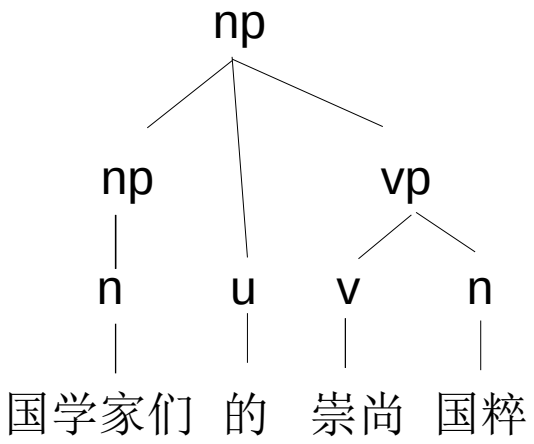
从形式文法歧义角度看短语范畴设置

詹卫东
zwd@pku.edu.cn
北京大学中文系

汉语句法结构中的非中心扩展模式：“np 的 vp”

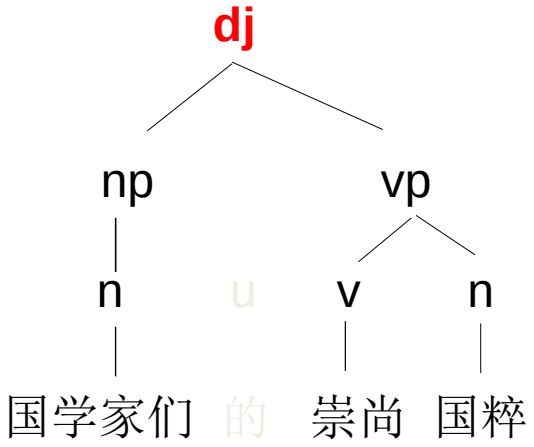
国学家们的**崇尚国粹**自是情理之中的事。



汉语句法结构中的非中心扩展模式：“np 的 vp”

国学家们的**崇尚国粹**自是情理之中的事。

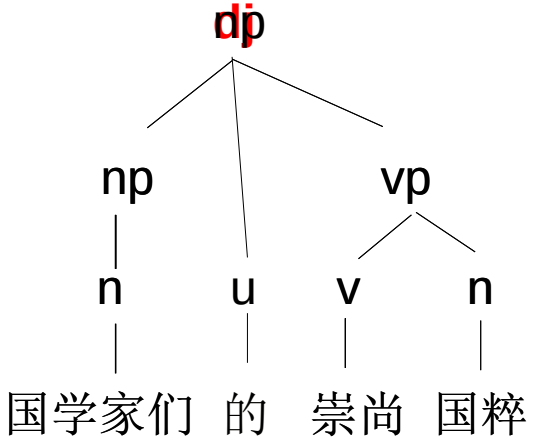
国学家们**崇尚国粹**自是情理之中的事。



汉语句法结构中的非中心扩展模式：“np 的 vp”

