup>为双谓核结构，主谓谓语结构则以支架顶起，形成一个新的句式层次。

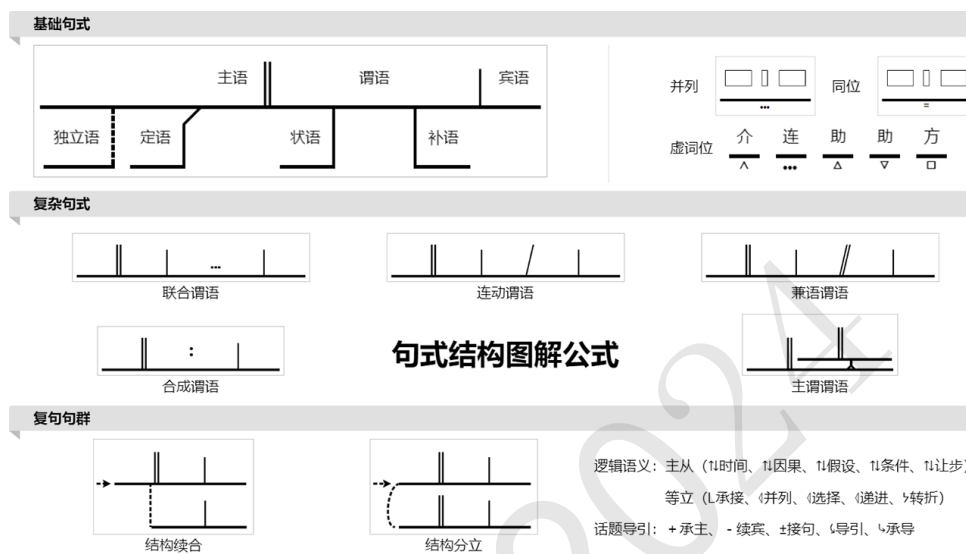


Figure 1: 句式结构图解公式

虚词位置于其所连接或附加的成分旁边，分四类：介词位、连词位、方位词位和助词位，前三类分别标示“^”、“...”、“□”于横线下方，助词位又分三种：①连接定、状、补语的助词“的/之”、“地”、“得”，置于折线与横线的连接处；②与 NP 或单词组合的助词，标示“△”于横线下方；③与 VP 或小句组合的助词，标示“▽”于横线下方。图解公式中各横线段均可进行横向扩展，但遵循如下约束规则：

- (1) 主、宾语为体词性横线段（简称体词段），谓语为谓词性横线段（简称谓词段）<sup>3</sup>；
- (2) 体词段只能扩展出定语、介词位、方位词位、△助词位、并列结构、同位结构；
- (3) 谓词段只能扩展出状语、补语、宾语、主语、▽助词位、连词位、双谓核结构；<sup>4</sup>
- (4) 附加成分横线段初始不限定体/谓性质，当其进行结构扩展时，再根据 (2) 和 (3) 确定；
- (5) 超出 (2) 和 (3) 限制的扩展，如主谓谓语、VP 主语等，需在原横线段上用支架顶起以造成新的结构层次；

