句法结构歧义的程度

詹卫东

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

句法结构歧义的程度：两个考察角度

给定文法规则集

考察角度1	输入：n^m 种**非终结符序列**（如：np ap vp） n是文法中的非终结符个数； m是格式中的符号个数； 输出：所有格式的歧义状况报告
考察角度2	输入：句子（**终结符序列**） 输出：句子结构分析结果的个数

考察角度1：对非终结符序列进行结构分析

np np np

np np vp

np np ap

np vp np

np vp vp

np vp ap

np ap np

np ap vp

np ap ap

vp np np

……

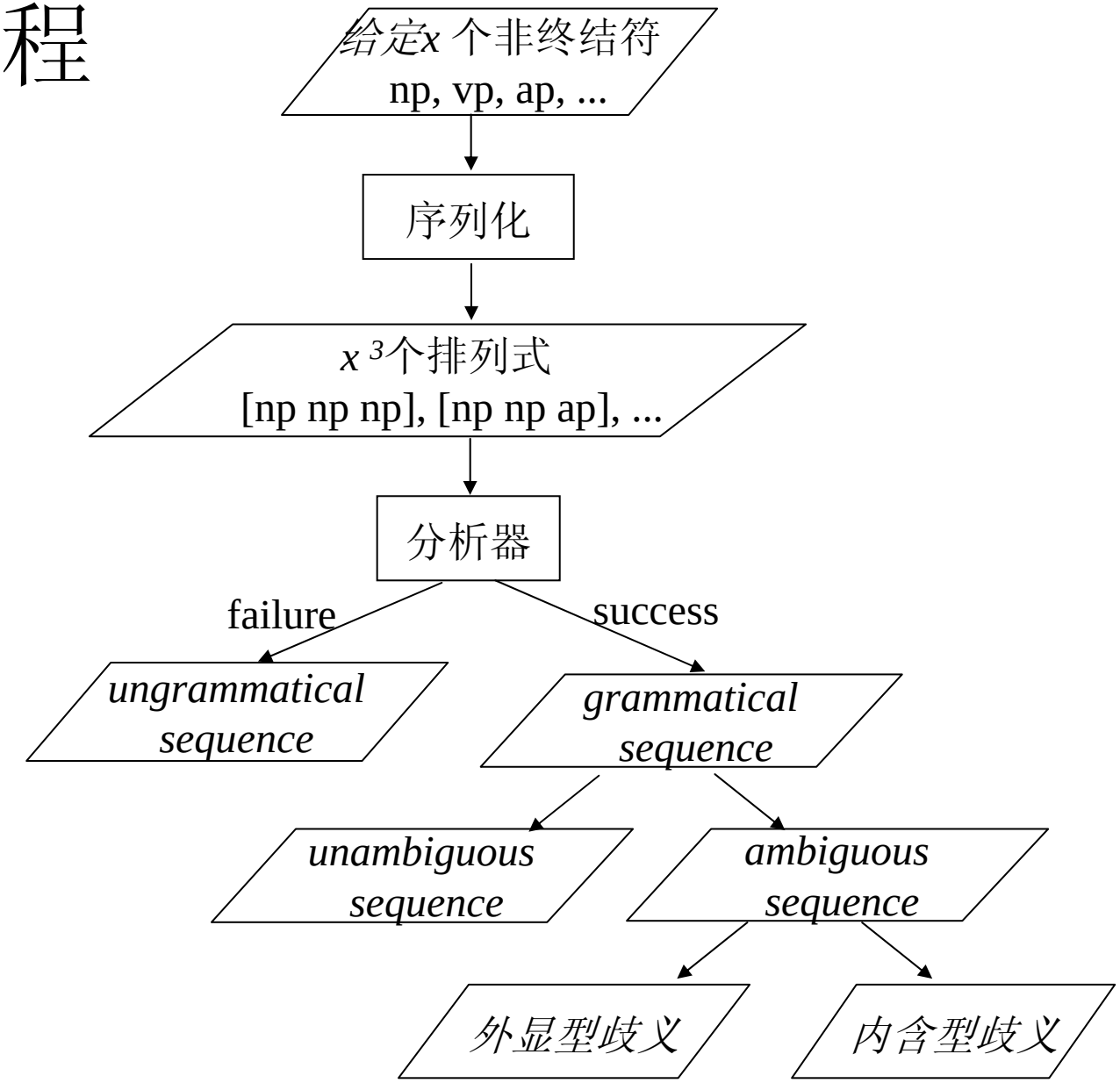
这**27**个非终结符序列中：

- 哪些格式有潜在歧义？
- 是外显型歧义还是内含型歧义？
- 每个有潜在歧义的格式歧义程度如何？

以 np，vp，ap三个非终结符的排列为例

詹卫东、常宝宝、俞士汶，汉语短语结构定界歧义类型分析及分布统计，《中文信息学报》1999年第3期

考察流程


给定x 个非终结符
np, vp, ap, ...
序列化
x 3个排列式
[np np np], [np np ap], ...
分析器
failure
success
ungrammatical sequence
grammatical sequence
unambiguous sequence
ambiguous sequence
外显型歧义
内含型歧义

