现代汉语句系系统的构建和研究

亢世勇　许小星

鲁东大学中文信息处理研究所 烟台　264025

E-mail: kangsy64@163.com，xuxx_2004@163.com

摘要：基于标注语料库，我们建立了现代汉语句子的句型系统、句模系统和句干系统，并将这三个系统有机地结合在一起，用句型统领句模和句干，构拟出现代汉语的句系系统。并通过将复杂的句型、句模结构解析为较小的简单的结构的组合，研究复杂句模的组合机制和规律。

关键词：语料库, 句型, 句模, 句干, 句系系统

The Construction and Research on the Sentence System in Modern Chinese

Shiyong Kang, Xiaoxing Xu

Institute of Chinese Information Processing, Ludong University　Yantai　264025

E-mail: kangsy64@163.com，xuxx_2004@163.com

Abstract: Based on the labeled corpus, this paper constructs the syntactic sentence pattern system, semantic sentence pattern system and sentence stem system, and combines them together to build the sentence system in modern Chinese by the way of syntactic sentence pattern guiding semantic sentence pattern and sentence stem. At the same time, this paper decomposes complex syntactic sentence pattern structure and semantic sentence pattern structure into the combinations of some simple structures, and employs these decompositions to research the mechanism and regulation of generating complex semantic sentence pattern.

Key word: Corpus, Syntactic sentence pattern, Semantic sentence pattern, Sentence stem, Sentence system

1．研究背景

根据“三个平面”理论，任何具体的句子都是句型、句模和句类的结合体。对一种语言的句子进行全面的调查以后，通过理性的抽象，可以建立该语言的句型系统、句模系统和句类系统，三个系统互相结合、纵横交错就形成一个句系网络系统。范晓先生提出：一旦将某种族语的句系建立起来，“不仅有利于不懂该族语的人们学习该族语，而且也能使懂得该族语的本族人更好地掌握和运用自己的母语，在现代高技术发展的信息社会里，还能促进机器翻译(自动翻译)和人工智能等方面的研究工作”。[1][2]因此范先生积极呼吁学界，希望能共同努力来构建现代汉语的句系系统，建立研究一门新的学科——句系学。

我们在 2005 年承担了国家社科规划项目“基于大规模标注语料库的现代汉语句子语义结构系统研究”，以中小学语文课文和对外汉语阅读材料为语料，共加工了 713430 个字、28669 个句子。以句子为单位标注了每个句子的句法结构和语义结构信息，建立了“现代汉语句法语义信息

语料库”。基于该语料库，分别提取和建立了句型系统、句模系统和句干系统，该系统包括句型5558类、句模13696类，句干14211类。为了进一步研究句型和句模的对应机制，也为了让相互独立的三个系统有机地结合起来，从而建立一个更有价值的研究体系，我们在原有成果的基础上，欲构拟出一个由句型系统、句模系统和句干系统组成的句系系统。尽管我们的语料还暂时缺少对句子的句类信息的标注，与范晓先生所提出的句系网络系统相比还不够全面，但仍可以说是对“句系学”理论的一次大胆的尝试性研究。

2. 句系构建的理论原则

在介绍句系构建的原则之前，先简单地介绍一下语料库标记的设置。我们共设置了24个语义成分标记（施事S、受事O、与事T、客事K、系事X、结果R、致事Z、当事D、领事L、分事F、共事Y、目的G、原因C、数量N、依据W、工具I、基准J、时间H、处所P、范围E、材料M、方式Q、方向A、同源B）和7个句法成分标记(主语S、谓语P、宾语O、状语D、补语C、兼语J、独立语T)。

标注样例：[S语言/n]D[P是/v]V[O人类/n最/d重要/a的/u交际/v工具/n]X。

每个句子用“[ ]”来划分语块，“[”后标记该语块的句法成分，“]”后标记该语块的语义成分。上例句的句型是[S][P][O]，句模是[]D[]V[]X，句干是[S]D[P]V[O]X。

通过对句型、句模、句干三个系统的观察发现，句型的种类远远少于句模和句干的种类，句型和句模存在着一对多的状况，且句型和句模的结合构成句干，所以我们采取以句型为纲，通过寻求句型和句模之间的对应关系来构建句系系统。首先我们对句型系统进行了全面细致地考察和整合。研究发现，复杂句在真实语料中占到绝对多数，但无论多复杂的句型，都可以切分为四个串组“P”“SP”“PO”“SPO”。如[S][P][P][O][P][O][P]可以切分为[S][P]+[P][O]+[P][O]+[P]。所以，可以把“P”“SP”“PO”“SPO”看作构成句型的基本结构，任何复杂句型都是这四类基本结构经过组合后再添加上状语、补语形成的。基于这样的想法，我们按照“P”“SP”“PO”“SPO”把句型系统分为四个子系统。对这四类句型的分类遵循以下原则：