句式结构有其独特的层次观。若只对主干成分进行规则 (2) 和 (3) 范围内的扩展，只会引起主干横线的延长，并不增加句式层次。增加句式层次的情况如图 2 所示：

<sup>1</sup>联合谓语：两个 VP 按复句逻辑紧缩而成，例如：母亲求神拜佛。他越想越生气。

<sup>2</sup>合成谓语：助动词/系动词+VP，例如：谁能帮助我？他昨天是从香港回来的。

<sup>3</sup>句本位语法不设“名词谓语句”，一般所谓“名词谓语”分析为缺省谓语“是”所带的判断宾语。

<sup>4</sup>许多从谓词段扩展出的成分并不与其直接相连，如主前状语、宾后补语，但应视为从谓词段而出。

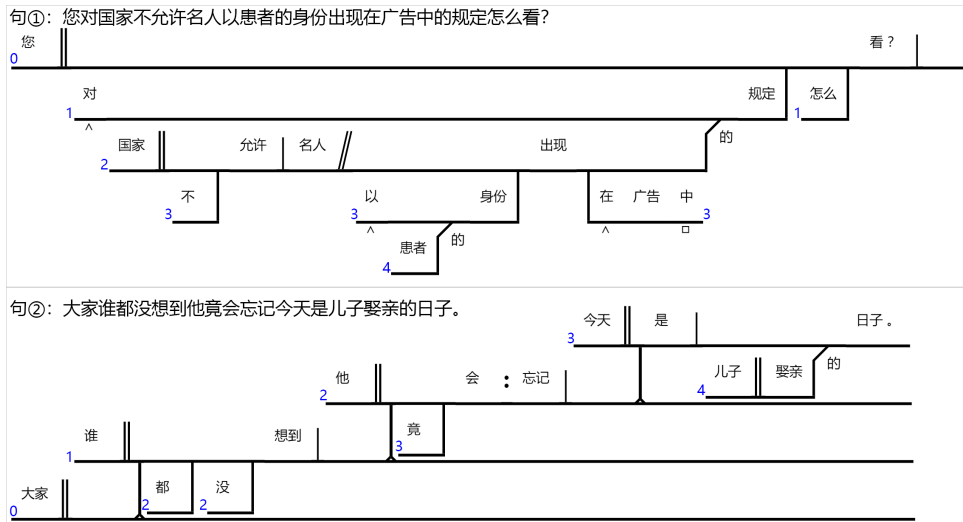


Figure 2: 句式结构的层次深度示例

### 3.2 词法结构

成分分析一般到词一级结束，但由于汉语词和短语边界不清，图解时就以一个基础词库为界，凡词库中收录的称作“词库词”，未收录而又不宜进一步句法切分的单位统称“动态词”。词库词直接标注词类和义项；对动态词，先从内部拆分出词素（不必拆至最小的语素，以能找到词库义项为界），分别标注词素类、义项、词法结构关系（如表 1）以及动态词的整体词类。

符号	关系	示例
...	并列	桌...椅；中...小学
↗	定中	鸡 ↗ 蛋；小 ↗ 白 ↗ 兔；群众 ↗ 接待 ↗ 站
→	状中	极 → 具；深 → 感；代 → 写；改 → 用
←	述补	赶 ← 跑；看 ← 清；拿 ← 下；举 ← 起
	动宾	驯   兽 □ 师
	主谓	你    争...我    夺
•	重叠	看 • 看；研究 • 研究；睡 • 睡觉；看 • 一 • 看；看 • 了 • 看；看 • 不 • 看
-	其它词法	桌-上；两-只；一-大-碗；看-了；同学-们；学习-者；拿-得-起；翩然-而-至
◇	超词法	十-元 ◇ 八-角 ◇ 六-分；江西 ↗ 省 ◇ 宁都 ↗ 县 ◇ 固村 ↗ 镇

Table 1: 词法结构关系

词法标注示例如图 3，XML 中为动态词整体设一个词类节点，词素类作为其孩子节点，词素类节点上设属性@sen 记录义项，词类节点上设属性@mod 记录如下 3 方面信息构成的词法结构模式：词素类、词素字数（为 1 时省略）和词法结构关系。

### 3.3 句间结构

篇章中的句子并非孤立存在，而是通过指称、结构衔接、逻辑连接等手段组织起来的。汉语语法具有“意合”特点，小句之间的结构衔接表现出显著的“流水句”特性。参考宋柔提出的广义话题结构及其堆栈模型，句式结构在处理小句的结构衔接时也引入缩进对齐的描写机制，在图解样式上分为两类（参见图 1）：

- **结构续合**：后续小句共享前面小句的句首成分序列，图解样式上缩进后续小句，并以虚直线相连，对齐至共享部分的末端，相当于（宋柔, 2013）中的堆栈模型。蒙后共享的情况即（宋柔, 2013）中的逆向堆栈模型，也归此类，但因数量较少，本文阐述均以承前共享为准。
- **结构分立**：不满足上述续合关系的其它衔接方式，图解以虚弧线相接两个小句主干横线的最左端。其中，后一小句必无缩进；前一小句若缩进，则以虚线延伸其主干横线至最左端。