np, tp, sp, mp, ap, dp, pp, vp, dj

$9^3 = 729$

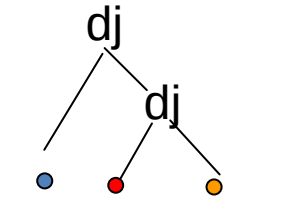
可能形成合法结构的排列:369个			不可能形成合法结构的排列:360个
np np np np np mp np np tp np np sp ……			np np dp np np pp np mp sp np mp dp ……
有歧义的排列式:285个		无歧义的排列式:84个	dj mp mp dj mp tp dj mp sp dj mp dp ……
外显型歧义格式:194个	内含型歧义格式:91个	np mp np np mp tp np mp dj np ap dj ……	pp tp sp pp tp dp pp tp pp
np np np np np ap np np vp np vp vp ……	np np mp np np tp np np sp np np dj ……		

外显型歧义格式（共 194 个）	歧义指数	内含型歧义格式（共 91 个）	歧义指数
[1] vp vp vp	43	[1] vp ap np	5
[2] vp vp ap	34	[2] dj vp vp	5
[3] vp ap ap	25	[3] np sp dj	4
……		……	
[194] pp sp vp	2	[91] pp pp pp	2
平均歧义数	6.55	平均歧义数	2.37

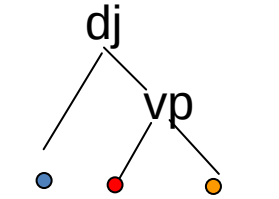
歧义格式示例

[1]	(dj:主谓(np,dj:主谓(np,vp)))	这事　校长　知道
[2]	(dj:主谓(np,vp:状中(np,vp)))	校长　现场　办公
[3]	(vp:状中(np,vp:状中(np,vp)))	?
[4]	(dj:主谓(dj:主谓(np,np),vp))	?
[5]	(dj:主谓(np:定中(np,np),vp))	奶油 面包 买不到
[6]	(vp:状中(np:定中(np,np),vp))	?
[7]	(dj:主谓(np:联合(np,np),vp))	眉毛 胡子 一把抓
[8]	(vp:状中(np:联合(np,np),vp))	?

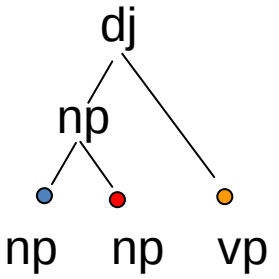
……



[1]



[2]



[5，7]

考察角度2：对终结符序列进行结构分析

基于简单CFG文法分析句子结构，可能产生的歧义结构的数量：Catalan number

Catalan number的计算公式

$$C_n = \frac{1}{n+1}\begin{bmatrix}2n\\ n\end{bmatrix} = \frac{1}{n+1}\times\frac{(2n)!}{n!(2n-n)!} = \frac{(2n)!}{(n+1)(n!)^2} = \frac{(2n)!}{(n+1)!\,n!}$$

