面向自动分词的三音节新词语构词法研究

徐艳华
鲁东大学 文学院，烟台 264025
E-mail:ysyh0401@163.com

摘　要：未登录词识别是汉语分词处理中的一个难点。在大规模中文文本的自动分词处理中，未登录词是造成分词错误的一个重要原因。为了解决自动分词的这一“瓶颈”问题，我们对未登录词中的三音节新词语的结构进行了分析，总结了新词语的构词类序，发掘新词语的构词规律，以期为未登录词的识别和标注提供一套规则。

关键词：新词语；构词模式；自动分词

The Study of Three Syllable New Words' Structure for Automatic Segmentation

Xu Yanhua
College of Arts，Ludong University, Yantai 264025
E-mail: ysyh0401@163.com

Abstract: Unknown words can cause segmentation mistakes in the automatic word segmentation processing of Chinese texts. Meanwhile the recognition of unknown words is a difficult point in word segmentation processing. In order to solve the "bottleneck" problem arisen in automatic word segmentation, we analyze new words' word-formation patterns and explore their formation rules in order to provide a set of principles for recognition and marking of unknown words.

Keywords: new word; word-formation pattern; automatic segmentation

1　引言

现代汉语新词语是现代汉语词汇发展中最为活跃的部分。新词语在丰富了现代汉语词汇的同时，也对中文信息处理尤其是自动分词带来了不便。为了解决自动分词的这一“瓶颈”问题，我们以收集的三万多条新词语为基础，开发了一个新词语构词法信息库，详细描述了新词语构词法信息。在此基础上对各种结构类型的新词语进行详细地统计、归纳，总结新词语的构词规则。我们希望计算机能够利用这些规则，对新词语进行有效地识别。这些规则是在大规模语料库的基础上总结出来的，具有可信度、代表性，相信这样的成果，能够为计算机识别未登录词提供一个基础依据，对中文信息处理的发展起一定的推动作用。

2　新词语构词法信息库属性的确立

2.1　新词语构词法信息库属性信息

新词语构词法信息库旨在研究用何种构词材料以何种方式组合构成何种词，从而总结新词语的构词规律，为了满足这一方面的需要，在确立属性时主要考虑了以下几个方面：

1. 构词部件。要研究新词的构词法，就要从词的构造成分入手。词的构造成分，从标注的过程中可以发现有语素，也有词甚至短语。对于它们的区分，采用的是现代汉语中对三者的区分标准。从收集新词的过程中发现，双音节词居多，三音节、四音节次之，四音节以上的较少。本文只对三音节的新词进行研究，构词部件设立了三个，分别称为构件1、构件2。

2. 构词法信息。这一属性主要是根据词素与词素的结合情况确定新词语的构词方式。主要有主谓、动宾、联合、偏正、补充、前接、后加等形式。

3. 词性。汉语词的构成方式与词性并不是一一对应的，而是复杂交错的，一种构词方式可能构成几种词。

2.2 采用的词类体系及相关标记

本库的词性标注采用的是北京大学计算语言学研究所信息处理用的现代汉语词类体系及标记集的研究成果，具体如下：

1. 词类体系及标记

名词 n，动词 v，形容词 a，时间词 t，处所词 s，方位词 f，区别词 b，副词 d，状态词 z，代词 r，数词 m，量词 q，叹词 e，拟声词 o，介词 p，连词 c，助词 u，语气词 y。

以上是基本词类的标注，此外还有非词的语言单位标记

前接成分 h，后接成分 k，简称略语 j，人名 nr，地名 ns。

2. 词素类型

“词素”在本文中指一个词的构词部件，有的词素在汉语中可以独立运用，可以称作“词”，如“公开教学”中的“公开”和“教学”可以单独成词，但在这里，它是某个词的构词部件，所以都称作“词素”而不称作“词”。在标注的过程中，主要涉及到以下几类词素：

