构式视角下的现代汉语词类范畴再认识

詹卫东
zwd@pku.edu.cn

北京大学中文系
2017.11

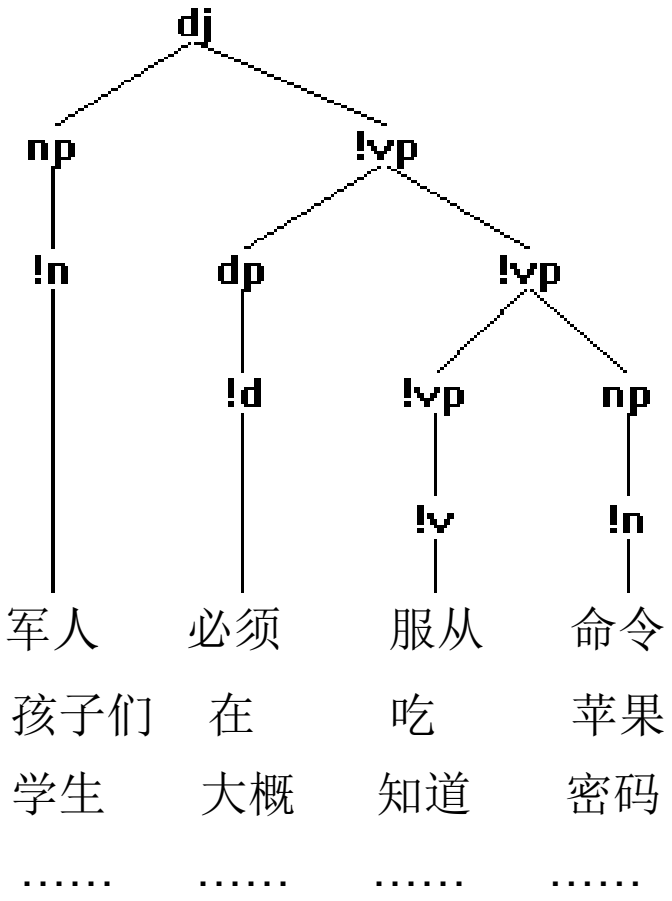
提纲

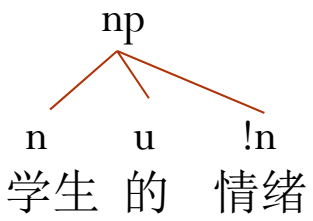
1. 背景
2. 现有的词类体系及主要争议问题
3. 从构式视角看词类
4. 结语

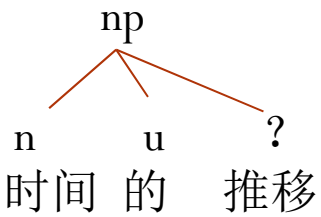
致谢：本文工作得到国家重点基础研究发展计划（2014CB340504）教育部人文社科重点研究基地重大项目（13JJD740001，15JJD740002）经费支持

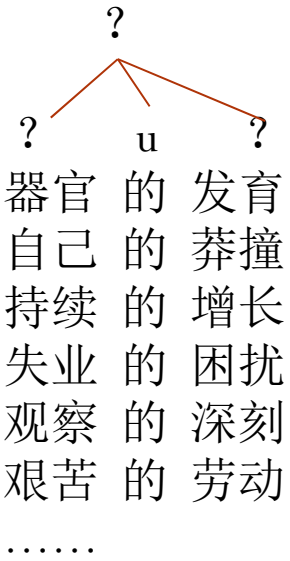
1 背景

- 面向中文信息处理，标注语料、构建语言知识资源（语料库、知识库）词类体系是“基础知识（基础的描述工具）”