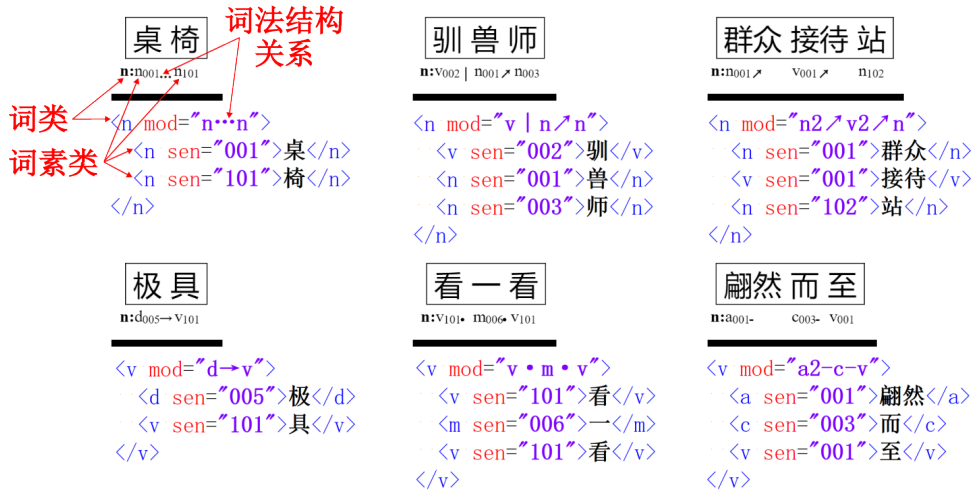


Figure 3: 动态词的词法结构及 XML 格式

例如图 4。XML 中，小句结构和词法结构编码在 xj 节点之下（限于篇幅此处省略，具体参见 (Peng W., 2015)），句间结构通过 xj 的属性 @upt (up to 的缩写，意为“向上对齐至”) 表明。取值格式为“m:n”，其中 m 表示相连 xj 的 id，n 表示接续至该 xj 的第 n 个词之后，若“结构分立”则取值“s”。结构续合仅限于同一整句内部，而分立小句则可能位于不同整句，此时 @upt 中加上表示整句 id 偏移量的 k 值（一般取负值），格式为“k:m:n”（详后图 5）。

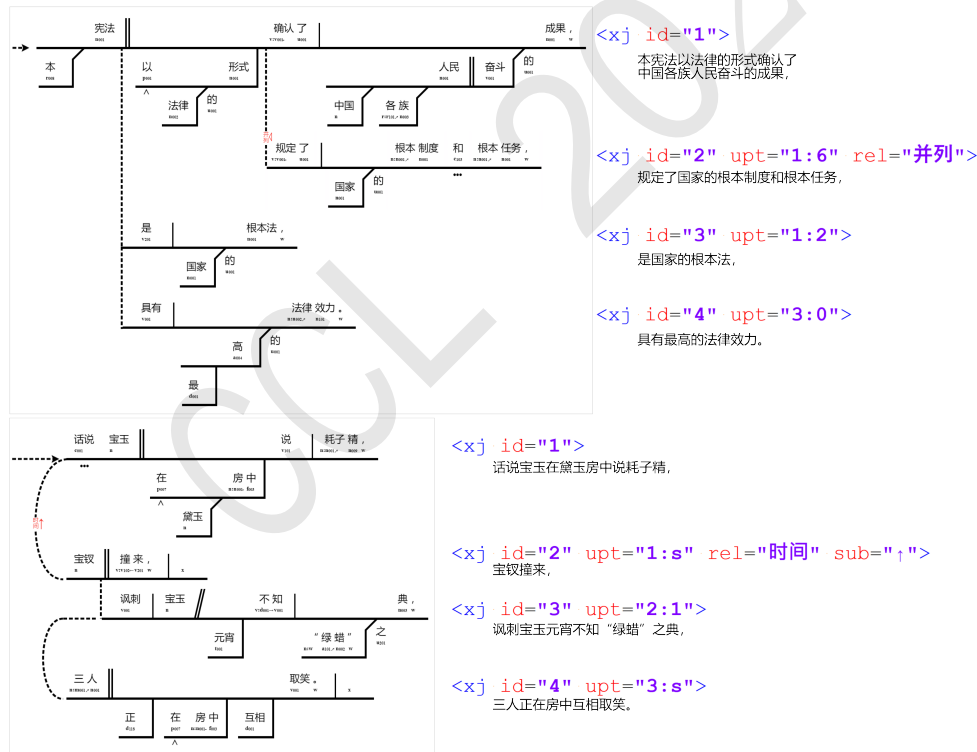


Figure 4: 句间结构的续合和分立

相连小句之间若存在逻辑语义关联，标注于虚线上，在 XML 中记录于 xj/@rel 属性。参照 (黎锦熙, 1962) 复句分类体系，句式结构中逻辑语义关系共分两大类、十小类：

- 等立：承接、并列、选择、递进、转折；
- 主从：时间、因果、假设、条件、让步。

主从复句中，由于从句一方前可后，故在关系旁加标“↑”或“↓”，以示从句所在位置（对应 XML 属性@sub）。仿此，等立关系旁也增加一个标示符，粗分为“承接 L、并列 4、选择 4、递进 4、转折 4”。由于“承接”可视为小句相连的缺省关系，故若无形式特征（关联词等），可以不显式标注。

小句间的“结构续合”是语法上经济原则的作用，即通过共享，后面小句避免了重复前面小句中出现的句首成分序列。然而，小句之间的共享信息并不限于此，比如（宋柔, 2013）中的“新支句/节栈模型”，即后面小句以前面小句的非开头部分作“话题”。这种情况均以“结构分立”的虚弧线图解，但加标“话题”关系：

- 承主（+）：后小句话题取自前小句的主语，或从更前小句继承来的话题。
- 续宾（-）：后小句话题取自前小句的宾语，包括状、补语中的介词宾语。
- 接句（±）：前小句整体作为后小句的话题。

此外，在涉及引语（包括直接引语和间接引语）的篇章中，引语后面的小句经常跨越引语部分（有时很长，涉及多个小句）而从前面的导语小句中延续话题。为此，再引入两类“话题导引”关系：

