基于标注语料库的现代汉语状元槽序研究*

周明海　　亢世勇

鲁东大学中文信息处理研究所，山东烟台，264025

E-mail:freer516@163.com　kangsy64@163.com

摘　要：句义的核心内容由命题和情态两部分构成，格关系和槽关系是对命题的深入研究，副词、能愿动词等状元则是情态的一部分，目前中文信息处理领域还缺乏深入、系统的研究。本文在介绍前人研究成果的基础上，从“事件描述块句法语义标注语料库”中抽取了副词、能愿动词连用的句子929个，对现代汉语状元的槽序进行了深入的研究，共得出现代汉语状元槽序120种，其中两个槽类连用的共774个、52种，三个槽类连用的共145个、58种，四个槽类连用的共10个、10种。现代汉语状元槽序虽然复杂，但使用频率较高的相对集中。最后我们根据120种槽序总结出了状元连用的级位链。

关键词：状元，槽序，级位链

Corpus-based Study of Modern Chinese Adverb Arguments Slot Orders

Zhou Minghai　　Kang Shiyong

Institute of Chinese Information Processing,Ludong University,Yantai,264025

E-mail:freer516@163.com　　kangsy64@163.com

Abstract: The core content of sentence meaning composes of Proposition and Modality, case relations and slot relations are deep study of Proposition. Adverb arguments, such as adverbs and modal auxiliaries are one part of Modality and it lacks deep and systematic study in the field of Chinese information processing at present. Based on reviewing the previous studies, this paper extracts 929 sentences in which adverbs and modal auxiliaries are used continually from Event Description Chunk Syntactic and Semantic Tagging Corpus. Slot relations of modern Chinese adverb arguments are studied and 120 slot orders are obtained. Among them, the number and category of two, three and four slot categories continually used are 774 and 52, 145 and 58, 10 and 10 respectively. Although slot relations of adverb arguments are complex, there is a tendency of centralization of high usage frequency. Finally, a rank chain for adverb arguments continually used is summarized according to 120 categories of slot orders.

Keywords: adverb argument, slot order, rank chain

1　引言

把语句的意义区分为主客观两部分从弗雷格就已经开始了，他在《算数基础》导论中就提出了研究逻辑语言哲学要“区分心理的东西和逻辑的东西、主观的东西和客观的东西”的基本原则，后来罗素、维特根斯坦进一步发展了该理论。在现代，俄罗斯当代语义学家则更深入的研究了这一问题，他们认为句子的意义可以分割为客观与主观两个性质不同的方面，其中句子的客观意义就是命题，句子的主观意义就是情态性。

在自然语言处理领域，菲尔默在20世纪60年代也提出了反映句义核心内容的经典公式：

*本文承863项目“智能感知与先进计算技术专题”（项目编号：2007AA01Z173）子课题“构建汉语句法语义标注库”的资助。

$S \rightarrow M+P$，$P \rightarrow V+C_1+C_2+\cdots\cdots+C_n$. $C \rightarrow K+NP$

正如林杏光（1999）先生所说，菲尔墨只研究了P，而没有研究M（情态），近些年来国内相关研究也大都集中在P上，对于M的研究则很少。林杏光、鲁川（1997）从宏观上提出了要深入研究句子语义平面的主客观信息，并系统深入地研究了语块内部以名词为中心的槽关系，即上面公式中C的内部关系，陈群秀（2000，2001）在此基础上则做了进一步完善。但相对以名词为中心的槽关系，我们认为与目标动词直接相关的修饰副词、能愿动词，简单地说就是语义上指向动词的成分，它们反映了相应事件内容的情态、否定、补充、时态等信息，对准确理解事件内容更有帮助，更需要深入研究，如："自大的他几乎成功了"，如果不研究"几乎"，则就不能完整地理解句意。出于这个原因，我们在综观前人对状元标注情况的基础上，从大规模标注语料库中抽取状元连用的句子进行了系统研究。

