现代汉语“名+名+名”组合的统计分析

王东波[1]，陈锋[2]

（南京师范大学文学院 04301 信箱，南京 210046；南京师范大学文学院 04301 信箱，南京 210046）

摘　要： 本文考察语料库中“名+名+名“组合。希望可以得到该组合在句法方面的统计规律，以期有利于自动句法分析和歧义类型的消解；首先从大规模已标注词性的语料库中提取“名+名+名”组合，统计出了 125 类“名+名+名”组合的出现次数，并且从这些组合中区分了合法组合和非法组合，然后重点分析了九类高频的“名+名+名”组合。从组合的合法性、组合的结构关系和组合的歧义类型三个方面详尽地分析了九类高频的“名+名+名”组合；分析结果表明：九类组合在句法上都有一种或一种以上的歧义，除“n+nr+nr”、“nr+nr+n”两组合外，歧义大多是句法层次切分不一致造成的；“n+nr+nr”、“nr+nr+n”、“ns+nz+n”、“nt+n+n”、“nz+n+n”语法结构关系比较单一，“n+n+n”的出现频率高，且语法结构较复杂需要进一步的研究。

关键词： “名+名+名”组合；语料库；歧义类型

Statistical Analysis on the"n+n+n" Phrases of Contemporary Chinese

Wang Dongbo[1], Chen Feng[2]

(1.Nanjing Normal University,Nanjing 210046; .2.Nanjing Normal University,Nanjing 210046)

Abstract: This article observed the "n+n+n" phases in corpus,in order to find statistical rule in syntax of Contemporary Chinese and make it available to automatically analyse the syntax and eliminate ambiguity; This article extractived the "n+n+n" phases from large and tagged corpus，calculated the number of each 125 types of "n+n+n" phases and told the differences between legal phases and illegal phases. From these statistical data ,this article observed the legality,structural relation and ambiguous types of 9 types of the "n+n+n" phrases which are high in frequency. The statistical result indicates that there is one or more ambiguity which is caused by different divided point in syntax except the phases of “n+nr+nr”and “nr+nr+n “in 9 types of the "n+n+n" phrases. There is simply structural relation in “n+nr+nr”,”nr+nr+n”, “ns+nz+n”, “nt+n+n” and “nz+n+n” , but the high frequent “n+n+n” is complicated in structural relation which is needed to more research.

Keywords: “n+n+n” prase; corpus; ambiguous type

1　引言

陆俭明先生《由指人的名词自相组合造成的偏正结构》一文中详细分析了“爸爸的爸爸的爸爸”这一结构的正确切分方式，本文在陆先生这篇文章的启发下，基于语料库的基础上全面考察了“名+名+名”这一组合。语料是北京大学计算语言学研究所的标注语料。本文所指的名词包括词（n）、地方（ns）、人名（nr）、机构团体（nt）

作者简介：王东波(1981-)，男，山东菏泽，南京师范大学文学院应用语言学专业 05 级硕士研究生。E-mail:jisuanyuyan@163.com
陈锋　(1982-)，男，山东滕州，南京师范大学文学院应用语言学专业 05 级硕士研究生 E-mail:ourchenfeng@hotmail.com

和其他专名（nz）五大类，本文所讨论的“名+名+名”组合指的是不含结构助词的三个名词的组合。

2 组合抽取和数据统计

本文用程序统计出了 125 类“名+名+名”组合的出现次数，具体数据如表 1 所示。

表 1 125 类“名+名+名”组合的数据

Table 1 The data of 125 types of “n+n+n”

	n	nr	ns	nt	nz
n+n	29714	6077	549	34	184
nr+n	1076	79	9	0	3
ns+n	13015	2945	528	24	98
nt+n	1036	1850	5	4	0
nz+n	2098	116	125	3	33
n+nt	169	1	0	3	0
nr+nt	21	0	1	0	0
ns+nt	135	5	2	0	1
nt+nt	17	0	0	0	0
nz+nt	3	0	0	0	0
n+nr	502	16920	13	5	9
nr+nr	7559	6308	268	22	1
ns+nr	82	310	3	0	0
nt+nr	2	59	3	0	0
nz+nr	5	5	0	0	0
n+nz	336	3	5	0	8
nr+nz	1	0	1	0	0
ns+nz	2001	1	13	11	53
nt+nz	6	0	0	0	0
nz+nz	85	1	1	0	5
n+ns	1310	4	239	31	56
nr+ns	64	5	164	0	18
ns+ns	2753	172	1022	26	254
nt+ns	136	0	15	0	0
nz+ns	71	0	5	0	0

注：表格内的数字为该组合在语料库中的出现次数

3 统计数据的分析

3.1 数据的初步分析

从统计的数据看，“n+n+n”是出现最多的一类短语组合，在整个组合占到 29%，而“n+n+n”、“nr+n+n”、“ns+n+n”、“nt+n+n”、“nz+n+n”、“ns+nz+n”、“nr+nr+n”、“n+ns+n”、“ns+ns+n”、“n+n+nr”、“ns+n+nr”、“nt+n+nr”、“n+nr+nr”、“nr+nr+nr”、“ns+ns+ns”这十五类组合共占了总量的 96.18%。而“nz+ns+nt”、“nz+ns+nz”、“nt+nt+nt”、“nt+nt+nz”、“nt+nr+nt”、“nt+nr+nz”等 39 个组合出现次数都是 0，其他的出现次数都在 600 次以下。