国学家们的**崇尚国粹**自是情理之中的事。

国学家们**崇尚国粹**自是情理之中的事。

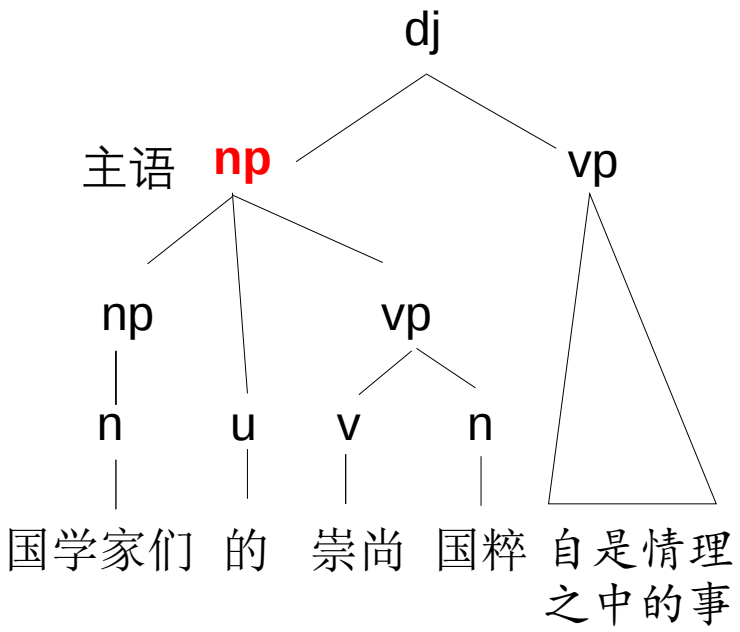


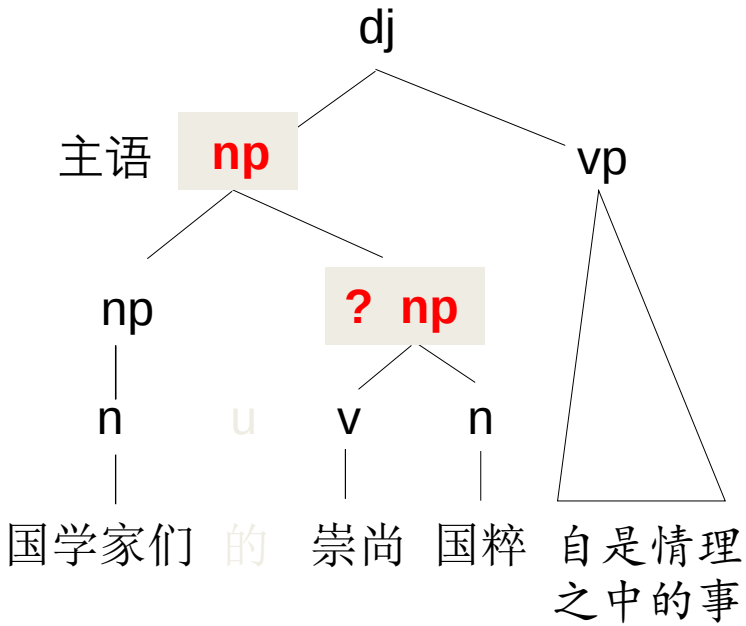
- 中心扩展模式是结构方式中的**主流，但非全部**。→ 功能稳定性
- 非中心扩展模式也是一种结构方式，这种模式是使表达简省的有效手段：**一个形式，多种功能**。→ 功能多样性