2 相关的几个概念

为了行文需要，我们先界定四个概念：

（1）状元：状元的通行定义是"状元即非核心论元、环境论元，如时间、处所、方式、工具、材料……"。一部分人将状元限定在与动核相联系的可有语义成分，可以由介词短语充当，但不包含副词；另有一部分人则认为副词也可以充当语义角色，是状元并对之进行标注，如"知网、CFN"等。本文的状元特指由副词和能愿动充当的语义角色。

（2）能愿动词做状语

能愿动词又叫助动词，能用在动词、形容词前边表示客观的可能性、必要性和人的主观意愿，对表达句子意思有着很重要的作用，在句子中做状语，我们将这部分词也看成状元，其语义角色为"评论"。

（3）状元槽序

林杏光、张庆旭（1998）认为根据偏词和正词的语义关系划分出来的语义类别就是槽类。多个槽类的排列有其一定的顺序，即槽序。为了和鲁川、陈群秀等研究的以名词为中心的槽类顺序相区别，我们把副词和能愿动词的槽类顺序称为状元槽序。

（4）级位链

槽类的连用并不是杂乱无章的，而是有一定顺序的，一般将出现在前面的称为高位，把出现在后面的称为低位。这样，把各个槽类按由高位到低位的顺序排列起来便成了状元的级位链。

3 目前对状元标注的概况

目前，语义角色的标注主要集中在动词和名词及介词短语的关系上，很少有人关注副词及能愿动词的系统标注，但各家标注却大都涉及副词，为了更好地汲取前人的经验、成果，我们首先对国内外状元标注的概况进行梳理。

3.1 国内对状元标注的概况

3.1.1 鲁川对状元的标注

林杏光、鲁川认为情态（modality）是表达言者由于主观的认识而引发的对事件的"情"绪

和“态”度以及对事件的评估，包括信念、情绪、观点、观察角度、态度、意图等；汉语的情态在语法上表现为虚词，即表达范畴，包括语气、情态、时态、动貌、语态、语式6类。在实际语料中，鲁川（2001）把情态表述为“事件的语意依附”，包括“必然、肯定、否定、照常、反常、遗憾、不足、过分、幸好、礼貌、揣测、理应、评价、能够”等十四类，其中只有“礼貌”和“能够”与副词无关，其余的都涉及对副词及能愿动词的角色标注。

这个体系对副词及能愿动词的处理是比较全面的，但对“不足”类、以及“究竟”、“毕竟”等词的处理不是很得当。另外，“亲自”一类的词很难归入上面的类。

3.1.2 CFN对状元的标注

汉语框架网（Chinese FrameNet，CFN）是由上海师范大学和山西大学合作开发的汉语语义描述系统。该语义知识库以Fillmore的框架语义学为理论基础，以加州大学的FrameNet为参照，对汉语语义进行了形式化描写。在标注中他们区分了核心框架元素和非核心框架元素，其中非核心框架元素中有一些是各个框架通用的，被称为“通用非核心元素”，共31个，如“物量、受益人、并行事件……”，其涉及的副词有：时间(time)、角色事件时间（time_role)、频率(freq)、特殊重复(part_iter)、形容(depic)、事件评价(eve_desc)、角色范围(sco_role)、程度(degr)、修饰(manr)等九类。

CFN并不刻意区分名词语义角色和副词语义角色，对能愿动词也未做出说明。通用非核心框架元素“修饰”是一个杂类，CFN把不能归为其他类的几乎全部归入了“修饰”，另外，他们不标注关联副词和否定副词。

3.1.3 知网对状元的标注

“知网”（HowNet）是由董振东、董强等创建的，是一个以汉语和英语的词语所代表的概念为描述对象，以揭示概念与概念之间以及概念所具有的属性之间的关系为基本内容的常识知识库。知网中共设计了89种动态角色，其中与副词、能愿动词有关的语义角色有8个：时距（timeRange)、评论(comment)、程度(degree)、强调(emphasis)、频率(frequency)、方式(manner)、幅度（range)、次序（sequence)，他们把能愿动词归在评论类里。