名词素 Ng，动词素 Vg，形容词词素 Ag，区别词词素 Bg，时间词词素 Tg，副词词素 Dg，方位词词素 Fg。

3. 短语类型

介词结构 pp，定中结构 dp，状中结构 zp，联合结构 lp，动宾结构 bp，补充结构 cp，主谓结构 wp，数量结构 mq。

4. 新词语的构词方式

联合式 L，动宾式 B，定中式 D，状中式 Z，主谓式 W，补充式 C，附加式（前缀 Q，后缀 H）。

3 三音节新词语构词方式研究

三音节新词语共 6502 个，占新词语总数的 21.4%，其中：

（1）定中式（D）4206 个，占三音节词语的 64.7%，共有 19 种构词模式：

Ng+Vg→N（共 21 个，占 0.5%）如：性侵犯、鱼采购、核扩散

PP+Ng→N（共 2 个，占 0.05%）如：对台戏、在室女

Ng+Ng→N（共 1800 个，占 42.8%）如：风筝城、国情车、核垃圾

Ag+Ng→N（共 753 个，占 17.9%）如：多层次、全方位、方便菜

DP+Ng→N（共 105 个，占 2.5%）如：白褶裙、大篷车、白眼病

Bg+Ng→N（共 72 个，占 1.7%）如：副班长、女强人、自动伞

Vg+Ng→N（共 845 个，占 20.1%）如：发射场、分手饭、按摩袜

BP+ Ng→N（共 286 个，占 6.8%）如：翻身仗、防寒服、扶贫款

ZP+Ng→N（共 101 个，占 2.4%）如：公休日、公用筷、优生法

LP+Ng→N（共 29 个，占 0.7%）如：港台剧、推拉车、吃喝风

NS+Ng→N（共 38 个，占 0.9%）如：中国风、欧洲军、印度风

WP+Ng→N（共 29 个，占 0.7%）如：自留山、人行道、手抛道

Ag+NR→N（共 2 个，占 0.05%）如：活雷锋、活鲁班

Fg+Ng→N（共 6 个，占 0.15%）如：地下军、掌上机、炉前工

M+Ng→N（共 46 个，占 1.1%）如：半成品、二老外、零缺陷

Ag+DP→N（共 13 个，占 0.3%）如：红大院、长防林、新税制

Ng+DP→N（共4个，占0.1%）如：核废料、车月票、泵排量
NR+Ng→N（共16个，占0.4%）如：琼瑶迷、雷锋卡、江青裙
定中式的三音节新词语全部是名词，在构词模式上，以Ng+Ng，Vg+Ng，Ag+Ng三类为主。
（2）状中式（Z）190个，占三音节词的2.9%，共有23种构词模式：
Ag+Vg→V（共72个，占38.4%）如：活读书、假出口、粗加工
Ag+Ag→A（共8个，占4.7%）如：穷大方、全自动、富小气
PP+Vg→V（共7个，占3.7%）如：朝钱看、往外挤、向右转
Fg+Vg→V（共3个，占1.6%）如：锅下愁、场外招、雪上飞
ZP+Vg→V（共1个，占0.5%）如：过劳死
BP+Vg （共11个，占5.8%）→ V（共10个，占99.9%）如：没戏唱、回头看
N（共1个，占0.1%）如：随身听
Vg+BP→V（共2个，占1.1%）如：会来事、伙种地
Ag+BP→V（共5个，占2.8%）如：假夺权、大换肩
Mg+Vg→V（共6个，占3.3%）如：半跳槽、二进宫、半残废
Dg+BP→V（共8个，占4.2%）如：不起眼、不走样、不称霸
Dg+Vg→V（共9个，占4.7%）如：不介入、不起诉、再教育
Vg+Vg→V（共22个，占11.6%）如：倒接班、倒发奖、热身唱
Ng+Vg→V（共22个，占11.6%）如：电捕鱼、双肩挑、口头纠
O+Vg→V（共2个，占1.1%）如：乓乓响、碰碰响
MQ+Vg→V（共3个，占1.6%）如：一刀砍、一刀切、一日游
Bg+Vg→V（共1个，占0.5%） 如：总动员
Ag+U+Vg→V（共2个，占1.1%）如：黑着干
Vg+U+Vg→V（共2个，占1.1%）如：走着瞧、对着干
Tg+Vg→V（共1个，占0.5%）如：生前葬
WP+Ag→A（共1个，占0.5%）如：自来红
CP+Vg→V（共1个，占0.5%）如：绑紧跳
D+Ag→A(共1个，占0.5%) 如：共同美
BP+Ag→A(共3个，1.6%) 如：上场慌、上场昏、上场怯
状中式的三音节词，以动词居多，形容词甚少，在结构模式上以Ag+Vg，Vg+Vg，Ng+Vg为主。
（3）动宾式（B）693个，占三音节词的10.7%，共有14种构词模式：
Vg+DP→V（共68个，占9.8%）如：翻老帐、赶热浪、走弯路
Vg+Ng→V（共575个，占83.1%）如：够意思、打电话、点码子
Vg+Ag→V（共4个，占0.6%）如：反腐败、够慈气、玩深沉
Vg+Vg→V（共12个，占1.8%）如：反分工、拉赞助、打埋伏
Vg+ZP→V（共4个，占0.6%）如：反冒进、放单飞、吃大富
Vg+MQ→V（共4个，占0.6%）如：顾一头、翻一番、送一程
Vg+U+Ng→V（共3个，占0.4%）如：挂了帅、乱了套、揭了壳
Vg+BP→V（共2个，占0.3%）如：反跳槽
Vg+J→V（共3个，占0.4%）如：除六害
CP+Ng→V（共9个，占1.3%）如：拖下水、拉下马、吃错药
Vg+Tg→V（共2个，占0.3%）如：赌明天、抢农时