非中心扩展模式“np 的 vp”在主语位置

国学家们的**崇尚国粹**自是情理之中的事。

国学家们**崇尚国粹**自是情理之中的事。

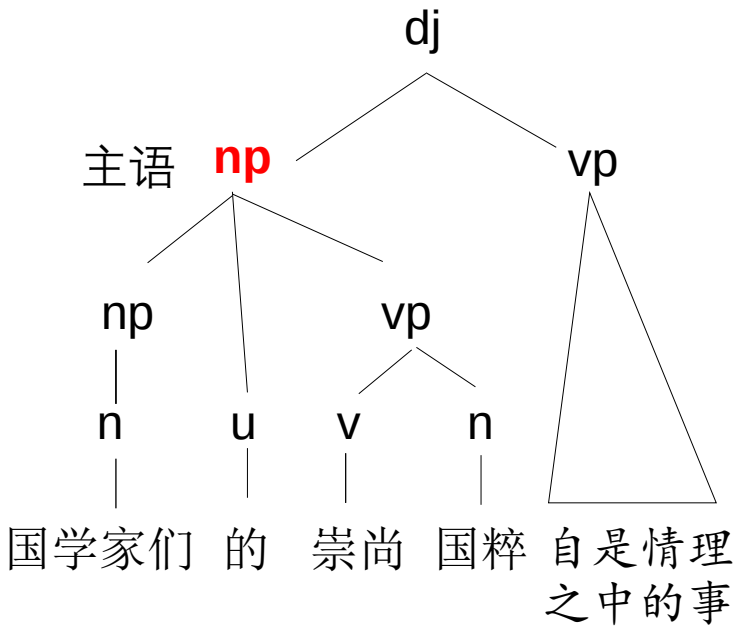


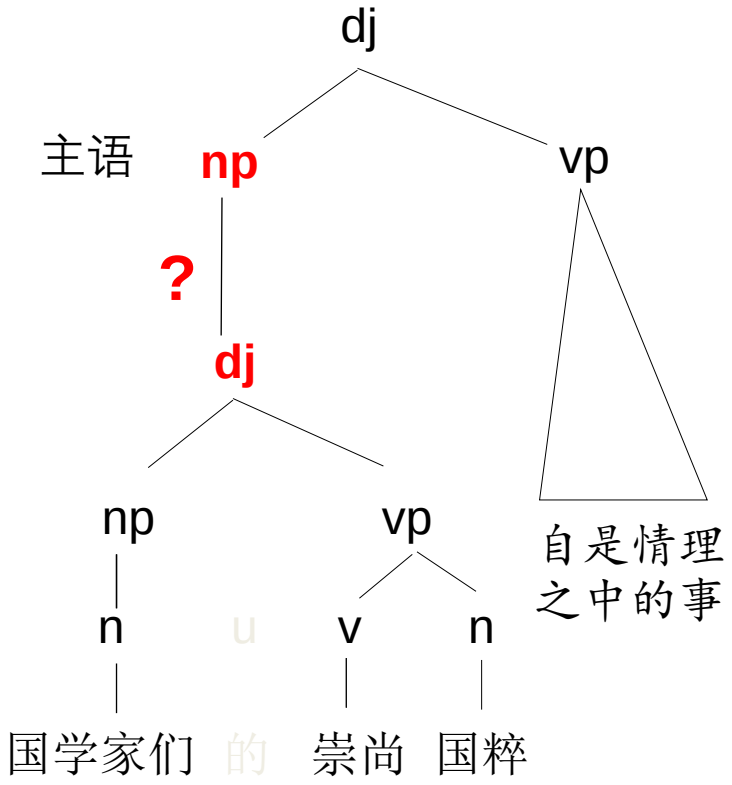


非中心扩展模式“np 的 vp”在主语位置

国学家们的**崇尚国粹**自是情理之中的事。

国学家们**崇尚国粹**自是情理之中的事。

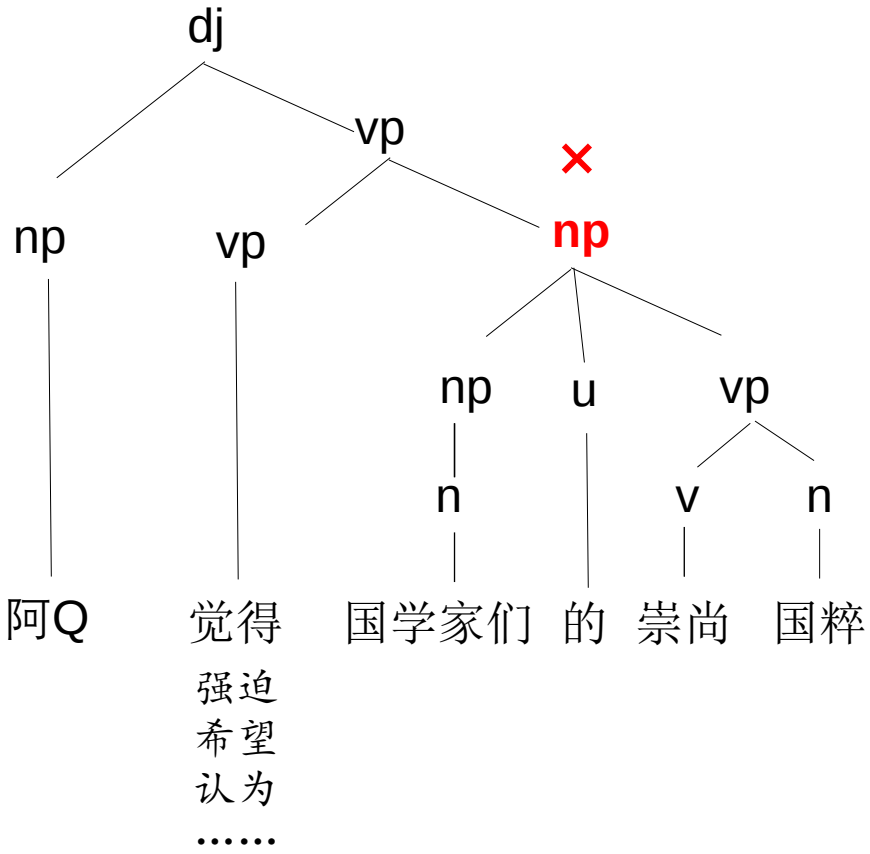


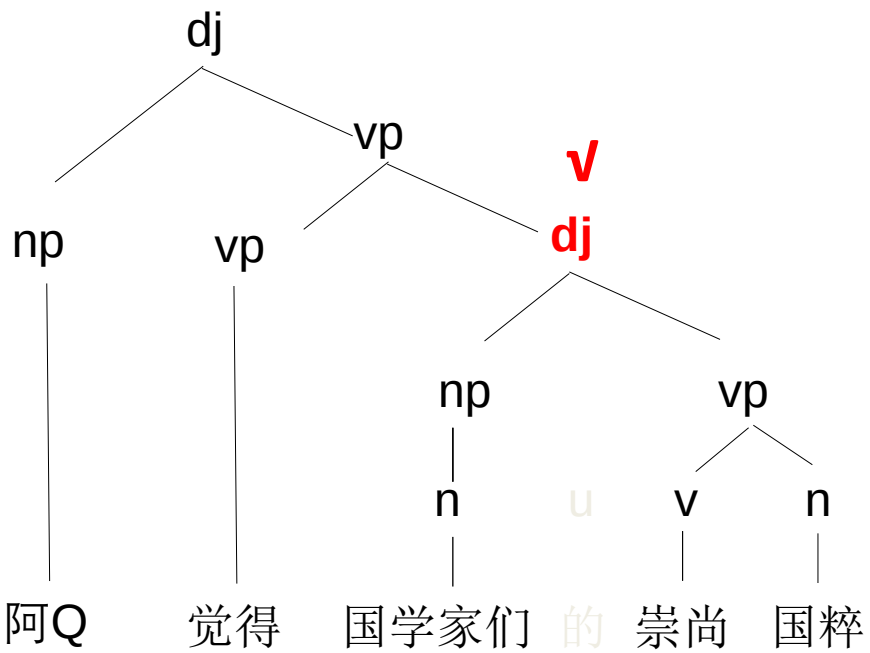


非中心扩展模式“np 的 vp”在宾语位置

阿Q 觉得 国学家们的**崇尚国粹** ×

阿Q 觉得 国学家们 **崇尚国粹** √

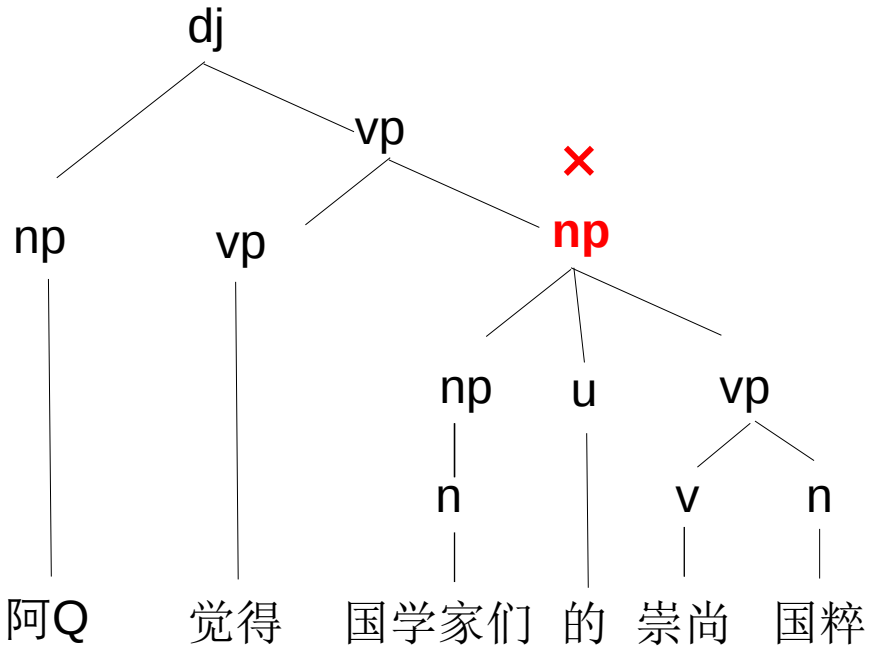


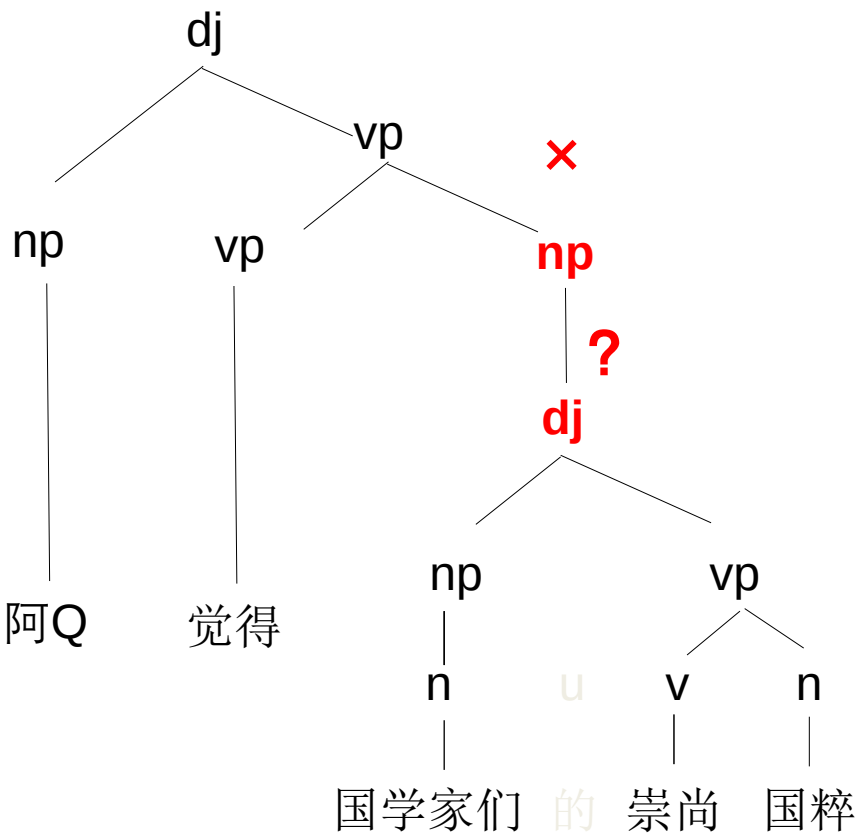


非中心扩展模式“np 的 vp”在宾语位置

阿Q 觉得 国学家们的**崇尚国粹** ×

阿Q 觉得 国学家们 **崇尚国粹** √





短语结构的歧义程度分析