- 导引（∩）：前、后小句分别为导语、引语，当引语涉及多个小句时，只标与导语相邻的小句。
- 承导（∪）：后小句话题取自前小句对应导语小句的话题。

图解样式中，“导语”关系的虚弧设为有向弧，而其它 4 类则加标记于虚弧线旁，对应 XML 编码设在 xj/@tpc 属性中，例如图 5 所示。

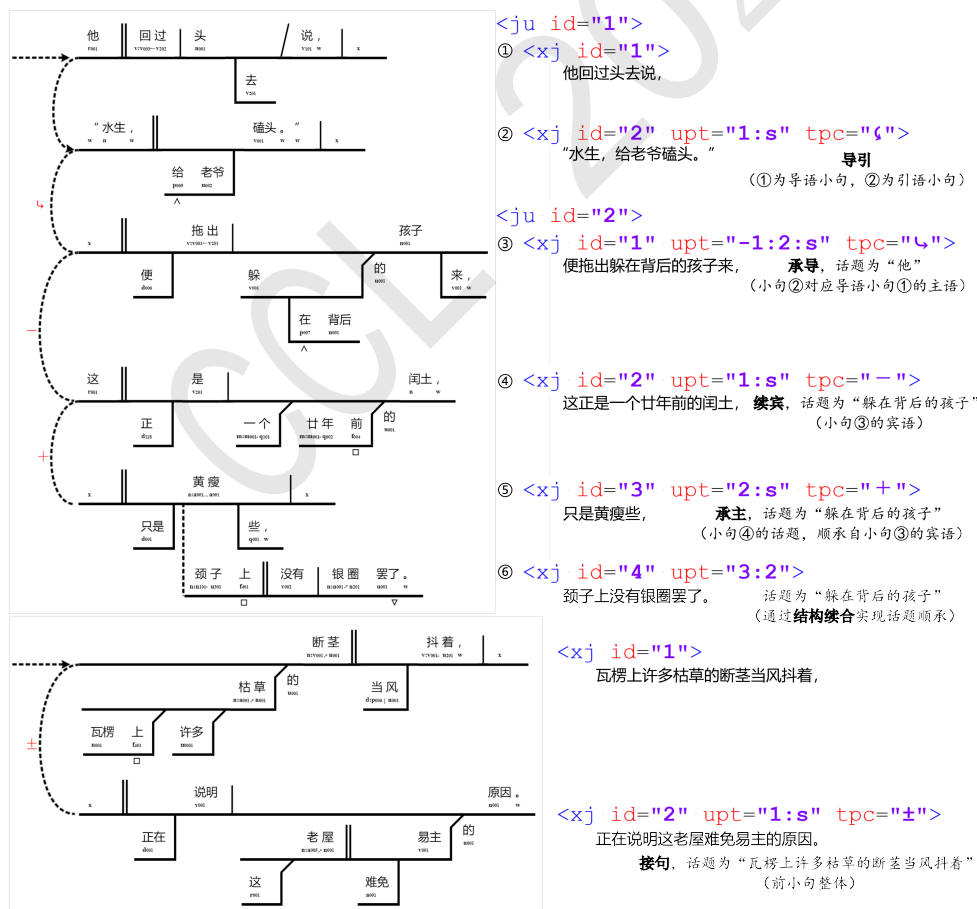


Figure 5: “话题导引”示例

## 4 设计理念和析句原则

句本位语法是一种教学语法，与强调分析程序严密性的结构主义语法相对。将图解法改造为句式结构体系，始终需要面对的一个问题就是兼顾其教学性质和信息化需求。从结构主义的角度看，句本位语法有许多不利于信息处理的特点，有些恰恰是其教学性质的突出反映，有着心理认知基础和教学实用目的。本节将从坚持教学性和实现信息化这一对矛盾的平衡处理入手，阐述句式结构体系的设计理念，并探讨该体系下句本位的析句原则。

### 4.1 直观呈现，提高空间利用

句式结构采用自定义的图解样式表达句子结构，直观性是其区别于现有形式化句法结构的最显著特征。以短语结构和依存结构为背景，句式结构的直观性来自于：

- 充分利用二维平面的空间性，不同层次深度的词在横、纵两个维度展开排布，通过横、竖、虚、实等线条实现层次对齐；而主流树库局限于“树”的数据结构，也受制于原文词序列的线性排列，空间利用率都不如句式结构。
- 以长横线分隔主干成分和附加成分，实现语法分析“抓主干”的实用目的。语法教学中最基本的“句子成分”在主流树库中隐藏在二元的层级结构或句法关系中，而在句式结构中它们是直接存在的实体，因此，获取句子主干和成分信息在句式结构中更为便捷。
- 图形样式和数据存储分离，由程序实现编、解码，标注人员只面对图形而不接触数据，避免了记忆标记集符号的负担。