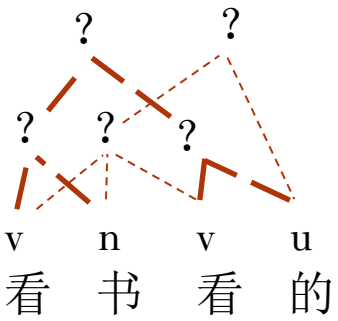




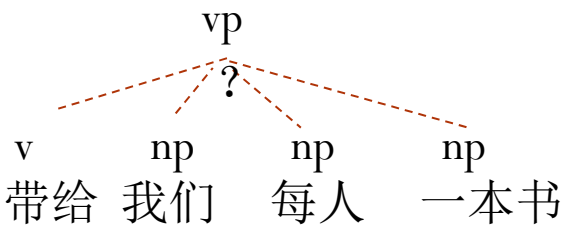




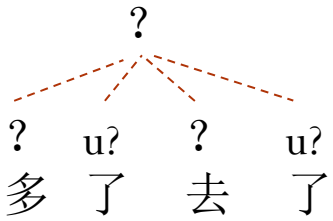
背景



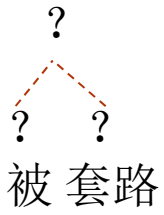
1. 他这眼睛是看书看的



2. 老张带给我们每人一本书



3. 天底下长得像的人多了去了



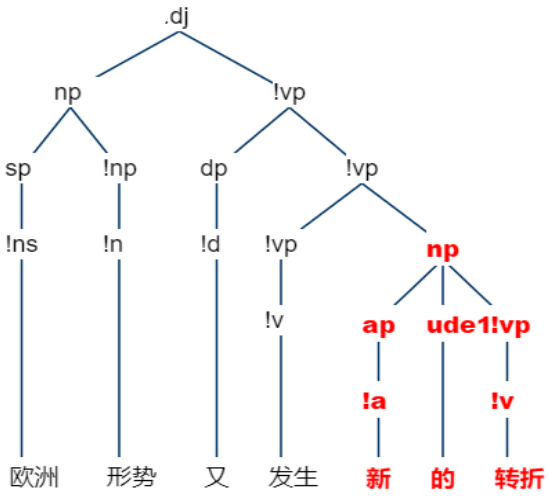
4. 承诺健康奇迹 老年人“被套路”

背景

- 从 短语结构描述 扩展到 构式描述

http://ccl.pku.edu.cn:8080/WebTreebank/

http://ccl.pku.edu.cn/ccgd



构式形式	构式类型	构式特征	释义模板	实例
关键+的+关键	凝固型	复现|主观大量	最+关键+的	关键的关键在第二天
a+着+r+的+a	短语型	语法错配|复现|主观大量|修辞	经历+r+的+a|体会+r+的+a	幸福着她的幸福
n+不+n+的+问题	短语型	复现|冗余	n+的+问题	这不是钱不钱的问题
v+帝	半凝固型	省略	擅长+v+的+人	抄袭帝|表演帝
半+v+不+v	半凝固型	复现	不+完全+v	半懂不懂|半信不信
v+也+得+v，不+v+也+得+v	复句型	省略|复现	无论+想+不+想+v，都+要+v	走也得走，不走也得走

组合：由小到大
递归：强

填槽：由大到小
递归：弱 或 无

构式中的内部成分，其范畴如何描述？
是否沿用短语结构描述中的词类体系？

2 现有的汉语词类体系及主要争议问题

- 理论导向的词类问题研究
  - 结构主义语法视角的词类划分
  - 生成语法视角的词类划分
  - 认知功能语法视角的词类划分
  - 语言类型学与汉语词类
- 应用导向的词类问题研究
  - 汉语教学
  - 辞书编纂
  - 中文信息处理

[±N ±V]

典型范畴理论

汉语无独立形容词类，形容词属于动词

1. 汉语词类有多少类？
2. 词类体系的设计是否有好与不好的分别？判断标准是什么？

现有的汉语词类体系及主要争议问题

考察：

- 一些中文著作中的词类体系 （20世纪、21世纪）
- 一些英文著作中的汉语词类
- 一些信息处理用词类标记集 （7家单位，10个标记集）

20世纪的11家词类体系

1.	马建忠《马氏文通》	（1898）	9类
2.	黎锦熙《新著国语文法》	（1924）	9类
3.	吕叔湘《中国文法要略》	（1942,1944）	9类
4.	王力《中国现代语法》	（1943,1944）	11类
5.	丁声树等《现代汉语语法讲话》	（1952-1953）	12类
6.	张志公《暂拟汉语教学语法系统》/《中学教学语法系统提要》	（1956，1984）	12类
7.	胡裕树《现代汉语》	（1981）	13类
8.	黄伯荣、廖序东《现代汉语》	（1985）	14类
9.	朱德熙《语法讲话》	（1982）	17类
10.	北大中文系《现代汉语》	（1993）	15类
11.	张斌《现代汉语》	（1996）	13类