从人的视角看歧义

- 鸡不吃了
- 咬死了猎人的狗
- 发现了敌人的哨兵
- 开刀的是他父亲
- 照片放大了一点儿
- 两个不同
- ……

从计算机的视角看歧义

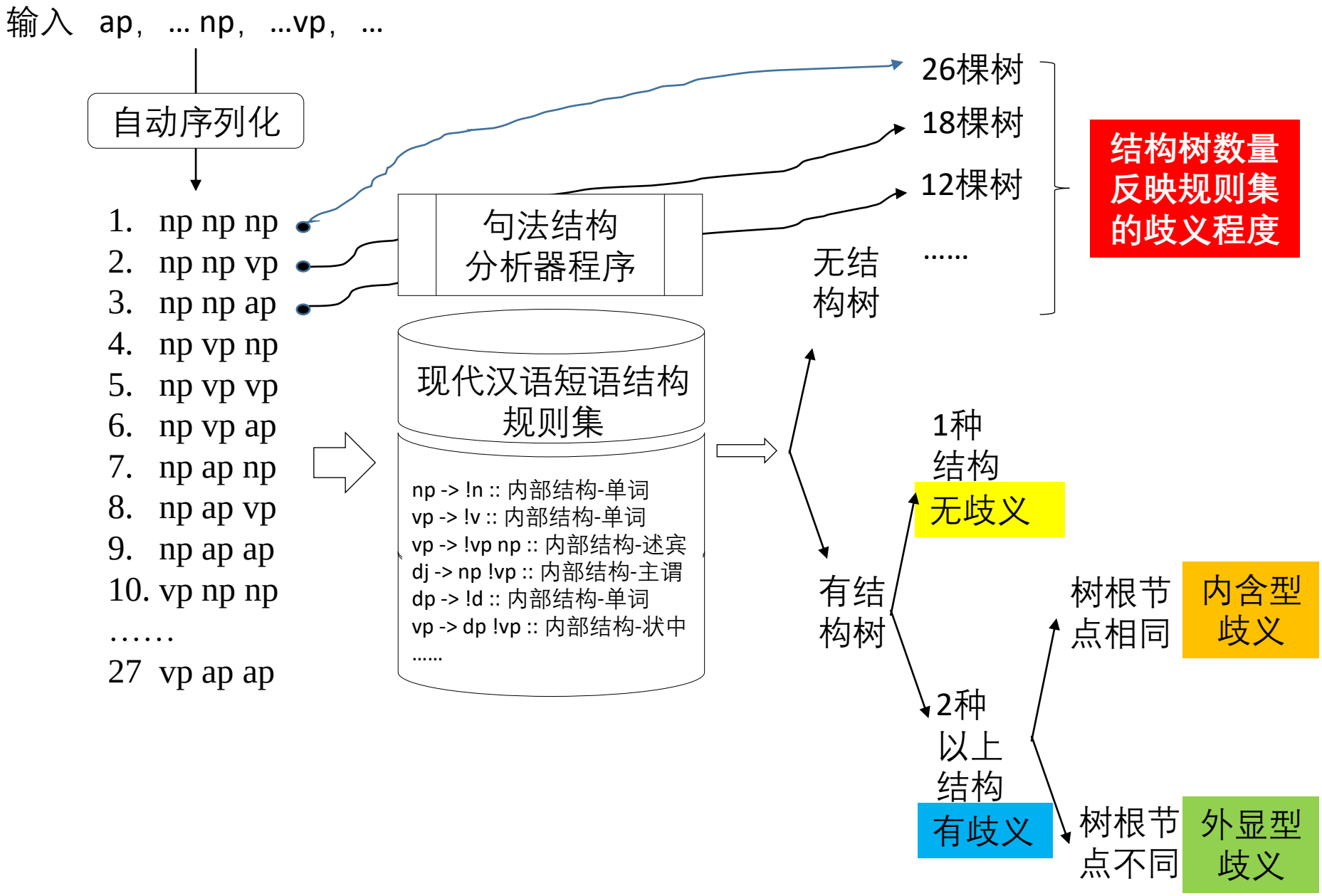
- 有多少歧义格式?
- 汉语组合模式的总体歧义度?
- 能否自动评估规则集的歧义度?
- ……

从树库抽取短语结构组合规则（形式文法）

编号	结构规则	频次	频率	累计频率
1	np -> !n	159778	0.121182749	0.121182749
2	vp -> !v	146250	0.110922511	0.23210526
3	vp -> !vp np	65751	0.049868486	0.281973746
4	dj -> np !vp	58724	0.044538896	0.326512642
5	dp -> !d	57281	0.043444461	0.369957102
6	vp -> dp !vp	50315	0.038161136	0.408118238
7	START -> zj	48956	0.03713041	0.445248648
8	np -> !rn	46171	0.035018142	0.48026679
9	np -> np !np	41435	0.031426149	0.511692939
10	ap -> !a	33370	0.025309294	0.537002233
11	mp -> !m	28331	0.021487492	0.558489725
12	vp -> !vp vp	25310	0.019196231	0.577685955
13	zj -> !dj wfs	20348	0.015432829	0.593118785
14	qp -> mp !q	19584	0.014853378	0.607972162
15	zj -> !fj wfs	19335	0.014664525	0.622636687
16	vp -> pp !vp	17959	0.013620905	0.636257592
17	np -> qp !np	17261	0.013091511	0.649349103
18	np -> np ude1 !np	13765	0.010439989	0.659789092
19	tp -> !t	12300	0.009328868	0.669117959
20	pp -> !p np	12158	0.009221168	0.678339128