从空间利用率角度出发，也就不难理解句式结构的如下设计：

- 缩进对齐手段的使用仅限于严格的结构续合，而没有如广义话题结构那样扩展至所有的流水模型。因为句式结构是完全的结构分析，若将“话题导引”关系也一律采用缩进形式，则会造成大片的空白区域。
- 在“话题导引”中，虚弧相接遵从就近原则。例如图 5 中，本来小句③可以从①中直接“承主”，小句⑤可以从③中直接“续宾”，但如此就会形成跨小句的虚弧线。“承导”关系的引入就是为了避免当引语小句较多时导语和引语后面小句之间的超长虚弧。

### 4.2 廓清层面，突显句式层次

句式结构的句式层次有别于结构主义语法的短语层次。直观看，短语层次倾向于二分，而句式层次则更加扁平化。事实上，句式结构层次观的独特性还首先表现在将不同层面的结构层次从句式层次中分离出来，具体为：

- 引入动态词，将词法结构的内部层次从句法层面剥离。由于动态词的词法组合与句法有着较高的一致性，主流树库中通常不做严格区分，进而如重叠、数量、词缀结构等词法结构也进入句法层面。在二分为主的层次结构中，这并不构成太大问题；但句式结构需要突显句式层次，讲究“抓主干”。如不设词法结构这一层，则介于词和短语之间的大量动态词结构将混入句法层，势必造成层次嵌套过于细碎，不利于突显主干和句式。
- 小句之间的结构续合和话题导引也具有一定的层次性，本质上都可归结为跨小句的广义话题统摄关系。它们在句式结构的描写机制中都没有采用类似短语结构的层次嵌套，就是为了使小句内部的结构分析能够维持在成分组合（包括虚词位）这一级别。

廓清了上述两个不同层面的层次，便能计算句式结构的分析层次。在单一句式层次的前提下，句子成分根据图解公式遵从特定的组合模式（彭炜明，2021），比如，不带补语的单谓核单宾语句式为：

- 连? 状\* 定\* 主 连? 状\* 谓 定\* 宾 助?（其中“?、\*”为正则元字符）

已知特定成分序列，如“定-定-主-状-状-谓-定-宾”，便可推定其短语结构层次：

- [[ 定-[ 定-主 ]]-[[ 状-[ 状-谓 ]]-[ 定-宾 ]]]

因此，句式结构不对短语层次做显式的描写，一些短语层次的结构歧义也不进行消歧。例如“他老板和哥哥是同学”的析句结果仅为“(他)老板..和.哥哥||是|同学”，其中定语辖域存在歧义。这确是句式结构层次的不足，但换个角度看，该句本身确实存有歧义，即所谓“真歧义”，因而不消歧也不失为一种存疑的做法。

此外，话题导引中“接句”和“导引”两类不按“结构续合”分析（如将后小句缩进对齐至前小句的最末端），除了空间利用率的原因外，也有结构层次的考虑：这两类结构中前后小句的成分序列不在同一句式层次。

### 4.3 依句辨品，建立正则规则

“依句辨品”，即根据词语充当句子成分的类型来辨别词类，是句本位语法的词类主张。在句式层次约束下，从句式到成分可以构造有限的正则规则；那么，“依句辨品”就是进一步构造从成分到词这一级规则，进而建立起完整析句规则系统的要求。

句式结构秉承句本位语法的词类观，力求在成分和词类之间维持一个相对稳定的对应关系：主、宾语对应体词，谓语对应谓词。现实中的矛盾来自于词库词类与语料标注词类的不一致，最典型的就是主、宾语中动词的名物化问题。我们主张从词库词到句中词之间有一个转类过程，当单个谓词进入图解中的体词段时，便转为名词。

根据我们的标注经验，词库（以《现汉》为例）的词类信息未能涵盖真实语料中句法功能的情况不在少数。并且对于同一类型的转类，有的词条为两类单独设置义项，有的则仅设一类，不尽统一。因此，拘定标注语料中的词类都来自词库的做法也不现实。

转类看似多一道手续，但收益大于麻烦，具有现实可行性。一来，主、宾语对应体词、谓语对应谓词的情况占多数，少数情况执行转类就可建立起完整的析句规则系统。二来，除名物化外，还有其它诸如“名↔量”、“动↔介↔连”等的转类，均有规律可循，黎氏语法中有一套系统的转类理论（黎锦熙，1959）可以借鉴。