和CFN一样“知网”也不区分名词语义角色和副词语义角色，且所标注的副词数量有限。

3.1.4 北大中文网库对状元的标注

“北大中文网库”(Peking University Chinese NetBank）是一个正在建设中的对汉语大规模真实文本（100万字级以上）进行多层次的语义关系标注的语料库，旨在通过对语料进行多层次的语义标注，来给汉语的论元结构、逻辑结构和篇章结构等语义关系及其句法实现建立文件，并为训练基于统计的自动语义分析系统提供数据。

北大中文网库将主观信息的标注放在逻辑语义关系中，其包括否定关系、模态关系、时体关系、称代关系和指示关系，主要涉及否定算子、模态算子和时体算子跟受其约束的成分之间的逻辑语义关系。否定算子主要是副词“不”和“没、没有”及助动词“别、甭”；模态算子主要是表示情态的助动词，如“能、能够、可以……”；时体算子包括时间副词、时态助词和语气词等。北大中文网库和鲁川的处理方法有相似之处，但标注的副词很有限。

3.2 国外对状元标注的概况

3.2.1 FrameNet对状元的标注

FrameNet（FN）是美国加州大学伯克利分校1997年开始构建的基于语料库的计算词典编纂工程，由Fillmore主持。FrameNet不仅认为动词、名词、形容词可以激活情景，就是副词也可以做为激活框架的元素，他们主要标注了那些表示说话者态度的目标副词。

CFN大致上可以看作FN的汉化，所以CFN里的角色大部分也是FN里有的。对能愿动词的处理，FN并未做出说明，但其有能力(Capability)这一框架。

3.2.2 PropBank对状元的标注

宾州大学命题库（PropBank）是在已有的句法分析上或者称为树库的基础上添加谓词—论元结构形成的。他有三个目标，一个重要的目标是对同一个动词通过不同的句法实现标注一致的论元标签；另一个目标是对动词的修饰语添加功能标签，如“方式、处所、时间”等，在语料库中统一格式为：ArgM-X，其中X有DIR、MNR、REC、PNC、DIS、MOD、TMP、LOC、EXT、PRD、CAU、ADV、NEG、STR等14种。

PropBank的辅助语义角色，既有由介词短语充当的，又有由副词充当的，副词充当的主要有MNR、TMP、EXT三类。另外，MOD指能愿动词的语义角色。

3.2.3 Chinese Proposition Bank对状元的标注

中文命题库（Chinese Proposition Bank）由Martha Palmer、Nianwen Xue等在宾州大学中文树库的基础上添加一层谓词—论元结构形成的。其方式大致和Propbank相同，但在处理ArgM-X上还是有所差异，他们称之为联结角色，共11个，即ADV、CND、CMP、EXT、LOC、PRP、TPC、BNF、DIR、FRQ、MNR、TMP。这十一个角色中FRQ、MNR、TMP涉及副词。

4 “事件描述块句法语义标注语料库”的建设及其对状元的处理

4.1 “事件描述块句法语义标注语料库”的建设

“事件描述块句法语义标注语料库”建设是863项目2007AA01Z173的一个子课题，其把研究的重点集中在对物质世界和人类社会中的几大类客观关系的实践内容分析和标注方面，通过真实文本句子中的事件情境内容的准确标注，在词汇层面上建立起句法关系与谓词-论元结构之间的内在联系，为进行大规模真实文本句子的事件内容信息分析提供重要的训练和测试语料库。

事件描述块的句法语义标注，使用清华大学开发的“事件描述块句法语义标注工具”，其目标是针对每个真实文本句子，在确定了目标动词义项描述的基础上，进一步确定该目标动词所反映事件情境的各个描述块，并对其进行句法语义标注。主要包括以下工作：（1）确定各个事件描述块在句子中的准确左右边界位置；（2）确定各个描述块在句子中的功能位置信息，即确定该描述块处于主语位置还是宾语位置等；（3）确定各个描述块的句法成分标记，即确定该描述块是名词短语还是动词短语等；（4）确定各个描述块的中心词；（5）确定各个描述块的语义角色信息。

