基于标注语料库以[S][P][O]为样本的句系研究

孙道功

（南京师范大学文学院，江苏南京，210097）

摘　要： 句系是语言中句样的总和。句样是在句干基础上选择各类语气类型形成的。文章首先制定句法、语义标注规范，建立标注语料库。基于大规模语料库，以[S][P][O]为典型样本，定量统计考察句型句模的对应关系，建立句干系统。语料库中，句型[S][P][O]共对应了30种句模类型，形成了30种句干。并根据使用频度差异，把句干分成三个层级。在句干的基础上，基于真实文本，考察句干对语气的选择情况，建立[S][P][O]对应的句样层级体系。各种句干对语气类型选择的自由度并不一致，形成的句样类型使用频度差异很大。其中，陈述语气是选择性最高的语气类型。并根据[S][P][O 的形成句系时表现，总结出现代汉语句系的特点。

关键词： 句系；句模；句干；句样；标注语料库

Sentence system studies of [S][P][O] on tagged corpus

Sun Dao-gong

(Chinese Language and Culture School, Nanjing Normal University, Jiangsu,210097;

Abstract: Sentence system is the total sentence samples in a language. Based on sentence stem, sentence samples take shapes of various tone types. The paper firstly makes out norms for syntactic and semantic marking, and sets up a tagged corpus. Based on this large scale tagged corpus, taking the sentence patterns of [S] [P] [O] as the typical samples, it quantitatively investigates the corresponding relation between syntactic and semantic sentence patterns of modern Chinese, and establishes a sentence-stem system. There are 30 types of semantic patterns and 30 sentence stems which correspond to[S] [P] [O] in corpus. According to frequency difference, sentence stems are divided into three levels. Based on sentence stems and real text, how the tone fits the sentence stem is also investigated. Hence the [S] [P] [O] corresponding level systems are set up. For various sentence stems, free degree which the tone fits the sentence stem is indifference, and frequency of sentence samples are also not the same. Statement tone is the easiest tone to be selected when setting up sentence samples. We briefly summarize the characteristics of sentence system of modern Chinese according to [S][P][O] .

Keywords: Sentence system; semantic pattern;sentence stem; sentence sample;tagged corpus

1　前言

范晓 1999 年在《略说句系学》中提出“句系”的思想，句系即句子系统。任何语言都是一个规则系统。但语言中句系是抽象的，是语言中句样的综合，是通过句样体现的。句样是从句法、语义、语用三个平面对句子进行概括的抽象常体。句系的建构，必须立足于三个平面，在句型、句模、句类三个子系统的基础上才能完成。句子

基金项目：国家社科项目“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”（编号 05BYY029）。

作者简介：孙道功（1977—），男，山东日照人，南京师范大学文学院博士生，专业为语言学及应用语言学。

E-mail:sundg9527@tom.com

作为人们日常交流、传递信息的载体，句系的构建有重要的理论意义和应用价值。不仅能反映汉语句子在各个层面上的真实面貌，丰富三个平面的语法理论，深化句法语义关系的研究。同时，可以弥补对外汉语教学中单纯句型教学的不足，提高句型练习的针对性和实用性，提高句子的使用水平、推动句子教学。基于大规模标注语料库建立的句法、语义模型及句样系统，也可以提高机器翻译、句子理解生成的准确度，推动自然语言处理的发展。