（1）在忽视各类句型中状语、补语标记的前提下对句型进行提取和分类。

（2）提取P类句型时，排除所有含SP、PO、SPO串的句型；在提取SP句型时排除所有含PO、SPO的句型；提取PO句型时排除所有含SP和SPO的句型。

（3）SPO类句型包含了剩余所有句型；

（4）层级性（系统性）。句系系统是有层级的，处在第一层级上的是P、SP、PO、SPO四个子系统，第二个层级是上位句型系统，每一个上位句型都对应着若干下位句型，下位句型是真实文本中的句子结构，上位句型是对真实文本句子结构的再抽象，剥离了构成句型的非核心成分（状语和补语），只保留了构成句型的核心成分（主语、谓语、宾语、兼语）。下位句型系统为第三个层级，第四个层级是下位句型对应的句模及该句型句模结合生成的句干。

真实文本句子的句型和句模是较为复杂的，句型句模的对应机制也是当今语言学研究的重点和难点问题。既然一个复杂的句法结构可以看作是几个简单结构的组合体，那么一个复杂的句模也应该可以切分为较小的单位。我们考虑是否可以通过研究简单句型和复杂句型、简单句模和复杂句模之间的组合映射规律，从而找到句型句模对应机制研究的一个新的突破点。我们按照“P”“SP”“PO”“SPO”对句型系统进行分类，也正是基于这样的考虑。

如上文所述，[P]、[P][O]、[S][P]、[S][P][O]是构成句型的基本结构，我们就把这四类确立为四类子句型系统的基础句型。这四类基础句型不仅是真实语料中最常见的简单句，其对应的句模的种类也是非常多的。句型[P]对应着一类句模[]V；句型[P][O]对应着 17 类句模，共计 1526 例；句型[S][P]对应着 16 类句模，共计 2490 例；句型[S][P][O]对应着 73 类句模，共计 5131 例。随着进一步的深入分析，我们的研究也取得了预想的突破，这也反过来验证了我们按照“P”“SP”“PO”“SPO”对句型系统进行分类的合理性。

3. 复杂句模产生的机制——叠加法

3.1 高频句干和高频句模的确立

句型同句模之间存在一对多的对应关系。同一类句型，它同其所对应的不同类型的句模结合所产生的句干的数目也不尽相同。我们按照下面的公式为每一种基础句型提取出高频句干，取高频句干公式如下：

将一个句型和语义对应的各类句模所构成的所有句干的个数记为数组 $\mathbf{n}$，该数组的长度记为 N，定义如下两个函数，

$$f_1(m)=\begin{cases}0 & if\ \mathbf{n}(m)=\min(\mathbf{n})或\mathbf{n}(m)=\max(\mathbf{n})\\ \mathbf{n}(m) & others\end{cases},\quad f_2(m)=\begin{cases}0 & if\ \mathbf{n}(m)=\min(\mathbf{n})或\mathbf{n}(m)=\max(\mathbf{n})\\ 1 & others\end{cases}$$

其中，$m=1,\Lambda\ ,N$。那么，当第 m 种句干的个数满足下式时就称为高频句干，

$$\mathbf{n}(m)\geq\frac{\sum_{m=1}^{N}f_1(m)}{\sum_{m=1}^{N}f_2(m)}$$

可以与基础句型结合成高频句干的句模我们称之为高频句模。

句型[P][O]与对应的高频句模结合成的高频句干有以下 5 类，这 5 类句干的数目占到总数的 92.73%。

句型	句模	句干	个数
[P][O]	[]V[]O	[P]V[O]O	689
[P][O]	[]V[]K	[P]V[O]K	276
[P][O]	[]V[]U	[P]V[O]U	233
[P][O]	[]V[]X	[P]V[O]X	165
[P][O]	[]V[]P	[P]V[O]P	52