陆俭明（2003）《现代汉语语法研究教程》北京大学出版社

十一家词类体系比较

	① 名词	② 时间词	③ 处所词	④ 方位词	⑤ 动词	⑥ 助动词	⑦ 形容词	⑧ 状态词	⑨ 区别词	⑩ 数词	⑪ 量词	⑫ 副词	⑬ 代词	⑭ 连词	⑮ 介词	⑯ 助词	⑰ 语气词	⑱ 叹词	⑲ 拟声词
1	+				+		+					+	+	+	+	+		+	
2	+				+		+					+	+	+	+	+		+	
3	+				+		+					+	+	+	+	+	+		
4	+				+		+			+		+	+	+	+	+	+	+	
5	+				+		+			+	+	+	+	+	+	+	+	+	
6	+				+		+			+	+	+	+	+	+	+		+	+
7	+				+	+	+			+	+	+	+	+	+	+	+	+	
8	+				+		+		+	+	+	+	+	+	+	+	+	+	+
9	+	+	+	+	+		+		+	+	+	+	+	+	+	+	+	+	+
10	+				+		+	+	+	+	+	+	+	+	+	+	+	+	+
11	+				+		+			+	+	+	+	+	+	+	+	+	+

21世纪的若干汉语词类体系

1.	黄伯荣、廖旭东《现代汉语》第三版，高等教育出版社2002年	14 类
2.	张斌 《新编现代汉语》 复旦大学出版社 2002年	16 类
3.	邵敬敏 《现代汉语通论》 第二版，上海教育出版社，2007年	14 类
4.	北大中文系现代汉语教研室编《现代汉语》（重排本） 商务印书馆 2004年	15 类
5.	陆俭明《现代汉语语法研究教程》北京大学出版社2003年	15 类
6.	郭锐《现代汉语词类研究》商务印书馆2002年	20 类
7.	沈阳、郭锐主编《现代汉语》高等教育出版社2014年	20 类
8.	《现代汉语词典》（第六版）商务印书馆	12/17 类

21世纪八种词类体系比较

	① 名词	② 时间词	③ 处所词	④ 方位词	⑤ 动词	⑥ 形容词	⑦ 状态词	⑧ 区别词	⑨ 数词	⑩ 量词	⑪ 数量词	⑫ 副词	⑬ 指示词	⑭ 代词	⑮ 连词	⑯ 介词	⑰ 助词	⑱ 语气词	⑲ 叹词	⑳ 拟声词
1	+	-	-	-	+	+	-	+	+	+	-	+	-	+	+	+	+	+	+	+
2	+	-	-	-	+	+	-	+	+	+	-	+	-	+[+]	+	+	+	+	+	+
3	+	-	-	-	+	+	-	+	+	+	-	+	-	+	+	+	+	+	+	+
4	+	-	-	-	+	+	+	+	+	+	-	+	-	+	+	+	+	+	+	+
5	+	-	-	-	+	+	+	+	+	+	-	+	-	+	+	+	+	+	+	+
6	+	+	+	+	+	+	+	+	+	+	+	+	+	+[+]	+	+	+	+	+	+
7	+	+	+	+	+	+	+	+	+	+	+	+	+	+[+]	+	+	+	+	+	+
8	+	(+)	-	(+)	+	+	(+)	(+)	+	+	-	+	-	+	+	+	+	(+)	+	+

面向人的词类划分：

- 词类数量呈现增加趋势
- 新设立的词类反映了对一部分词的语法特点的认识更加精细

20世纪的词类体系中，各家都认可的词类是8类。到21世纪的词类体系，各家都认可的词类增加到14类

一些英文著作中的汉语词类

1. Yuen Ren Chao, 1968, A Grammar of Spoken Chinese, University of California Press.

2. Charels N. Li, Sandra A. Thompson, 1981, Mandarin Chinese: A Functional Reference Grammar, University of California Press.