2 句系建构的思路

现代汉语句系的构建必须基于句型、句模、句类三个层面。这就要求我们首先要制定句法、语义的标注规范，建立句法、语义标记集。我们的技术路线是根据标注规范对大规模的语料进行句法、语义标注，建立标注语料库。在此基础上研究句型、句模、句类之间的对应关系。具体言之，首先考察句型、句模的对应关系，统计句干（句型句模结合体）的类型、数量，然后在大规模的真实文本中考察句干对句子语气的选择情况，抽象出汉语中句样类型，从而构拟现代汉语句系。在具体的建构中，必须充分考虑到句样的抽象性、层级性、有限性。因为归纳抽象时，句子已经失去了具体的理性意义。同时还要考虑各种类型之间上位和下位的派生关系。从理论上讲，句系应该是一种语言中所有句子的综合体。但从研究现实性和可操作性来看，基于标注语库所抽象出的句样类型应该是有限的，本文句系的建构正是在受限语料基础上的受限分析。仅以汉语中最典型的[S][P][O]类型句子为研究对象，进行句系建构。信息处理用的语法、语义模型，是一种受限的语言符号标记集。可以首先建立一个典型样本子集，然后逐步扩充和完善，从而完成整体的建构。之所以以[S][P][O]为典型样本，因为从语料库实际看，它具有典型性、原型性、底层性。以[S][P][O]为典型样本的分析，基本能够反映整个句系建构的过程和特点。在100万字的语料中，共有15358个句子，单句类型句子共有9519个。其中[S][P][O]句型的句子有1175个，占单句总量的12.34%。

3 句法成分、语义成分、句类的标记及标注方法

3.1 句法成分标记及含义

吸收现有句法研究成果，采用清华大学周强拟定的句法标注体系，确定句法成分的标注对象为主语语块（S）、述语语块（P）、宾语语块（O）、状语语块（D）、补语语块（C）、兼语语块（J）、独立语块（T）七大类。括号内为语义标记符号。

3.2 语义成分标记及含义

语义成分的标注属于深层句法分析。依据美国语言学家菲尔墨的格语法（case grammar）理论，借鉴国内学者在"格"系统方面的研究成果，最终以林杏光提出的6大类22个格为基础，进行适当改造，共分出23个语义角色。分别是施事(S)、当事(D)、 受事(O)、客事(K)、共事(Y)、系事(X)、类别(B)、对象(T)、结果(R)、方式(Q)、数量(N)、范围(E)、时间(H)、领事(L)、分事(F)、基准(J)、工具(I)、材料(M)、处所(P)、方向(A)、依据(W)、原因(C)、目的(G)。括号内为语义标记符号。

3.3 句类标记及含义

根据已有成果，从语气上把句类标记分为四类：陈述语气(L1)、疑问语气（L2)、祈使语气（L3)、感叹语气（L4)，扩号内为句类标记符号。

3.4 标注方法

用中括号标注出句中的句法成分部分（主、谓、宾、状、补)，在"["后文字前标注句法成分标记：S、P、O、D、C等；然后在"]"后直接标注语义成分标记，谓语语块则标明谓词符号"V"。例如： [S 我国选手张军高凌]S1S2 [D 以2：1]Q[P 击败]V1 [O 丹麦选手]T1[P 获得]V2[O 决赛权]R2。

4 基于标注语料库的[S][P][O]句型的句系建构

句系建构必须首先建立句型句模的对应体系即句干系统，然后考察句干对句子语气的选择情况。同一句干可能选择不同的语气类型从而形成了不同的句样。这也正彰显一个句子的生成过程。经过定量统计，句法结构同为[S][P][O]的 1175 个句子共对应了 30 种句模类型，形成了 30 种句干类型。

4.1 [S][P][O]与句模的对应体系考察

在语料库中，[S][P][O]共对应了 30 种句模，分别是：DVX、SVR、SVO、DVK、DVR、DVO、SVP、SVE、LVF、DVE、SVK 、DVP、SVT、DVT 、SVN、HVO、SVG、SVX、DVH 、HVX、DVN、PVO、EVK、DVG、DVA、PVK、SVB、EVX、DVF、PVX、HVD、TVX、DVJ、SVJ。句模与句型相结合形成了 30 类句干，使用频度差异很大。为了考察更加细致，根据频度分成三个层级。频度 100 次以上为第一层级，频度 10—99 为第二层级，频度 1—9 为第三层级。[S][P][O] 对应的句模中，第一层级有 4 种，分别是 DVX、SVR、SVO、DVK，共覆盖句子 1022 个，占本类型句子总量 86.98%。共形成了[S][P][O]/ DVX、[S][P][O]/ SVR、[S][P][O]/ SVO、[S][P][O]/ DVK（"/"前代表句型，其后代表句模，以下亦然）4 种句干，其中[S][P][O]/ DVX 频度最高，共对应了 394 个句子。