句型[S][P]与高频句模结合成的高频句干有以下 2 类，这 2 类句干的数目占到总数的 99.08%。

句型	句模	句干	个数
[S][P]	[]D[]V	[S]D[P]V	1080
[S][P]	[]S[]V	[S]S[P]V	1074

句型[S][P][O]与高频句模结合成的高频句干有以下 9 类，这 9 类句干的数目占到总数的 90.08%。

句型	句模	句干	个数
[S][P][O]	[]D[]V[]X	[S]D[P]V[O]X	1446
[S][P][O]	[]S[]V[]O	[S]S[P]V[O]O	850
[S][P][O]	[]D[]V[]U	[S]D[P]V[O]U	842
[S][P][O]	[]S[]V[]U	[S]S[P]V[O]U	470
[S][P][O]	[]D[]V[]K	[S]D[P]V[O]K	340
[S][P][O]	[]L[]V[]K	[S]L[P]V[O]K	282
[S][P][O]	[]P[]V[]K	[S]P[P]V[O]K	161
[S][P][O]	[]S[]V[]P	[S]S[P]V[O]P	138
[S][P][O]	[]U[]V[]X	[S]U[P]V[O]X	93

3.2 对复杂句模结构的分析

在确立了基础句型和其对应的高频句模后，着手展开对复杂句模结构的研究。先考察了由基础句型简单叠加而成的新句型，我们把这类新句型称为**典型句型**（典型句型是下位句型中的一类，如 SPO 类句型下的典型句型有[S][P][O][S][P][O]、[S][P][O][S][P][O][S][P][O]等）。在考察时我们使用了**解析法**，解析法是指先对一个句型进行分解，如典型句型[S][P][O][S][P][O]可以分解为[S][P][O]+[S][P][O]，那么[S][P][O][S][P][O]对应的句模也可以相应地分解为两个小句模。通过分解，可以直观地了解基础句型对应的句模的构成情况。

SPO 类句型的典型句型[S][P][O][S][P][O]对应着 161 种语义模式 528 例（528 例指由该句型同 161 种句模相结合构成的 161 种句干的总例句数）。这 528 例句模就可以解析为 1056 个小句模。基础句型[S][P][O]对应的高频句模共计 921 个，占到总量 1056 个的 87.22%。并且以上这些句模的出现的频度高低基本与它们在基础句型[S][P][O]出现的频度高低基本一致。此外还发现由两个相同语义模式叠加构成的句模有 21 类 182 例，占到总数的 34.47%。

我们还考察了其他典型句型如[S][P][O][S][P][O][S][P][O]、[S][P][O][S][P][O][S][P][O][S][P][O]等，在这些句型中由基础句型[S][P][O]对应的高频句模在新句模中的出现率是相当高的，也就是说典型句型的句模基本上由基础句型对应的几类高频句模组合而成的。而且，由同类型高频句模叠加构成新句模的比例也是相对比较高的。为了验证这一结论，我们还考察了 SP 类、PO 类句型中的典型句型的句模情况，均支持以上结论。

通过分析由基础句型叠加组合构成典型句型的构成情况，我们得出以下结论：基础句型对应的高频句模是构成典型句型对应的句模的主体。而且这些高频句模在典型句型对应的句模中的出现率（即出现频度）基本与其在基础句型中出现的频度一致。所以说，由相同句模叠加构成新句模的方法（简称叠加法）是构成典型句型对应的句模的一个非常重要的方法。

我们又用解析法抽查检验了句系系统中除典型句型之外其他句型的情况。

在[S][P][O][P][P][P]句型中，共有句模 7 类 11 例，这 11 例中，S 均是四个 P 的共同主语，我们将这个句型解析为[S][P][O]+[S][P]+[S][P]+[S][P]的组合，观察这四个小句型对应的句模的结构。其中[S][P][O]对应的句模涉及到[]D[]V[]X（有 3 个），[]D[]V[]K（有 1 个），[]L[]V[]K（有 1 个），[]S[]V[]O（有 6 个），这四类句模均是句型[S][P][O]对应的高频句模。句型[S][P] 的语义模式只有[]S[]V 和[]D[]V 两类，与基础句型[S][P]对应的高频句模一致。

经随机取样和分析，均可以验证基础句型所对应的高频句模是构成复杂句模的一个重要基础的判断。尽管汉语句子语义结构模式复杂，多达上万种类型，但动名语义关系主要集中在有限的几种类型。此外，还得知，当一个句型是[S][P]或[S][P][O]与[P]、[P][O]的结合体时，在[S][P][P][O]、[S][P][O][P][P]等这类句型中，处在句首的 S 经常充当后面多个 P 共同的主体性语义成分，这一比例高达 85%以上；在[P][O][S][P]、[P][S][P][O]这类句首为动词的句型中，句首的主体性语义成分由位置在其后的 S 兼任的比例大约在 20%左右。