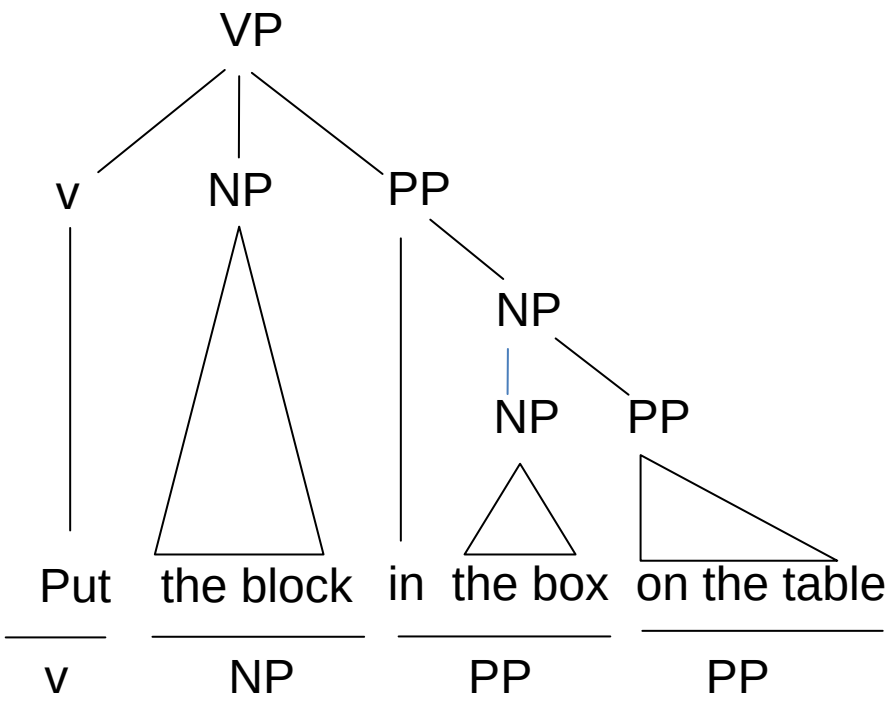
Church, K.W. & Patil, R. 1982, Coping with syntactic ambiguity (or How to put the block in the box on the table), American Journal of Computational Linguistics,8(3-4),pps.139-149.

Catalan number

- n=2 (2*2)!/(2+1)!*2! = 4! / 3!*2! = 2
- n=3 (2*3)!/(3+1)!*3! = 6! / 4!*3! = 5
- n=4 (2*4)!/(4+1)!*4! = 8! / 5!*4! = 14
- n=5 (2*5)!/(5+1)!*5! = 1(! / 6!*5! = 42
- n=6 (2*6)!/(6+1)!*6! = 12! / 7!*6! = 132
- n=7 (2*7)!/(7+1)!*7! = 14! / 8!*7! = 429
- n=8 (2*8)!/(8+1)!*8! = 16! / 9!*8! = 1430
- n=9 (2*9)!/(9+1)!*9! = 18! / 10!*9! = 4862
- …

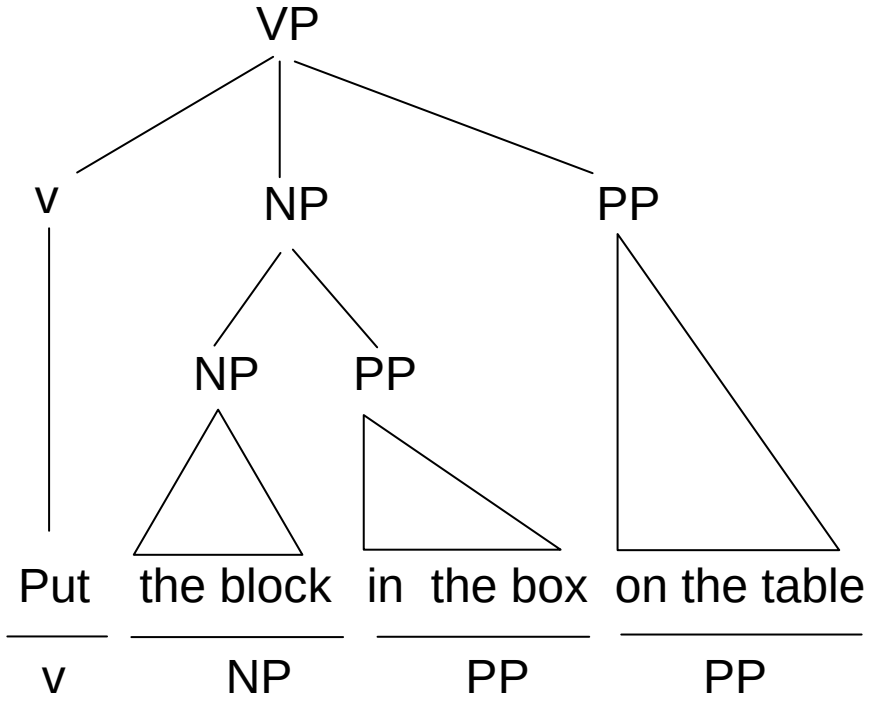
pp-attachment歧义实例

pp 个数为2



I

把积木放进桌上的盒子里



II

把盒子里的积木放到桌上

Put	the block	in the box	on the table	in the kitchen
v	np	**pp**	**pp**	**pp**
v	n	**p n**	**p n**	**p n**