第二层级共对应了 DVR、DVO、SVP、SVE 四种句模，共形成了[S][P][O]/ DVR、 [S][P][O]/ DVO、[S][P][O]/ SVP 、[S][P][O]/ SVE 4 种句干类型，仅占[S][P][O]类型句子总量的 13.33%。第三层级的成员最多，共对应了 22 种句模，分别是 LVF、DVE、SVK、DVP、SVT 、DVT 、SVN 、HVO、SVG、DVH 、HVX、DVN、PVO、EVK、DVG、DVA、PVK、SVB、EVX、PVX、DVJ、SVJ，形成了 22 种句干类型，但是仅覆盖 81 个句子，占本类型句子总量的 6.89%，不再列出。根据原型论观点，第一层级成员使用频度高，覆盖句子多。在形成新句干时具有优先选择性，属于原型成员。同时原型成员还具有底层性、派生性的特点，在句干系统中，原型成员是形成其他类型的基础。

4.2 基于标注语料库[S][P][O]句系建构

句干系统的建立仅为句系的构建提供了可能。但二者还有很大区别。句干是句型句模的结合体，就是通常所说的静态句。句干只有经过语用指派后获得语气、语调等语用因素，才能由静态变为动态，成为交际中具体的句子。从理论上讲，任何一个句干都潜在同四种语气类型搭配的可能性。本文是基于标注语料库的真实文本进行考察的，不采用内省法。[S][P][O]所形成的句样层级系统如下：

4.2.1 第一层级

在第一层级共包括 4 个高频句干，分别是[S][P][O]/ DVX、[S][P][O]/ SVR、[S][P][O]/ SVO、[S][P][O]/ DVK。从真实文本来看，各类句干同各种语气类型搭配的倾向性不一样，陈述语气是搭配度最高的语气类型。句干[S][P][O]/ DVX 进行语气选择形成 3 类句样，其中句样[S][P][O]/ DVX/L1 频度最高，为 372 次，[S][P][O]/ DVX/L2 为 20 次，[S][P][O]/ DVX/L4 为 2 次，由于语料规模等限制，未发现同祈使语气搭配的类型。句样[S][P][O]/ DVX/L1 自由度最高。句干[S][P][O]/ SVR 形成了 3 种句样， [S][P][O]/ SVR/ L1 频度为 365 次，[S][P][O]/ SVR/ L2 为 6 次，[S][P][O]/ SVR/ L4 仅有 1 次。句干[S][P][O]/ SVO 形成 4 种句样，[S][P][O]/ SVO/L1 频度为 120 次、[S][P][O]/SVO/L2 为 13 次，[S][P][O]/ SVO/ L3 为 1 次，[S][P][O]/ SVO/ L4 为 13 次。句干[S][P][O]/ DVK 形成 3 种句样，[S][P][O]/ DVK/ L1 频度为 101 次，[S][P][O]/ DVK/ L2 为 3 次，[S][P][O]/ DVK/ L4 为 5 次。

4.2.2 第二层级

第二层级共有四个句干，分别是[S][P][O]/DVR、[S][P][O]/DVO、[S][P][O]/SVP、[S][P][O]/SVE。从语料库的实际考察来看，陈述语气被选择性仍然最高，这与第一层级的表现一致。句干[S][P][O]/DVR 可形成 2 种句样