3.2 合法组合和组合的歧义类型

本文的合法组合是从层次分析的角度看能彼此构成句子直接组成成分和语义上能够搭配的组合。所谓非法组合是从层析分析的角度看彼此不能构成直接组成成分关系和语义上不能搭配的组合。通过统计发现在选取的十五类高频（出现次数大于1000）的组合中“n+n+nr”、“nr+n+n”、“ns+n+nr”、“ns+n+nr”、“nt+n+nr”、“n+ns+n”等为非法组合。

对于歧义类型，本文主要从定界歧义和结构关系歧义来看组合短语结构歧义。所谓定界歧义，也就是短语结构的层次切分歧义。层次切分歧义通常会伴随着结构关系歧义。而所谓结构关系歧义，则是指两个成分发生组合能以不同的关系形成一个组合体。

本文主要对 “n+n+n”、“nr+n+n”、“ns+n+n”、“nt+n+n”、“nz+n+n”、“ns+nz+n”、“nr+nr+n”、“n+ns+n”、“ns+ns+n”、“n+n+nr”、“ns+n+nr”、“nt+n+nr”、“n+nr+nr”、“nr+nr+nr”、“ns+ns+ns”等十五类高频组合进行详细的分析。

对于合法组合具体分析如下：

（1）n+nr+nr 根据从语料库中抽出的“n+nr+nr”分析，其主要是同位关系，例如“记者/n 何/nr 伟/nr”、“同伙/n 陈/nr 伟 nr”。普通名词（n）大多表示职务角色，如“主任”、“妻子”、“嫌疑人”等。

（2）nr+nr+n 由于“nr+nr”组合成一个完整的人名，所以“nr+nr+n”基本上全部是合法的组合，大部分是同位结构，例如“高/nr 铁/nr 同志/n”，同时有一部分定中结构例如“雷/nr 锋/nr 精神/n”。

（3）n+n+n 这一组合是“名+名+名”组合中出现最多的。结构关系也最复杂，数量最多的是定中关系，例如“工商/n 银行/n 队/n”、“石油/n 化工/n 总公司/n”；其次是联合关系，如 “区/n 地/n 县/n”、“省/n 市/n 区/n”等。

（4）ns+n+n 这三个名词连接成的组合基本上是定中关系，例如“中国/ns 服装/n 服饰/n”、“南京/ns 大学/n 出版社/n”等。由于定中关系的切分点不同，歧义结构大量存在，例如“中国/ns 人口/n 新闻奖/n”、“北京/ns 图书/n 订货会/n”等。

（5）ns+ns+n 在统计的语料中，这一组合都是合法的，结构关系主要集中在定中关系上，例如“西藏/ns 那曲/ns 地区/n”、“美国/ns 宾夕法尼亚/ns 大学/n”等。

（6）ns+ns+ns 由于中文表达几个地点相连时大都是有层次的，所以这一组合是合法的，并且大部分是定中结构，例如“江西省/ns 乐平市/ns 礼林镇/ns”；少量的联合结构，如“大连/ns 厦门/ns 南京/ns”。

（7）ns+nz+n 这一组合结构关系全部是定中关系，例如“北太平庄/ns 京京/nz 肉食厂/n”、“山东/ns 金贵/nz 酒厂/n”等。

（8）nz+n+n 这一组合基本上全部是合法的组合，在结构关系上都是定中结构,例如“长虹/nz 科技/n 公司/n”、“创佳/nz 电子/n 有限公司/n”等，由于“nz”是专有名词，因此在层次关系上切分不一致的问题不多，所以歧义结构很少。

（9）nt+n+n 这一组合基本上都是合法的组合，在结构关系上都是定中结构，例如“国务院/nt 新闻/n 办公室/n”、“北京大学/nt 国际/n 关系/n”等，本组合在层次关系上切分不一致的问题很少。

总之，合法组合的结构关系可以组成联合关系、同位关系、主谓关系、定中关系等各种类型，并且定中关系是出现最多的一种关系，而“nz+n+n”、“ns+nz+n”、“ns+ns+n”、“ns+n+n”、“nt+n+n”基本上只有定中关系一种。

在选取的这九类合法的组合中全部有歧义，“nz+n+n”和“nt+n+n”的歧义数量比较少。歧义类型大多是切分不一致造成的，只有“n+nr+nr”、“nr+nr+n”这两类有一部分歧义是由于结构关系不同造成的。

4 结束语

本文所分析的“名+名+名”，只不过是“np+np+np”中最简单的一种形式。但对于所用语料库也有一些要解决的问题，如对于人名是否可以不拆分，而作为一个分词单位来处理，这样也许对提高短语和句子的自动识别有帮助。同时对普通名词(n)进行更详细的分类，如分化出称呼、职称和职业等，这样有助于详细地考察“名+名+名”这一结构，从而也有助于短语和句法的自动分析。

参考文献：

[1]陈小荷.现代汉语自动分析——Visual C++实现[M],北京:北京语言文化大学出版社,2000.3.P132-134

[2]戴海胜,杨波,颜伟.现代汉语“名+名”组合的统计考察[A].第二届全国学生计算语言学研讨会论文集[C].北京:北京语言文化大学,2004.8 P.163-165.

[3]黄伯荣,廖序东.现代汉语[M].第二版,北京:高等教育出版社,2002.12.P59-69.

[4]詹卫东.面向中文信息处理的现代汉语短语结构规则研究[M],北京:清华大学出版社出版,2000.P118-128.

[5]陆俭明.由指人的名词自相组合造成的偏正结构[J].中国语言学报,1985,第二期.

[6]王珏.现代汉语名词研究[M],上海:华东师范大学出版社,2001.1.P236-247.

[7]朱德熙.朱德熙文集（一）[M],北京:商务印书馆,1999.P107-197.