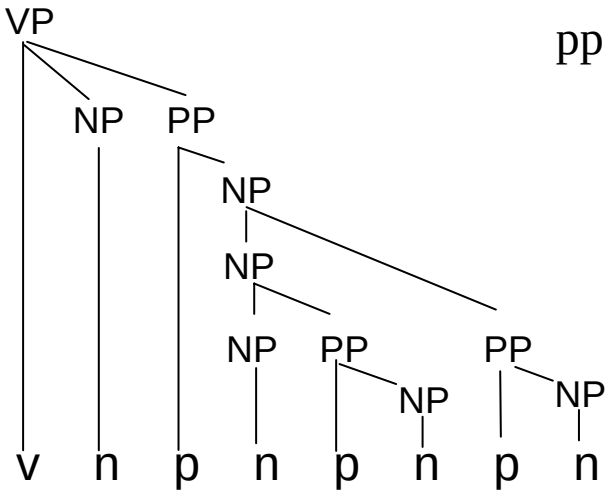
NP -> NP PP

NP -> n

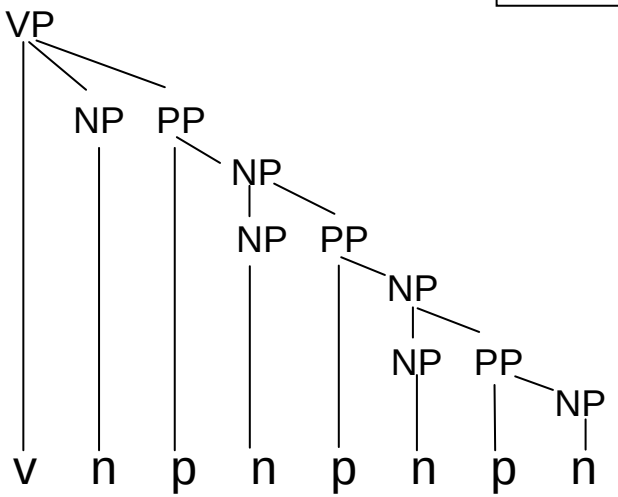
PP -> p NP

VP -> v NP PP

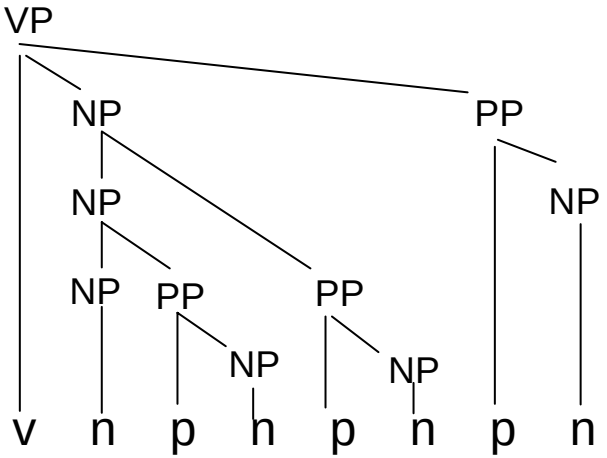
pp 个数为3



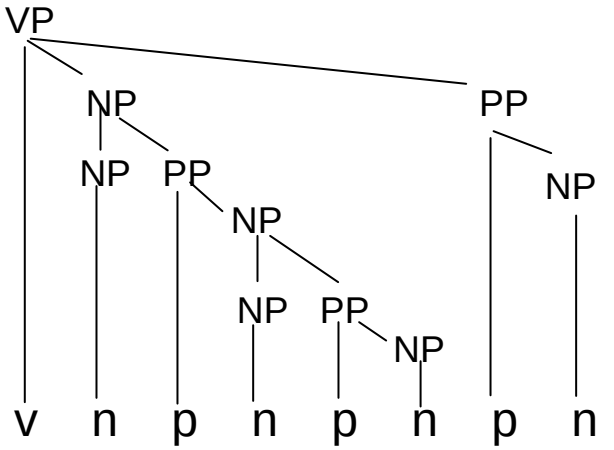
A



B



C



D

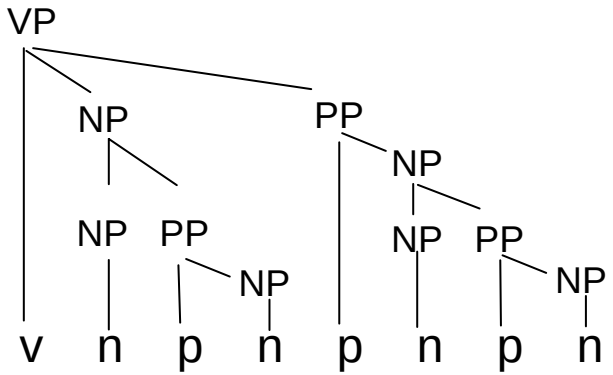
Put	the block	in	the box	on the table	in	the kitchen
v	np	**pp**		**pp**		**pp**
v	n	**p**	**n**	**p**	**n**	**p**	**n**