3. Ying-Che Li, Robert L. Cheng, Larry Foster, Shang H. Ho, John Y. Hou, Moira Yip, 1984, Mandarin Chinese: A Practical Reference Grammar for Students and Teachers (Vol. I, II), The Crane Publishing Co. Chinese Materials Center Publications

4. C.T. James Huang, Y.H. Audrey Li, Yafei Li, 2009, The Syntax of Chinese, Cambridge University Press.

5. Chu-Ren Huang, Dingxu Shi, 2016, A Reference Grammar of Chinese, Cambridge University Press, Chapter 2 Syntactic Overview, 2.1.3 Word Classes. pp.19-42.

Chao(1968)

词类篇幅：324/819 = 39.5% (cf.《语法讲义》 106/251 = 42.23% ）

- noun, proper names, place words, time words, D-M compounds, N-L compounds, Determinatives, Measure, Localizers, Pronoun 10 类
- **verbs（包括 adjectives）**, prepositions, adverbs, conjunctions, particles, interjections 6类

Huang & Shi(2016)

verbs, nouns, numerals, classifiers, adjectives, adverbs, prepositions, coordinators, interjections, sentence final particles, onomatopoeia 11类

信息处理用汉语词性标记集

1. 北大计算语言所词性标记集（1999版， 2002版，2003版）
2. 清华大学树库词性标记集
3. 北京语言大学与清华大学精加工语料库词性标记集
4. 中科院计算所（ICTCLAS2.0， ICTPOS3.0）
5. 社科院语用所 （信息处理用词性标注规范 - 国家标准）
6. 台湾中研院词库小组
7. 美国宾州大学中文树库（CTB）词性标记集

信息处理用汉语词性标记集主体标记

序号	词类	标记	序号	词类	标记	序号	词类	标记
1	名词	n	10	数词	m	19	简称略语	j
2	时间词	t	11	量词	q	20	习用语	l
3	处所词	s	12	副词	d	21	成语	i
4	方位词	f	13	介词	p	22	前接成分	h
5	动词	v	14	连词	c	23	后接成分	k
6	形容词	a	15	助词	u	24	语素字	g
7	区别词	b	16	叹词	e	25	非语素字	x
8	状态词	z	17	语气词	y	26	标点符号	w
9	代词	r	18	拟声词	o	27	连接语	(l)

信息处理学界所用的词性标记集主体部分跟汉语语法学界的词类划分基本一致

七家词性标记集比较（1）

	① 名词	② 时间词	③ 处所词	④ 方位词	⑤ 动词	⑥ 形容词	⑦ 区别词	⑧ 状态词	⑨ 代词	⑩ 数词	⑪ 量词	⑫ 副词	⑬ 介词	⑭ 连词	⑮ 助词	⑯ 叹词	⑰ 语气词	⑱ 拟声词
1	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
2	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
3	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
4	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
5	+	-	-	-	+	+	+	-	+	+	+	+	+	+	+	+	-	+
6	+	-	-	-	+	-	+	-	-	-	-	+	+	+	+	+	-	-
7	+	-	-	+	+	+	-	-	+	+	+	+	+	+	+	+	+	+

1 北大计算语言所词（ICL） 2 清华大学树库词性标记集 3 北京语言大学与清华大学精加工语料库词性标记集
4 中科院计算所（ICT） 5 社科院语用所 （信息处理用词性标注规范 - 国家标准） 6 台湾中研院词库小组
7 美国宾州大学中文树库（CTB）词性标记集

七家词性标记集比较（II）

	(19) 简称略语	(20) 习用语	(21) 成语	(22) 前接成分	(23) 后接成分	(24) 语素字	(25) 非语素字	(26) 标点符号	(27) 连接语
1	+	+	+	+	+	+	+	+	-
2	-	-	+	+	+	+	+	+	+
3	-	-	+	+	+	-	+	+	+
4	-	-	-	+	+	-	+	+	-
5	+	+	-	+	+	+	+	+	-
6	-	-	-	+	+	+	+	+	-
7	-	-	-	-	-	-	+	+	-