4.2 “事件描述块句法语义标注语料库”对副词及能愿动词的标注

语义角色与目标动词的关系亲疏远近并不一样，根据原型论的观点，可以把语义角色分为：核心语义角色、外围语义角色和边缘语义角色。一些与目标动词直接相关的修饰副词和紧密联系补语成分，由于反映了相应事件内容的情态、肯定/否定、补充、时态等信息，对准确理解事件内容很有帮助。在借鉴“知网”语义角色体系的基础上我们补充了九个辅助语义角色来标记它们，

其中与状元有关的是八个，下表是这八个辅助语义角色的相关信息及其在语料库中的标注情况：

	名称	英文	代号	含义	个数	百分比
1	程度	degree	DG	事件或属性值的程度	1120	5.66%
2	范围限定	scope restrictive	SR	对事物或动作行为进行范围的限定	1480	7.47%
3	肯定/否定	affirmative/negative	KF	表示对事件的肯定或否认	1837	9.28%
4	方式	manner	FS	事件发生或动作行为实施的方式	3876	19.57%
5	频率	frequency	F	事件发生或进行的频率	1306	6.59%
6	评论	comment	CM	表示对动作行为的意愿或看法等	3516	17.75%
7	情态	modality	QT	表示动作行为进行的情貌或状态	1877	9.48%
8	时态	tense	TE	说话者观测事件或状态的时间角度	4791	24.19%
合计					19803	100%

注：在语料标注中辅助语义角色前面都加“0-”以和核心语义角色区别

5 现代汉语的状元槽序

我们这个课题共标注单义动词459个、71261个句子，多义动词119个、29626个句子，最后形成了由578个目标动词共100887个句子组成的“事件描述块句法语义标注语料库”。该语料库共标注副词语义角色19083次，我们抽取了所有副词连用的句子共929个，共120种槽序，其中两个连用的774个、52种，三个连用的145个、58种，四个连用的10个、10种。

5.1 两个槽类连用

两个槽类连用的共774个，占总数的83.32%，是最常见的状元连用形式，具体类型见下表：

状元1	状元2	个数	百分比	排位	状元1	状元2	个数	百分比	排位	状元1	状元2	个数	百分比	排位
TE	TE	93	12.02%	1	KF	TE	13	1.68%	19	QT	CM	3	0.39%	37
CM	FS	92	11.89%	2	DG	SR	11	1.42%	20	DG	FS	2	0.26%	38
CM	TE	65	8.40%	3	F	F	11	1.42%	21	DG	QT	2	0.26%	39
CM	DG	47	6.07%	4	CM	SR	9	1.16%	22	KF	FS	2	0.26%	40
KF	CM	41	5.30%	5	DG	KF	9	1.16%	23	SR	FS	2	0.26%	41
TE	CM	34	4.39%	6	TE	QT	9	1.16%	24	SR	KF	2	0.26%	42
SR	TE	32	4.13%	7	F	QT	9	1.16%	25	FS	TE	2	0.26%	43
CM	QT	30	3.88%	8	F	SR	9	1.16%	26	QT	FS	2	0.26%	44
TE	FS	27	3.49%	9	F	TE	9	1.16%	27	QT	QT	2	0.26%	45
F	FS	21	2.71%	10	KF	SR	7	0.90%	28	QT	SR	2	0.26%	46
KF	QT	19	2.45%	11	TE	SR	7	0.90%	29	DG	F	1	0.13%	47
CM	CM	18	2.33%	12	F	CM	7	0.90%	30	KF	DG	1	0.13%	48
SR	CM	17	2.20%	13	QT	TE	6	0.78%	31	KF	F	1	0.13%	49
CM	KF	16	2.07%	14	DG	TE	5	0.65%	32	F	DG	1	0.13%	50
FS	FS	15	1.94%	15	F	KF	5	0.65%	33	FS	SR	1	0.13%	51
CM	F	14	1.81%	16	FS	F	5	0.65%	34	QT	KF	1	0.13%	52
TE	F	14	1.81%	17	SR	SR	4	0.52%	35					
TE	KF	14	1.81%	18	DG	CM	3	0.39%	36					