[J]是一个兼语成分，在前期考察基础句型时，我们把含[J]的句型分化在各类下位句型中。如[P][J][P]是 P 类句型下的含两个 P 的下位句型中的一种。我们把[P][J][P]作为含成分[J]句型的基础句型，单独考察[J]同语义成分之间的映射关系。在句型[P][J][P]中，[J]对应着的高频语义成分组合有以下几类，O1+S2，K1+S2，O1+D2，K1+D2，这四类在该句型对应的句模中占到 86.60%。按照前面的研究我们推论这四种语义成分的组合应该是所有含[J]句型中的 J 对应的语义成分组合中高频组合。我们单独抽取出含[J]的所有句型对应的 2130 类句模对上述结论加以验证，经验证，结论与推论一致。[P][J][P]对应的高频句模在所有包含[J]的句型所对应的句模中的出现率占到 85%以上。

小结：通过验证，我们主要得出以下两点结论。

（1）将汉语复杂的句法结构和语义结构解析为较小结构的组合，基础句型的高频句模在组合构成复杂的语义结构中占到较大的比重。

（2）在句子中兼语成分[J]优先映射为 O1+S2，K1+S2，O1+D2，K1+D2 这几种语义组合。

3.3 补语、状语与语义成分的对应情况

前文对复杂句模的产生机制的考察没有考虑句子结构的非核心成分状语和补语。实际上，从简单句模到复杂句模的生成，补语和状语对应的语义成分是不可或缺的重要因素。下面分别考察补语、状语同语义成分的对应情况。

我们首先对补语位置出现的语义成分进行了单独地考察。从四大类句系中分离出了“PC”、“SPC”、“SPOC”、“POC”和“PCO”五类动补组合，考察补语位置上的语义成分的情况。

在 PC 组合中，C 主要映射为数量成分（N）、时间成分（H）、处所成分（P）和谓词性成分（V），各种成分出现的比例如下：

C⟶P 66.9%
C⟶H 14.9%
C⟶V 7.8%
C⟶N 7%

在 SPC 组合中，C 主要映射为数量成分（N）、时间成分（H）、处所成分（P）、谓词性成分（V），各种成分出现的比例如下：

C⟶P 53.6%
C⟶V 21.43%
C⟶N 8.3%
C⟶H 7.9%

在 SPOC 组合中，C 主要映射为数量成分（N）、时间成分（H）、处所成分（P）、基准成分（J）和谓词性成分（V），各种成分出现的比例如下：

C——▶N 31.30%
C——▶H 27.59%
C——▶J 13.8%
C——▶P 12.1%
C——▶V 8.6%

在POC组合中，C主要映射为数量成分（N）、时间成分（H）、处所成分（P）、谓词性成分（V），各种成分出现的比例如下：

C——▶N 50%
C——▶P 21.4%
C——▶H 14.3%
C——▶V 14.3%

在PCO组合中，C主要映射为数量成分（N）、时间成分（H）、谓词性成分（V），各种成分出现的比例如下：

C——▶V 50%
C——▶N 29%
C——▶H 21%

此外还考察了状语位置上的语义成分的出现率情况。我们从数据库中提取出所有的包含[D][P]、[D][P][O]、[D][S][P]、[D][S][P][O]字段的句型对应的语义模式，考察在这四类中状语同语义成分的对应情况。