1 北大计算语言所词（ICL） 2 清华大学树库词性标记集 3 北京语言大学与清华大学精加工语料库词性标记集
4 中科院计算所（ICT） 5 社科院语用所 （信息处理用词性标注规范 - 国家标准） 6 台湾中研院词库小组
7 美国宾州大学中文树库（CTB）词性标记集

北大ICL《人民日报》语料库标准（1999）	北大ICL语料标注新标准（草稿2002）	北大ICL语料库加工规范（2003）	清华大学树库标准	北语-清华精加工语料库标记规范（1999）	计算所标记集ICTCLAS 2.0	计算所标记集ICTPOS3.0	语用所标准（GB/T 20532-2006）	Sinica Academia树库词性标记集	宾州大学中文树库ChinesePennTreeBank 3.0（2000）
39	40	106（26个一级类；72个二级类；8个三级类）	51+标点符号	86+标点符号	50（26个一级标记；24个二级标记）	99（22个一级标记，66个二级标记，11个三级标记）	48（普通词类：13个一级类，15个二级类；其他切分单位：7个一级类，13个二级类）	48+（11个一级标记，36个二级标记）（不含外文字符，标点）二级标记主要是动、名、副的次类	33

面向信息处理的词类标记集：

- 大类主体词类基本参考汉语语法学界的词类划分；
- 细类（二级、三级标记）更多，标记数量 50 – 100 范围；
- 进一步细分的词类主要是：n（名），v（动），a（形），d（副），m（数），q（量），p（介），r（代），u（助）

一些词类的功能（分布）进一步明确

代词

拟声词

→ 根据分布特点不同，分别归入现有的其他词类

助词 （助词的唯一共性恐怕是类中的每个词都很有个性）

↘

细分：一词一类？

根据义项（用法）不同分别标记。

比如：宾州树库“被”区分为LB，SB；

“的”区分为DEC，DEG；

“了”区分为AS，SP 等等

3 从构式视角看词类

平凡组合
（多数）

构式中“范畴”变异（泛化）的现象：

（甲） 构式中的某个变项位置要求A类词进入，
但是，B类词进入到该位置

超常组合
（少数）

（乙） 构式中的某个变项位置允准“引语”成分

（丙） 构式中的某个变项位置允准“指代”成分 （例：概说构式）

构式压制

从构式视角看词类

情况甲 例1 我 也 **大款** 一回

N 一回

被V

例2 这种赛制明显是极度不合理的，太容易被**田忌**了。

（他们）不安排**田忌**你们。。。。。。居然把最差的情况安排给我们。

V N

例3： a 你写你的

b 你走你的独木桥

c 我**胖**我的，关你什么事

RVR 的

情况乙 例4： a 一天到晚**啊**来**啊**去，啊个没完

b **漂亮**什么**漂亮**，难看死了

V 来V 去

X 什么 X

情况丙 例5： a “**四个全面**”战略布局

b 交通事故后「**三不一没有**」的处理方式靠谱吗？

- 这不是是不是的问题
- 这不是想不想的问题
- 这不是敢不敢的问题
- 这不是要不要的问题
- 这不是快不快的问题
- 这不是灵不灵的问题
- 这不是钱不钱的问题
- 这不是嫁妆不嫁妆的问题

情况甲

例6 这不是v不v的问题
这不是a不a的问题
这不是n不n的问题

这不是X不X的问题

情况甲

例7　为v而v
　　为n而n
　　为a而a

- 为工作而工作
- 为读书而读书
- 为上市而上市
- 为写文章而写文章
- 为教育而教育
- 为学术而学术
- 为文学而文学
- 为幽默而幽默
- 为大而大
- 为新而新
- 为新奇而新奇