从概率上来讲，8 种槽类应有 64 种槽序，但实际却只出现了 52 种，其中 DG+DG、FS+CM、FS+DG、FS+QT、FS+KF、KF+KF、QT+F、QT+DG、SR+F、SR+QT、SR+ DG、TE+DG 在真实语料中没有出现，这 12 种应是现实中不能说或不常说的槽序。表中前十种占总数的 62.28%，是优势槽序，我们可以根据这十种槽序排出一个简单的级位链：SR→TE→KF→CM→F→FS，由于 DG、QT 都只出现一次，还无法归入这个级位链里。

上表中同槽类连用的有 6 个，缺少 DG+DG、KF+KF。另外，前位 CM、F 可以任意后接副词，后位 FS、TE 可以任意前接副词，但他们没有一个槽类是完全自由的，其中 CM 缺少前位 FS，F 则缺少前位 QT、SR。相反，DG 灵活度最差，只能出现在 CM、F、KF 之后，这说明 DG 的级位是较高的，但 DG 在这三类副词后的数量又比出现在前位时的数量多，这又说明 DG 也是一个较低位的槽类，也就是 DG 的槽位兼有较高和较低的性质，这应与其内部分类有关，同样情况的还有 SR、TE。

5.2 三个槽类连用

三个槽类连用的共 145 个，占总数的 15.61%，具体类型见下表：

状元1	状元2	状元3	个数	百分比	排位	状元1	状元2	状元3	个数	百分比	排位	状元1	状元2	状元3	个数	百分比	排位
TE	KF	CM	17	11.72%	1	SR	KF	CM	2	1.38%	21	CM	CM	F	1	0.69%	41
KF	CM	FS	10	6.90%	2	CM	TE	CM	2	1.38%	22	CM	CM	QT	1	0.69%	42
TE	CM	FS	8	5.52%	3	F	KF	CM	2	1.38%	23	CM	F	FS	1	0.69%	43
TE	CM	QT	8	5.52%	4	F	SR	CM	2	1.38%	24	CM	KF	FS	1	0.69%	44
KF	CM	F	7	4.83%	5	TE	CM	CM	1	0.69%	25	CM	SR	TE	1	0.69%	45
CM	FS	FS	7	4.83%	6	TE	KF	TE	1	0.69%	26	CM	TE	QT	1	0.69%	46
SR	CM	FS	6	4.14%	7	TE	QT	KF	1	0.69%	27	CM	TE	TE	1	0.69%	47
TE	CM	F	5	3.45%	8	TE	QT	TE	1	0.69%	28	F	CM	CM	1	0.69%	48
KF	CM	QT	4	2.76%	9	TE	TE	QT	1	0.69%	29	F	CM	TE	1	0.69%	49
SR	CM	DG	4	2.76%	10	TE	TE	FS	1	0.69%	30	F	F	FS	1	0.69%	50
CM	KF	CM	4	2.76%	11	TE	TE	TE	1	0.69%	31	F	KF	F	1	0.69%	51
TE	CM	TE	3	2.07%	12	TE	TE	F	1	0.69%	32	F	SR	QT	1	0.69%	52
TE	KF	SR	3	2.07%	13	TE	FS	FS	1	0.69%	33	F	TE	CM	1	0.69%	53
KF	CM	CM	3	2.07%	14	TE	F	FS	1	0.69%	34	QT	KF	CM	1	0.69%	54
CM	CM	FS	3	2.07%	15	KF	CM	TE	1	0.69%	35	QT	TE	CM	1	0.69%	55
F	TE	KF	3	2.07%	16	KF	KF	CM	1	0.69%	36	FS	TE	CM	1	0.69%	56
TE	CM	DG	2	1.38%	17	KF	QT	QT	1	0.69%	37	FS	FS	FS	1	0.69%	57
TE	CM	SR	2	1.38%	18	SR	DG	QT	1	0.69%	38	DG	CM	FS	1	0.69%	58
TE	KF	QT	2	1.38%	19	SR	TE	KF	1	0.69%	39						
KF	CM	SR	2	1.38%	20	SR	CM	QT	1	0.69%	40						