[D][P]类中状语位置上语义成分的出现率构成不等式如下：

P>O>H>T>J>Q>A>Y>S>E>W>C>I>D>N>G>M>K>L>F>R

[D][P][O]类中状语位置上语义成分的出现率如下：

H>P>T>O>Q>E>I>W>J>Y>D>C>N>M>S>G>A>L>K

[D][S][P]类中状语位置上语义成分的出现率如下：

H>P>E>W>C>T>J>G>Q>N>K>I>O>M>A

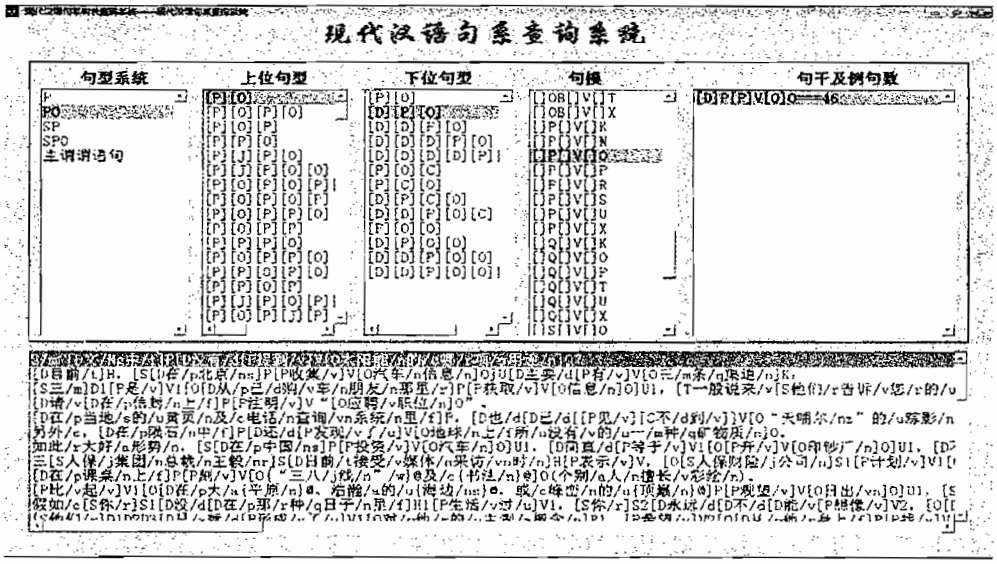
[D][S][P][O]类中状语位置上语义成分的出现率如下：

H>E>P>W>G>C>T>J>Q>N>D>K>I

上述研究中对补语、状语位置出现的语义成分的优先序列的考察及结论是比较粗疏的，具体到实际句子中状语、补语位置对应的究竟是什么语义成分，受到核心动词的价、动词以及动词所控制的名词的语义类，还有句子句式等多种因素的制约。这也是我们进一步研究的方向。

4. 句系系统

我们结合“现代汉语句系查询系统”界面的一个截图，直观地了解句型系统的层级体系和句系系统的概貌。下图中“句型系统”列表框中显示出第一层级的四个子系统；以子系统[P][O]为例，包含[P][O]、[P][O][P][O]、[P][P][O]、[P][O][P]等共计184类上位句型；其中上位句型[P][O]下属[P][O]、[D][P][O]、[D][D][P][O]、[P][O][C]、[P][C][O]等13类下位句型；下位句型之一[D][P][O]对应着78类句模，[]D[]P[]O和句模之一[P][V][O]结合生成的句干[D]P[P]V[O]O在语料库中共有例句46个，例句集显示在界面下端的图框中。



将整个句系系统构建成一个有层级的体系，使得把三个相对独立的句型、句模、句干系统有序整合在一起成为可能，从而形成更有研究价值的语义资源。按照基础句型把庞大的句系系统分离为四个子系统，每个子系统中通过提取真实文本句子句法结构的共性特征（核心句法成分）设立上位句型，从而将句型系统中 5558 类句型纳入 2304 类上位句型体系，为句型和句模对应关系、复杂句模的生成机制等研究铺平了道路。

5. 展望

当今的语言学研究中，语义与句法如何接口的问题，或者说探索语义结构转化/投射为句法结构的理论和方法是语法学家、计算语言学家们十分关心的问题[3]。通过我们的考察研究，了解了现代汉语句子结构系统的复杂性，同时也认识到这种复杂的本质特征中存在着某种规律性。既然对一个完整的句法结构和语义结构的对应关系进行研究较为困难，试将完整的句法结构/语义结构分解为较小的单位寻求组合的规律，未尝不是一个办法。尽管我们对句法结构和语义结构之间的映射机制的研究还处于起步阶段，本文还没有提出一套完整的供计算机自动识别和分析句子语义结构的规则，但我们的研究还会持续下去，并坚信一定会取得较大地突破。

参考文献

[1] 范晓. 略说句系学. 汉语学习, 1999(6).

[2] 范晓. 汉语的句子类型. 上海: 书海出版社, 1998.

[3] 陆俭明. 句法语义接口问题. 外国语, 2006（3）.

[4] 亢世勇, 许小星. 现代汉语句子语义成分标注研究. 中文计算技术与语言问题研究——第七届中文信息处理国际会议论文集. 北京: 电子工业出版社，2007.

[5] 许小星, 亢世勇. 基于标注语料库的现代汉语句法成分和语义成分对应机制研究初探. 中文计算技术与语言问题研究——第七届中文信息处理国际会议论文集. 北京: 电子工业出版社，2007.

[6] 亢世勇. 面向信息处理的现代汉语语法研究. 上海: 上海辞书出版社，2004.