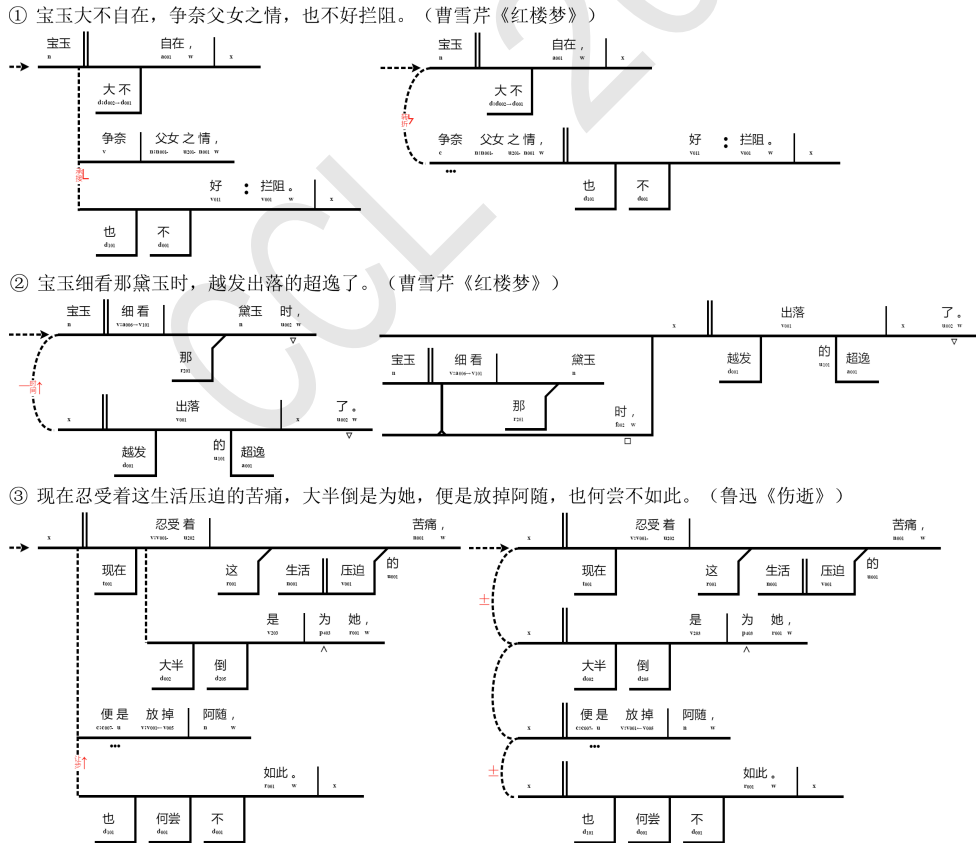


Figure 6: “两可”的结构分析



#### 4.4 服务章法，活看句法结构

句本位语法图解的最终目标是走到篇章结构分析，也就是说，词法分析、句法分析都只是手段，都是为“章法”服务的。句本位语法著述（黎锦熙,1959;1962）经常提到“活看”，即同一句子作“两可”的结构分析，根据上下文语境有不同取舍。试举几例，如图6所示：

例①左图断作三小句，以“续合”相接；右图断作两个“分立”小句，形成“转折”的逻辑关联。例②左图分析为“时间”关系主从复句，右图则分析为带时间状语的单句。句中“争奈”、“时”的词类认定也随句法结构而变。例③左右断句相同，两种分析源于对第二、四小句主语/话题的不同认定，左图以人（省略的“我”）为话题，右图以事（前一小句）为话题，两种分析逻辑上都能自洽。取舍需看更大范围的上下文，着眼于汉语表达的“简明、连贯、得体”，这就是所谓“章法”了。

### 5 树库工程

图解句式结构提出至今，我们已开发出多个版本的语料标注工具（杨天心, 2014; 赵敏, 2014; Guan, 2020）。由于体系结构一直处于升级调整状态中，树库语料的发布滞后于实际的构建进程。截至目前，共标注了 30 余万句规模的现代汉语树库和 2.5 万句规模的古汉语树库。北京师范大学语言与文字资源研究中心、汉字汉语实验室、北京大学计算语言学研究所（ICL）、北京汉雅天诚教育科技有限公司、宾夕法尼亚大学语言数据联盟（LDC）等单位先后参与构建。表??为目前树库的语料分布情况。最新版包含 5000 余句古今篇章语料的树库已提交至 LDC 发布，后续语料也将以篇章或书籍为单位陆续发布。

语言	类别	文本来源	规模（句）	进度（√: 完成; Δ: 进行中）		
				小句	词法	句间
白话	二语教材	《长城汉语》等国际汉语教材	140,833	√	√	Δ
白话/文言	语文教材	高中语文教材	12,015	√	√	√
白话	文学作品	《毛泽东选集》 《鲁迅选集》 《红楼梦》 《四世同堂》等	174,953	√	√	Δ
白话	学生作文	中学生考场作文	5,590	√	√	√
文言	经典古籍	《左传》《论语》《孟子》	25,388	√	√	Δ