类型，[S][P][O]/DVR/ L1 使用频度为 26 次；[S][P][O]/DVR/ L2 使用频度为 3 次。[S][P][O]/DVO 可形成 3 种句样，[S][P][O]/DVO/L1 使用频度高，为 14 次，[S][P][O]/DVO/ L2 为 1 次，[S][P][O]/DVO/ L4 为 1 次。[S][P][O]/SVP 可形成 2 种句样，[S][P][O]/SVP/ L1 为 12 次，[S][P][O]/SVP/ L2 为 4 次。[S][P][O]/SVE 仅形成了一种句样类型，[S][P][O]/SVE/ L1 频度为 11 次。

4.2.3 第三层级

第三层级的成员最多，共有 22 个句干，但使用频度低，语气选择的自由度低。除句干[S][P][O]/DVE 外，其他都仅形成一种句样类型，其频度与句干是一致的，仅指出句样模型。句干[S][P][O]/DVE 形成 [S][P][O]/DVE/L1、[S][P][O]/DVE/L2 两种。[S][P][O]/LVF 形成了[S][P][O]/LVF/ L1 一种；[S][P][O]/SVK 形成了[S][P][O]/SVK/L1 一种；[S][P][O]/DVP 形成了[S][P][O]/DVP/L1 一种；[S][P][O]/SVT 形成了[S][P][O]/SVT/L1 一种；[S][P][O]/DVT 形成了[S][P][O]/DVT / L1 一种；[S][P][O]/SVN 仅形成了[S][P][O]/LVF/L4 一种；[S][P][O]/HVO 仅形成了[S][P][O]/HVO/L1 一种；[S][P][O]/SVG 仅形成了[S][P][O]/SVG /L1 一种；[S][P][O]/DVH 仅形成了[S][P][O]/DVH/L1 一种；[S][P][O]/HVX 仅形成了[S][P][O]/HVX/ L1 一种；[S][P][O]/DVN 仅形成了[S][P][O]/DVN / L1 一种；[S][P][O]/PVO 仅形成了[S][P][O]/PVO/L2 一种；[S][P][O]/EVK 仅形成了[S][P][O]/EVK/L1 一种；[S][P][O]/DVG 仅形成了[S][P][O]/DVG / L1 一种；[S][P][O]/DVA 仅形成了[S][P][O]/DVA / L2 一种；[S][P][O]/PVK 仅形成了[S][P][O]/PVK / L1 一种；[S][P][O]/SVB 仅形成了[S][P][O]/SVB/L1 一种；[S][P][O]/EVX 仅形成了[S][P][O]/EVX/L1 一种；[S][P][O]/PVX 仅形成了[S][P][O]/PVX / L1 一种；[S][P][O]/DVJ 仅形成了[S][P][O]/DVJ / L1 一种；[S][P][O]/SVJ 仅形成了[S][P][O]/SVJ / L1 一种。

5 从[S][P][O]对应的句样体系看现代汉语句系特点

第一，现代汉语句系是一个相互关联的层级系统。其层级性是通过句样去体现的。句样是从三个平面对句子进行抽象的结果。新句样是以上位句干为基础形成下位句干再经过语用指派形成的。

第二，句干经过语用选择形成句样时，陈述语气的优先选择性最高。陈述语气的句样类型使用频度最高，往往成为派生新句样类型的基础，属于原型成员。而其他语气类型的句样使用频度低。

第三，在形成新的句样的过程中，递归性原理起了主导作用。新句样的形成是在新句干基础上语用指派完成的。句样的形成首先表现在句干的形成上。就是在上位句型的基础上，在句首或句中不断添加状语，在高频句模的基础上不断增加可有论元从而形成了各种下位句干。

参考文献：

[1] C.J.Fillmore.*The case for case*《"格"辨》（中译本）[M].北京:商务印书馆,2002.

[2] 范　晓. 略说句系学[J]. 汉语学习, 1999（6）:2~5.

[3] 亢世勇[,孙茂松,田珍都]. 基于标注语料库的现代汉语句子语义结构研究[J]. Advances in Computation of Oriental Language,北京:清华大学出版社,2003.

[4] 朱晓亚. 现代汉语句模研究[M]. 北京:北京大学出版社, 2001.