情况甲

例8 有一种 n1 叫 n2
有一种 n 叫 a
有一种 a 叫 n
有一种 a 叫 v
有一种 v 叫 v
有一种 v 叫 a
有一种 n 叫 dj

有一种 X 叫Y

有 一种 婚姻 叫 事实婚姻
有 一种 毒药 叫 成功
有 一种 从容 叫 范冰冰
有 一种 力量 叫 平静
有 一种 爱 叫 放手
有 一种 爱 叫 永不放弃
有 一种 幸福 叫 转
有 一种 强拆 叫 公平
有 一种 心疼 叫 “随便你”
有 一种 失望 叫 “算了”
有 一种 误差 叫 数据造假
有 一种 倒下 叫 站起

被 自杀

被 吸烟

被 涨工资

被 开会

被 没有资格

被 奥数

被 高铁

被讨论

被 正态

被 平均

被 幸福

被 满意

被 67%

情况甲/乙

例9 被 v / vp

被 n

被 a

被 m

被 X

- 什么侵权不侵权的
- 什么值得不值得的
- 什么解脱不解脱的
- 什么顺利不顺利的
- 什么幸福不幸福的
- 什么难不难的
- 什么面子不面子的
- 什么美人不美人的
- 什么法不法的
- 什么文明不文明的
- 什么道德不道德的

情况甲/乙

例10 什么v不v的
什么a不a的
什么n不n的

什么X不X的

情况乙

例11： a 那位日本游客**“哈依哈依”地**直点头。

b 你傻呀，别老**大姐大姐**的，要叫就叫姐。

c 昔日老点头哈腰**局长长局长短**打招呼的人，现在见了都躲着走

刘丹青（2009）实词的拟声化重叠及其相关构式，《中国语文》2009年第1期，22-31页。

情况乙

- 张口政治闭口政治
- 张口民主闭口民主
- 张口李总闭口李总
- 张口修炼闭口养生
- 张口是权闭口是钱
- 张口委屈闭口委屈
- 甄嬛体走红网络 观众张口“极好”闭口“真真”
- 我最讨厌张口“高富帅”闭口“白富美”的