NP -> NP PP

NP -> n

PP -> p NP

VP -> v NP PP



E

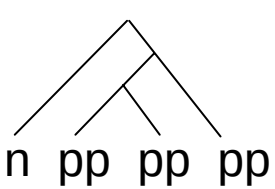
A 把积木放进盒子里（盒子在桌上，在厨房）

B 把积木放进厨房桌上的盒子里

C 把积木放到厨房（积木在盒子里，在桌上）

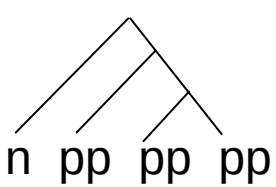
D 把桌上盒子里的积木放到厨房

E 把盒子里的积木放到厨房桌上



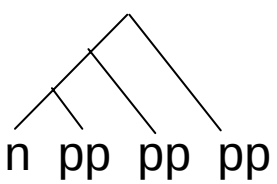
A

(n ((pp pp) pp))



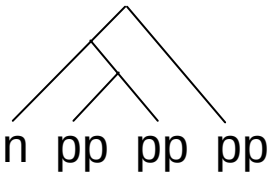
B

(n (pp (pp pp)))



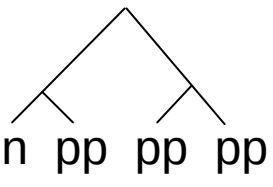
C

(((n pp) pp) pp)



D

((n (pp pp)) pp)



E

((n pp) (pp pp))

关于Catalan数计算公式的说明

n 个左括号跟n个右括号排列成2n项，在任意位置，左括号数不少于右括号数。这样的排列式的个数为Catalan数。

$$\binom{2n}{n} - \binom{2n}{n-1} = \frac{(2n)!}{n!\times n!} - \frac{(2n)!}{(n-1)!(n+1)!}$$

$$= \frac{1}{n+1}\binom{2n}{n}$$

Donald E.Knuth著 苏运霖译，《计算机程序设计艺术》（第三版）第一卷。508页。国防工业出版社。

n = 3

√	1	((( )))	11	) ((( ))
√	2	(( ) ( ))	12	) (( ) ( )
√	3	(( )) ( )	13	) (( )) (
	4	(( ))) (	14	) ( ) (( )
√	5	( ) (( ))	15	) ( ) ( ) (
√	6	( ) ( ) ( )	16	) ( )) ((
	7	( ) ( )) (	17	)) ((( )
	8	( )) (( )	18	)) (( ) (
	9	( )) ( ) (	19	)) ( ) ((
	10	( ))) ((	20	))) (((

$C_3 = 20 - 15 = 5$

关于Catalan数计算公式的说明（续）

条件1：左右括号数相等
条件2：在任意位置，左括号数不少于右括号数

- 满足第1个条件的序列个数为：$\binom{2n}{n}$
- 违背第2个条件的序列记作S
  （1）设在序列S的i位置，右括号数多于左括号数；
  （2）将i位置的右括号换成左括号；
  （3）从i位置依次往左，将括号方向“反转”，得到S’序列；
  （4）结果：S’序列中左括号数为n+1，右括号数为n-1
- 对S’序列，i位置左边（含i位置）的左括号数比右括号数多1；
  （1）将i位置左边的所有括号方向都“反转”，即恢复为S序列；
  （2）S’ 与 S 有一一对应关系。S’的个数就是S的个数。
- S’的个数是：在2n个位置上，放置n+1（或n-1）个左括号（或右括号）的可能的个数，即$\binom{2n}{n+1}$ 或 $\binom{2n}{n-1}$。

n = 3

0 表示左括号，1表示右括号

√	1	0 0 0 1 1 1	11	0 0 0 0 1 1
√	2	0 0 1 0 1 1	12	0 0 0 1 0 1
√	3	0 0 1 1 0 1	13	0 0 0 1 1 0
	4	0 1 0 0 0 1	14	0 0 1 0 0 1
√	5	0 1 0 0 1 1	15	0 0 1 0 1 0
√	6	0 1 0 1 0 1	16	0 0 1 1 0 0
	7	0 1 1 0 0 1	17	0 1 0 0 0 1
	8	0 1 1 0 0 1	18	0 1 0 0 1 0
	9	0 1 1 0 1 0	19	0 1 0 1 0 0
	10	0 1 1 1 0 0	20	0 1 1 0 0 0

$C_3 = 20 - 15 = 5$