……

基于树库文法考察短语结构歧义



基于树库文法考察短语结构歧义

ap dp mp np pp qp sp tp vp dj w

$11^3 = 1331$ 基于657条规则

规则集1

基于树库文法考察短语结构歧义

新增 np → !dj

ap dp mp np pp qp sp tp vp dj w

$11^3 = 1331$ 基于658条规则

规则集2

可能形成合法结构的排列:731个（54.92%）			不可能形成合法结构的排列:600个（45.08%）
ap ap ap ap ap mp ap ap np ap ap pp ……			ap ap dp ap ap w ap dp dp ap dp tp ……
有歧义的排列式:663个（49.81%）		无歧义的排列式:68个（5.11%）	dp ap dp dp ap pp dp ap sp dp ap tp ……
外显型歧义格式:569个（42.75%）	内含型歧义格式:94个（7.06%）	ap ap pp ap dp mp ap dp qp ap mp qp ap np sp ap np tp ……	pp ap dp pp ap pp pp ap sp ……
dj dj dj np dj dj dj np dj dj dj np ……	dj dj sp dj dj tp np dj sp np dj tp ……		

基于树库文法考察短语结构歧义

删除4条规则 ~~dj → dj !vp~~

新增 np → !dj

ap dp mp np pp qp sp tp vp dj w

$11^3 = 1331$ 基于654条规则

规则集3

可能形成合法结构的排列:731个（54.92%）		不可能形成合法结构的排列:600个（45.08%）
ap ap ap ap ap mp ap ap np ap ap pp ……		ap ap dp ap ap w ap dp dp ap dp tp ……
有歧义的排列式:663个（49.81%）	无歧义的排列式:68个（5.11%）	dp ap dp dp ap pp dp ap sp dp ap tp ……
外显型歧义格式:569个（42.75%）	内含型歧义格式:94个（7.06%）	pp ap dp pp ap pp pp ap sp ……
dj dj dj np dj dj dj np dj dj dj np ……	dj dj sp dj dj tp np dj sp np dj tp ……	ap ap pp ap dp mp ap dp qp ap mp qp ap np sp ap np tp ……

基于树库文法考察短语结构歧义

删除 ~~np → np !vp~~

新增 np → !vp

ap dp mp np pp qp sp tp vp dj w

$11^3 = 1331$ 基于657条规则

规则集4

可能形成合法结构的排列:672个（50.49%）			不可能形成合法结构的排列:659个（49.51%）
ap ap ap ap ap mp ap ap np ap ap pp ……			ap ap dp ap ap w ap dp dp ap dp sp ……
有歧义的排列式:561个（42.15%）		无歧义的排列式:111个（8.34%）	dp ap dp dp ap pp dp ap sp dp ap tp ……
外显型歧义格式:456个（34.26%）	内含型歧义格式:105个（7.89%）	ap ap pp ap dp mp ap dp qp ap mp qp ap np sp ap np tp ……	pp ap dp pp ap pp pp ap sp ……
vp vp vp vp vp ap vp ap vp vp vp np ……	np np dj np mp dj np np sp np np tp ……		

基于树库文法考察短语结构歧义

删除13条规则 ~~np → np !vp~~

新增 np → !vp

ap dp mp np pp qp sp tp vp dj w

$11^3 = 1331$ 基于645条规则

规则集5

可能形成合法结构的排列:672个（50.49%）		不可能形成合法结构的排列:659个（49.51%）
ap ap ap ap ap mp ap ap np ap ap pp ……		ap ap dp ap ap w ap dp dp ap dp sp ……
有歧义的排列式:561个（42.15%）	无歧义的排列式:111个（8.34%）	dp ap dp dp ap pp dp ap sp dp ap tp ……
外显型歧义格式:456个（34.26%）	内含型歧义格式:105个（7.89%）	pp ap dp pp ap pp pp ap sp ……
vp vp vp vp vp ap vp ap vp vp vp np ……	np np dj np mp dj np np sp np np tp ……	ap ap pp ap dp mp ap dp qp ap mp qp ap np sp ap np tp ……