例12 张口n闭口n
张口v闭口v
张口a闭口a

思考：词类信息如何帮助句法分析？

例1： a 你 说 的 **好**，他 说 的 不 好

b 你 说 的 **好**，做 起来 完全 不 是 那么 回 事

c 你 说 的 **好**，还 不 是 真 的 好

d 你 说 的 **好**，声调 有些 不 太 准确

例2： a 这本书 的 **出版** 令 学术界 感到 很 振奋

b 老王 的 **出版** 不 到 一 年 就 卖 完 了

c 关于 北大 的 **出版** 当然 比较 容易

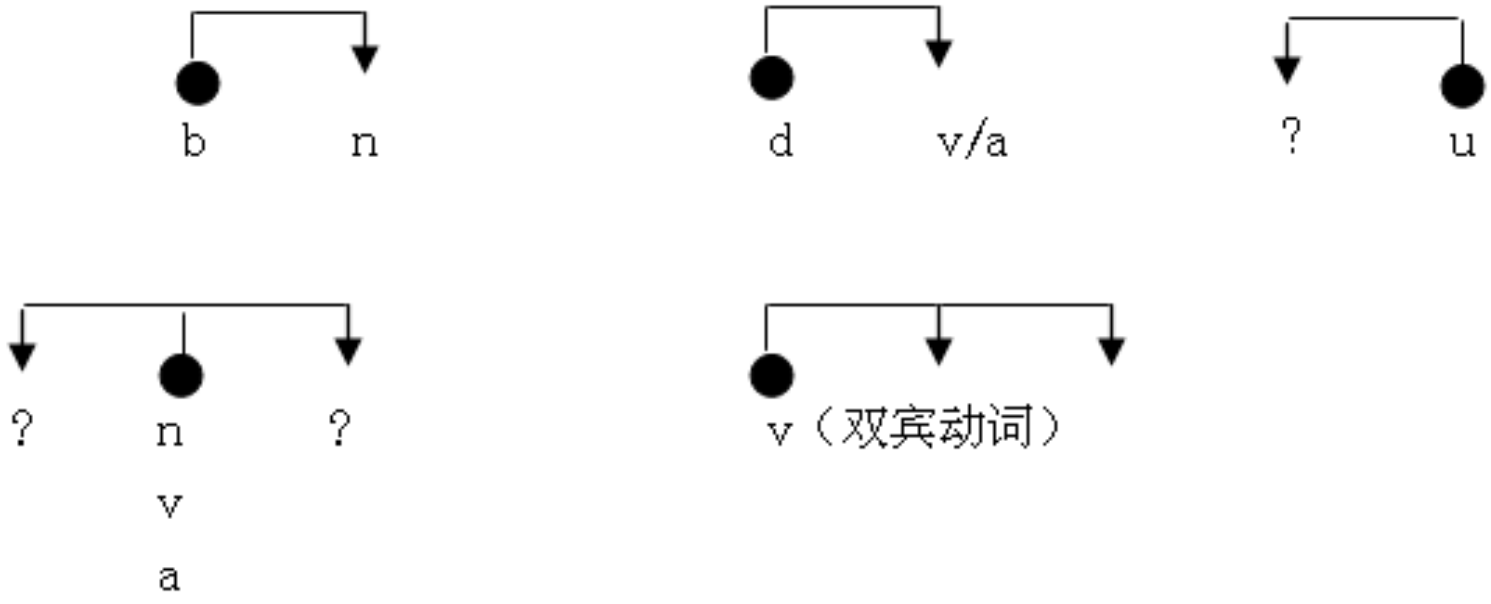
d 关于 北大 的 **出版** 意见 相当 多

如何给出组合约束条件，是最重要的！

仅知道“好”是形容词，“出版”是动词，对句法（语义）分析能起到多大的帮助作用？

“词类/词性”可以反映（一部分）分布

词类（分布）信息 = 组合方向 + 对组合对象的约束



b: 区别词　d：副词　u：助词　v：动词　a: 形容词　n：名词

词类信息如何帮助句法分析？

问题1：　更细？ ← 词类 → 更粗？

问题2：　如果要更细，怎么个细法？ → 增加词类数量？

→ 增加特征数量？

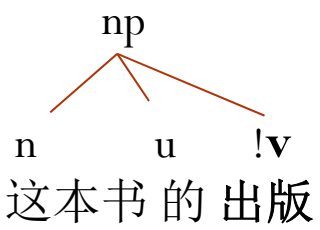
(1) 语义标准定类

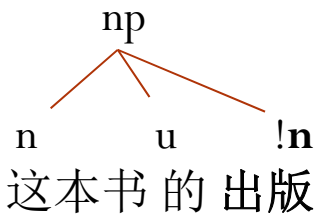
(2) 形式（分布）标准定类

(3) 语义×形式标准定类

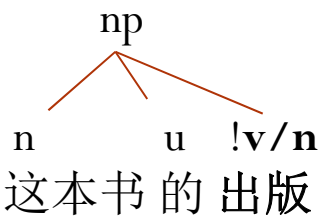
越分越细

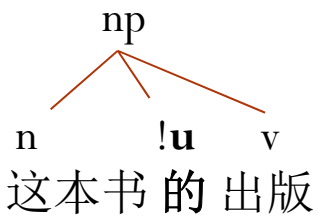
构式槽（变项）允准多个词类





对比：
n 的 老张 的
a 的 白 的
v 的 吃 的
↓
X 的





器官 的 发育
自己 的 莽撞
持续 的 增长
失业 的 困扰
观察 的 深刻
艰苦 的 劳动
……

作家（中） 的 作家
教授（里） 的 教授
？……

万一 的 万一
？……

X 的Y → X 的 X

从意义到形式：分类的细化过程

形式 / 意义	F_1	F_2	F_3	……	F_j	……
M_1	$W_{1,1}$	$W_{1,2}$	-		-	
M_2	-	-	$W_{2,3}$		-	
M_3	-	-	$W_{3,3}$		-	
……						
M_i	-	-	-		$W_{i,j}$	
……						

例1：$W_{1,1}$ $W_{1,2}$ $W_{1,3}$ $W_{1,4}$ $W_{1,5}$ $W_{1,6}$ $W_{1,7}$ ……
爸爸 父亲 老豆 老爸 爹 爹哋 家父 ……