Table 2: 树库语料分布

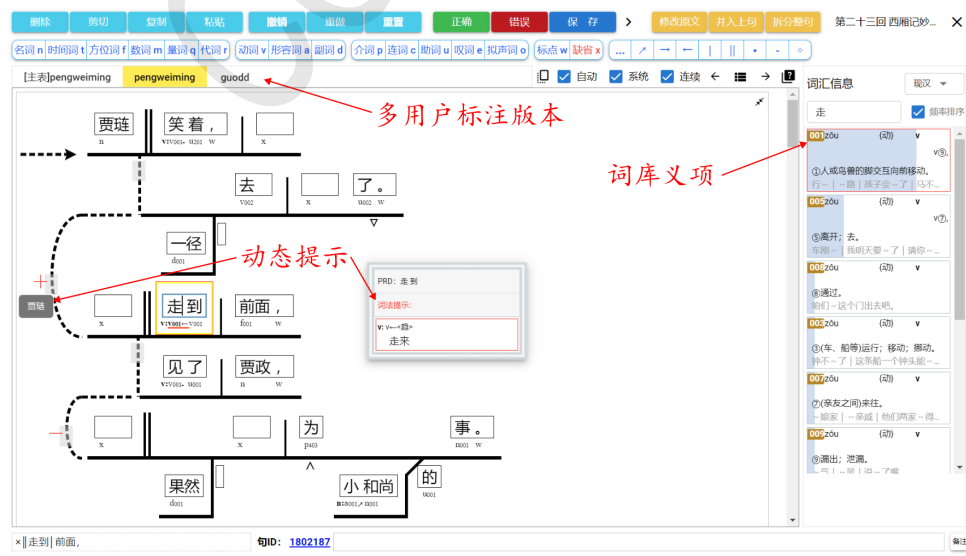


Figure 7: 图解标注系统界面

最新版的图解标注系统地址为：<http://www.jubenwei.com/>，标注界面如图 7 所示。

图解系统背后集成了基础词库、补充词库、词法结构模式库、句式结构规则库、关联词语模式库等一系列知识库，可以动态提示词法、句法结构以及话题导引信息，因而树库标注的效率持续提升。新版树库在增加句间结构信息的情况下，平均标注效率仍能达到 350 句/天的较高水平，略高于 (Guan, 2020) 的 345 句/天。

由于图解结构相对直观的层次性以及由此而来“抓主干”效果，句式结构的正误判断一目了然。权衡标注成本和质量两方面因素，我们没有选择众包、多人盲标的试验模式来验证一致性，而是采用一标多审的质量控制机制：一位标注员首先标注至副表（个人版本），再经多人共同审核、校改后存入主表。审校人员由本团队内部相对稳定的专职人员充当，分为主审和陪审两种角色。由经验丰富的主审员主持审校，从多位陪审员中选择一人负责操作，其他人同步观看并提出不同意见。陪审员只有取得和主审员超过 80% 的整句一致率（即小句、词法、句间完全一致）时才可成为主审员。当遇到“两可”标注意见时，记录后由专家仲裁。疫情期间，我们还组织专家和审校人员对部分经典篇章进行过小规模线上集体标注。如果“两可”标注也计入“一致”的范围，则审校人员在小句、词法、句间这三方面的平均一致率分别可达 92%、95%、84%。

此外，由于我们的语料取材多来自教材和文学作品，并且标注、审校不只是单一枯燥的句法操作，而是综合了词汇、语法、章法的语篇学习过程，因而用户体验要优于一般的树库标注。这也是多人共审和集体标注能够持续开展的原因之一。

## 6 结语

句子结构的图形表示早在上世纪 80 年代的析句方法大讨论中就有人做过对比分析，吕冀平 (1984)、廖序东 (1986) 都指出图解法的优点，“它清晰，醒目，犹如工业上的制图，在讨论句子结构时，双方只要把自己的理解用图解表示出来，就可以省去许多语言。如果要对很长的句子进行分析，那末这种图解法几乎成为 唯一能够使用 的手段，其他各种图示都不免黯然失色。”可惜的是，这种教学上的实用工具在后来很长时间内未能实现信息化，以至错失了在中文信息处理中发挥应有作用的机会。在人工智能和大语言模型盛行的当下，句法分析在计算语言学研究中的关键核心地位受到很大冲击，能否与面向人的语言教学结合起来或是一条值得探索的出路。本文研究正是在这一方向上做出的积极尝试，我们清楚，目前的体系设计中多有迁就传统教学语法，而在信息处理特别是自动分析方面考虑不周的地方，诚恳希望得到有关专家的指正。