基于树库文法考察短语结构歧义

删除39条规则 ~~np → np !vp~~

新增3条规则 np → !vp

ap dp mp np pp qp sp tp vp dj w

$11^3 = 1331$ 基于618条规则

规则集6

可能形成合法结构的排列:752个（56.50%）			不可能形成合法结构的排列:579个（43.50%）
ap ap ap ap ap mp ap ap np ap ap pp ……			ap ap dp ap ap w ap dp dp ap dp w ……
有歧义的排列式:714个（53.64%）		无歧义的排列式:38个（2.85%）	dp ap dp dp ap pp dp ap w dp dp w ……
外显型歧义格式:623个（46.81%）	内含型歧义格式:91个（6.84%）	ap dp sp ap dp tp ap w ap ap w np mp dp mp mp dp np ……	pp ap dp pp ap pp pp ap w ……
vp vp vp ap vp vp vp vp ap vp ap vp ……	dj dj sp dj dj tp np dj sp np dj tp ……		

基于树库文法考察短语结构歧义

ap dp mp np pp qp sp tp vp dj w

$11^3 = 1331$ 基于657条规则

vp可以作定中结构的中心语，dj可以作主语 **规则集1**

外显型歧义格式（共386个）	歧义指数	内含型歧义格式（共134个）	歧义指数
[1] vp vp vp	43	[1] dj vp vp	6
[2] vp vp ap	34	[2] np np dj	5
[3] vp ap vp	27	[3] dj vp ap	5
[4] np np np	26	[4] np mp dj	4
[5] vp ap ap	24	[5] np np sp	4
……		……	
[385] dj sp np	2	[133] dj dj vp	2
[386] dj tp np	2	[134] dj dj dj	2
平均歧义数	6.61	平均歧义数	2.29

基于树库文法考察短语结构歧义

新增 np → !dj

ap dp mp np pp qp sp tp vp dj w

$11^3 = 1331$ 基于658条规则

vp可以作定中结构的中心语，dj不能作主语 **规则集2**

外显型歧义格式（共569个）	歧义指数	内含型歧义格式（共94个）	歧义指数
[1] dj dj dj	112	[1] dj dj sp	8
[2] np dj dj	82	[2] dj dj tp	8
[3] dj np dj	76	[3] np dj sp	6
[4] dj dj np	72	[4] np dj tp	6
[5] tp dj dj	67	[5] dj np sp	6
……		……	
[568] vp sp tp	2	[93] dj tp mp	2
[569] vp w vp	2	[94] dj tp qp	2
平均歧义数	15.57	平均歧义数	2.62

基于树库文法考察短语结构歧义

删除4条规则 ~~dj → dj !vp~~

新增 np → !dj

ap dp mp np pp qp sp tp vp dj w

$11^3 = 1331$ 基于654条规则

vp可以作定中结构的中心语，dj不能作主语 **规则集3**

外显型歧义格式（共569个）	歧义指数	内含型歧义格式（共94个）	歧义指数
[1] dj dj dj	112	[1] dj dj sp	8
[2] np dj dj	82	[2] dj dj tp	8
[3] dj np dj	76	[3] np dj sp	6
[4] dj dj np	72	[4] np dj tp	6
[5] tp dj dj	67	[5] dj np sp	6
……		……	
[568] vp w vp	2	[93] dj tp mp	2
[569] dj w vp	2	[94] dj tp qp	2
平均歧义数	14.83	平均歧义数	2.60

基于树库文法考察短语结构歧义

删除 ~~np → np !vp~~

新增 np → !vp

ap dp mp np pp qp sp tp vp dj w

$11^3 = 1331$ 基于657条规则

vp不能在np之后作中心语 **规则集4**

外显型歧义格式（共456个）	歧义指数	内含型歧义格式（共105个）	歧义指数
[1] vp vp vp	304	[1] np np dj	5
[2] vp vp ap	137	[2] np mp dj	4
[3] vp ap vp	116	[3] np np sp	4
[4] vp vp np	114	[4] np np tp	6
[5] np vp vp	105	[5] qp ap dj	4
……		……	
[455] dj tp np	2	[104] dj dj np	2
[456] dj vp pp	2	[105] dj dj dj	2
平均歧义数	14.28	平均歧义数	2.24

基于树库文法考察短语结构歧义

删除13条规则 ~~np → np !vp~~

新增 np → !vp

ap dp mp np pp qp sp tp vp dj w

$11^3 = 1331$ 基于645条规则

vp不能作定中结构中心语 **规则集5**

外显型歧义格式（共456个）	歧义指数	内含型歧义格式（共105个）	歧义指数
[1] vp vp vp	304	[1] np np dj	5
[2] vp vp ap	137	[2] np mp dj	4
[3] vp ap vp	116	[3] np np sp	4
[4] vp vp np	114	[4] np np tp	6
[5] np vp vp	105	[5] qp ap dj	4
……		……	
[455] dj tp np	2	[104] dj dj np	2
[456] dj vp pp	2	[105] dj dj dj	2
平均歧义数	13.95	平均歧义数	2.24