三个同槽类连用的有 FS、TE，其中 FS 可以合并成一个并列的副词词组，这样实际上只有 TE 可以三次连用，如下例：

[S-np 他/r 的/u 唱片/n]PN [D-dp 至今/d]O-TE [D-dp 还/d]O-TE [D-dp 在/d]O-TE [P-vp 出售/v]Tgt。

上表中占据状元 1 位置的槽类数量由多到少顺序为 TE>KF>CM>SR>F>FS=QT>DG，占据状元 2 位置的槽类数量由多到少顺序为 CM>KF>TE>FS>SR>F=QT>DG，占据状元 3 位置的槽类顺序为 FS>CM>QT>F>TE>SR>DG>KF，占据状元 3 位置的槽类 FS、QT、TE、DG 数量的增多说

明他们更容易占据靠后的位置。根据表中优势槽序，我们可以推断出如下的级位链：

TE→KF→CM→FS

5.3 四个槽类连用

四个槽类连用的共10个，占总数的6.90%，具体类型见下表：

状元1	状元2	状元3	状元4	个数	状元1	状元2	状元3	状元4	个数
TE	KF	CM	QT	1	KF	CM	F	FS	1
CM	KF	CM	TE	1	TE	KF	CM	F	1
CM	KF	CM	FS	1	TE	KF	CM	TE	1
KF	CM	FS	SR	1	F	TE	CM	QT	1
CM	KF	CM	CM	1	DG	TE	CM	FS	1

前八个槽序遵循三个状元槽类连用的级位链TE→KF→CM→FS，后两个也是部分遵守，这说明前面总结的级位链是正确的。总结上面三个表得出的级位链，再根据前、后位某个槽类出现的频率我们可以得出如下总的级位链：

DG→SR→TE→KF→CM→F→FS→DG→SR→TE→QT

6 结语

状元槽关系研究是句义情态研究的一部分，也是自然语言理解的一部分，其对状元的自动标注起着一定的指导作用，随着句处理研究的进一步深入必将得到重视。本文是在真实语料基础上的一个初探，对一些问题的处理还有待进一步思考，如“形容词+地”和成语修饰目标动词做状语标还是不标、如何标，“否定副词+能愿动词” 是标在一起还是分开标，我们在借鉴前人研究成果的基础上将状元设为八类是否囊括了所有的副词、能愿动词，情态和方式等易混的类该如何区分、有没有形式上的标志……。另外，现有的语料是针对具体目标动词抽取的，在全息语料中状元槽序又是什么样子？我们下一步打算在现有研究基础上，对中小学语文课本语料库标注辅助语义角色，以期得到更好的结果。

参 考 文 献

[1] Honglin Sun and Daniel Jurafsky.Shallow Semantic Parsing of Chinese. In Proceedings of NAACL-HLT, 2004.

[2] 陈群秀. 现代汉语名词槽关系系统研究初步进展. 语言文字应用, 2000(1).

[3] 林杏光. 词汇语义和计算语言学. 北京:语文出版社, 1999.

[4] 林杏光、张庆旭. 现代汉语槽关系研究. 汉语学习, 1998(6).

[5] 鲁川. 汉语语法的意合网络. 北京:商务印书馆, 2001.

[6] 袁毓林. 基于认知的汉语计算语言学研究. 北京:北京大学出版社, 2008.

[7] 张谊生. 副词的连用类别和共现顺序, 烟台大学学报(哲学社会科学版), 1996(2).