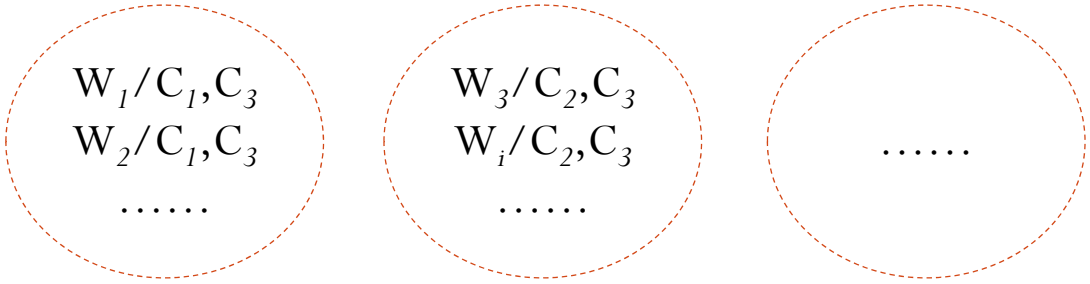
例2：$W_{1,1}$ $W_{1,2}$ $W_{1,3}$ $W_{1,4}$ $W_{1,5}$ $W_{1,6}$ $W_{1,7}$ ……
桌子 椅子 板凳 床 沙发 书柜 壁橱 ……

例3：$W_{1,1}$ 迅速　a 他迅速销毁了密码本　　b 他销毁密码本很迅速

$W_{1,2}$ 立即　a 他立即销毁了密码本　　b * 他销毁密码本很立即

从词组到构式：分类的细化过程

构式 / 词	C_1	C_2	C_3	……	C_j	……
W_1	+	-	+		-	
W_2	+	-	+		-	
W_3	-	+	+		-	
……						
W_i	-	+	+		-	
……						



词类1　　词类2　　词类$_i$

语义分析模式：从 论旨角色 到 框架元素

a） 王老师 在 考 小明
b） 小明 在 考 语文
c） 王老师 在 考 小明 语文
d） 小明 今年 考 大学
e） 语文 只 考 了 一道题

格角色

施事	考	受事
施事	考	内容/范围
施事	考	受事 内容/范围
施事	考	目的/结果
内容?	考	内容?

针对全体动词的设计

框架元素

核心元素	考试（examination）
	被试（examinee）
	主考（examiner）
	知识（knowledge）
	资质证明（qualification）
非核心元素	方式（manner）
	方法（means）
	考场（place）
	目的（purpose）
	时间（time）

针对部分动词的设置

4 结语

- 从 短语结构描述 扩展到 构式描述

http://ccl.pku.edu.cn:8080/WebTreebank/

- 语言是一个规则系统
- 规则刻画了语法成分组合的递归性
- 大的语法单位的意义由小的语法单位组合得到
- 基本语法单位的范畴性质（词类）在语言系统的全体组合关系（分布）基础上得到定义。

http://ccl.pku.edu.cn/ccgd

- 构式是“形式-意义”的配对
- 构式是有限能产的，有组合性但无递归性
- 构式有独立的意义，不能由其组成部分的意义组合得到
- 构式中成分的范畴性质既要从全局性组合关系的层面认识，更要从构式自身组合关系的层面认识。因为构式中的成分是“不自由的、不透明的”。

结语

- 词类划分依赖句法结构关系。
  假设语言系统有普遍的针对所有单位的结构关系集合。
  在该结构关系集合的基础上，定义词的分布。
  以上述方式得到的词类，是普遍的，针对所有词的一个词类集。

- 构式语法的观念：无法给出一个普遍适用的结构关系的清单
  语言系统中存在的是一个一个的独立的构式。
  出现于某一构式中的词，所形成的范畴（词类），对其他构式不完全（精确）适用。

- 对人：词类不宜过多；倾向于语义制导的词类划分。

  对机器：词类可以相对多； 倾向于分布制导的词类划分
  （分布式语义表示）

  实现策略：类 + 特征　更容易进行知识管理

参考文献

- 略

欢迎访问

- http://ccl.pku.edu.cn/ccgd （现代汉语构式知识库）

敬请批评指正

谢谢大家！