基于树库文法考察短语结构歧义

删除39条规则 ~~np → np !vp~~

新增3条规则 np → !vp

ap dp mp np pp qp sp tp vp dj w

$11^3 = 1331$ 基于618条规则

ap、vp不能作定中结构中心语、主语，dj不能作主语 **规则集6**

外显型歧义格式（共623个）	歧义指数	内含型歧义格式（共91个）	歧义指数
[1] vp vp vp	354	[1] ap dj sp	8
[2] ap vp vp	280	[2] ap dj tp	8
[3] vp vp ap	278	[3] dj dj sp	8
[4] vp ap vp	266	[4] dj dj tp	8
[5] ap ap vp	224	[5] ap np sp	7
……		……	
[622] vp sp pp	2	[90] dj tp np	2
[623] vp tp pp	2	[91] dj tp qp	2
平均歧义数	35.24	平均歧义数	3.27

基于树库文法考察短语结构歧义

从规则集1到规则集6，各种尝试将谓词性成分直接上升为np的做法，都导致歧义的显著增加

规则集	规则数量	语法特征	形式特征	外显歧义格式数量	外显歧义指数	内含歧义格式数量	内含歧义指数	歧义格式总数	歧义格式占比
规则集1	657	vp可以作定中结构的中心语；dj可以作主语	np → np !vp dj → dj !vp	386	6.61	134	2.29	520	39.07%
规则集2	658	同上；并且dj可以直接实现为np	跟规则集1相比，新增一条规则：np → !dj	569	15.57	94	2.62	663	49.81%
规则集3	654	同上；但dj不再直接作主语	跟规则集2相比，删去dj → dj !vp等4条dj作主语的规则	569	14.83	94	2.60	663	49.81%
规则集4	657	vp不能在np之后作定中结构中心语；vp可以直接实现为np	跟规则集1相比，新增 np → !vp 删去 np → np !vp	456	14.28	105	2.24	561	42.15%
规则集5	645	vp不能作定中结构的中心语；vp可以直接实现为np	跟规则集1相比，新增 np → !vp 删去 np → np !vp等13条vp作定中结构中心语的规则	456	13.95	105	2.24	561	42.15%
规则集6	618	vp、ap不能作定中结构中心语，不能作主谓结构主语；dj不能作主谓结构主语	跟规则集1相比，新增 np → !vp，np → !ap，np → !dj等3条规则 删去 vp、ap在定中结构中心语和主语位置的规则、删去dj在主语位置的规则（共计删除42条规则）	623	35.24	91	3.27	714	54.63%

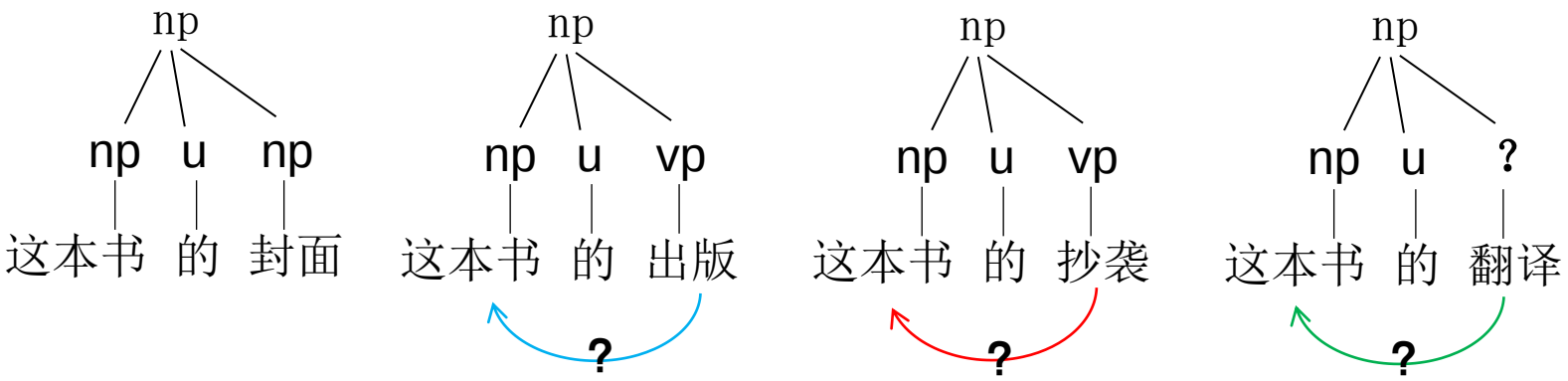
“np 的 vp” 不是“np 的 np”

结构树是区分语句含义的有效手段

np 的 vp

np 的 np

把vp跟np区别开，更便于观察和解释 “np 的 vp”结构中，np和vp的不同语义关系



真正需要思考的问题：

（1） 这本书的出版 → ____ 出版 这本书

（2） 这本书的抄袭 → 这本书 抄袭 ____

（3） 这本书的翻译 → ___ 把 这本书 翻译(成)____

___ 翻译 这本书