## 参考文献

- Che W., Li Z., Liu T. 2019. Chinese dependency treebank 1.0 (LDC-2012T05) [DB/OL]. Philadelphia: Linguistic Data Consortium, <https://catalog.ldc.upenn.edu/LDC2012T05>.
- Guan S., Peng W., Song J., et al. 2020. The Construction of Interactive Environment for Sentence Pattern Structure Based Treebank Annotation. Proceedings of 20th Chinese Lexical Semantics Workshop (CLSW2019), Beijing.
- Guo D., Zhu S., Peng W., et al. 2016. Construction of the Dynamic Word Structural Mode Knowledge Base for the International Chinese Teaching. Proceedings of 17th Workshop on Chinese Lexical Semantics (CLSW 2016), Singapore.
- He J., Peng W., Song J., et al. 2013. Annotation schema for contemporary Chinese based on JinXi Li's grammar system. Proceedings of 14th Workshop on Chinese Lexical Semantics (CLSW2013), Zhengzhou.
- Peng W., Song J., Sui Z., et al. 2015. Formal Schema of Diagrammatic Chinese Syntactic Analysis. Proceedings of 16th Workshop on Chinese Lexical Semantics (CLSW2015), Beijing.
- Song T., Peng W., Song J., et al. 2016. The Construction of Sentence-Based Diagrammatic Treebank. Proceedings of 17th Workshop on Chinese Lexical Semantics (CLSW 2016), Singapore.
- Xue N., Xia F., Chiou F., et al. 2005. The Penn Chinese Treebank: Phrase structure annotation of a large corpus. Natural Language Engineering, 11(2):207-238.
- Wang Y., Guo, J., Che, W., et al. 2016. Transition-based Chinese semantic dependency graph parsing. Proceedings of China National Conference on Chinese Computational Linguistics, Yantai, 12-24.
- 褚晓敏, 奚雪峰, 蒋峰等. 2020. 宏观篇章结构表示体系和语料建设. 软件学报, 31(2):321-343.

- 郭丽娟, 彭雪, 李正华等. 2019. 面向多领域多来源文本的汉语依存句法树库构建. 中文信息学报, 33(02): 34-42.
- 胡壮麟. 1994. 语篇的衔接与连贯. 上海: 上海外语教育出版社.
- 李斌, 闻媛, 宋丽等. 2017. 融合概念对齐信息的中文 AMR 语料库的构建. 中文信息学报, 31(6): 93-102.
- 黎锦熙. 2001. 新著国语文法. 北京: 商务印书馆.
- 黎锦熙, 刘世儒. 1959. 汉语语法教材 (第二编 词类和构词法). 北京: 商务印书馆.
- 黎锦熙, 刘世儒. 1962. 汉语语法教材 (第三编 复式句和篇章结构). 北京: 商务印书馆.
- 廖序东. 1986. 论句子的图解——句子结构分析结果的表示法. 汉语学习, (01).
- 卢露, 矫红岩, 李梦等. 2022. 基于篇章的汉语句法结构树库. 自动化学报, 48(12): 2911-2921.
- 吕冀平. 1984. 句法分析和句法教学. 《中国语文》杂志社. 汉语析句方法讨论集. 上海: 上海教育出版社. 出版: 216-235.
- 邱立坤, 金澎, 王厚峰. 2015. 基于依存语法构建多视图汉语树库. 中文信息学报, 29(3):9-15.
- 彭炜明. 2012. 句本位语法数字化平台的建设和应用研究. 北京师范大学博士学位论文.
- 彭炜明. 2021. 句本位语法的中文信息处理理论与实践. 北京: 外语教学与研究出版社.
- 彭炜明, 宋继华, 王宁. 2014. 基于句式结构的汉语图解析句法设计. 计算机工程与应用, 50(6):11-18.
- 宋柔, 葛诗利, 尚英等. 2017. 面向文本信息处理的汉语句子和小句. 中文信息学报, 31(02): 18-24.
- 宋柔. 2022. 小句复合体的语法结构. 北京: 商务印书馆.
- 奚雪峰, 褚晓敏, 孙庆英等. 2017. 汉语篇章微观话题结构建模与语料库构建. 计算机研究与发展, 54(08): 1833-1852.
- 谢晨晖, 胡正升, 杨麟儿等. 2023. 句式结构树库的自动构建研究. 中文信息学报, 37(2): 15-25.
- 荀恩东. 2023. 自然语言结构计算——意合图理论与技术. 北京: 人民邮电出版社.
- 杨天心, 彭炜明, 宋继华. 2014. 基于句式结构的高效语法图解标注系统. 中文信息学报, 28(04):43-49.
- 张志公. 1956. 语法和语法教学——介绍“暂拟汉语教学语法系统”. 北京: 人民教育出版社.
- 赵敏, 彭炜明, 宋继华等. 2014. 面向图解树库的标注工具开发与优化. 中文信息学报, 28(06): 26-33.
- 周强. 2004. 汉语句法树库标注体系. 中文信息学报, 18(04):1-8.