Vg+Ag+U→V（共1个，占0.15%）如：说白了

ZP+Ng→V（共4个，占0.6%）如：不信邪、不懂电、不进鳞

Vg+Q+Ng→V（共1个，占0.15%）如：打把伞

以动宾方式构成的三音节词全部是动词，在构词模式中，Vg+Ng占绝对优势，Vg+DP次之。

（4）主谓式（W）104个，占三音节词的1.6%，共有11种构词模式：

Ng+Ag→A(共32个，占30.8%) 如：根子正、性开放

Ng+Vg+Ng→V(共31个，占29.8%) 如：房改房、利改税

Ng+Vg （共24个，占23.1%）

→ V（共16个，占66.7%）如：核辐射、肠梗阻

↘ N（共8个，占33.3%）如：同窗恋、全球通

Ng+ZP→A(共4个，占3.8%) 如：性早熟、心太软

Mg+Ng+Ag→A(共2个，占3.8%) 如：一头热、一头沉

Vg+Ag→A(共3个，占2.9%) 如：抓药难、劳动美、开门红

R+Vg→V(共1个，占1%) 如：大家拿

Bg+Vg+Bg→V(共1个，占1.9%) 如：单改双、专转本

Fg+Vg+Fg→V(共1个，占1.9%) 如：内转外

M+Vg+Mg→V(共1个，占1.9%) 如：二过一

M+Ng+Vg→V(共1个，占1.9%) 如：两手抓

该类型动词最多，形容词次之，名词最少，以Ng+Ag，Ng+Vg+Ng，Ng+Vg三种结构模式为主。

（5）并列式（L）50个，占三音节词的0.8%，共有7种构词模式：

Ag+Ag+Ag→A(共19个，占39.6%) 如：高精尖、快准狠、脏乱差

Vg+Vg+Vg→V(共9个，占18.8%) 如：产运销

Ng+Ng+Ng→N(共8个，占16.7%) 如：山江湖、责权利、人财物

Ng+Ng→N(共7个，占9.4%) 如：葱韭菜、工副业

Ag+C+Ag→A(共5个，占8.3%) 如：威而刚、少而精、大而全

Vg+Vg→V(共3个，占3.2%) 如：离退休、玩儿闹

R+R+R→R(共1个，占2.1%) 如：你我他

以并列方式构成的新词中，形容词最多，占总数的47.9%；动词次之，占20.9%；再者是名词，还有个别的代词。

（6）补充式（C）47个，占三音节词的0.7%，共有7种构词模式：

Vg+U+Vg→V(共14个，占29.8%) 如：谈得拢、玩得转

Vg+U+Ag（共11个，占23.8%）

→ V(共10个，占90%) 如：过得硬、黑得好

↘ N（共1个，占10%）如：热得快

Vg+Vg→V(共10个，占19.1%) 如：热昏头、逗咳嗽

Vg+ZP→V(共4个，占9.5%) 如：谈不拢、搞不通、吃不开

Vg+Dg+Ag→V(共3个，占7.1%) 如：搞不通、摆不平、说不准

Vg+Ag→V(共3个，占7.1%) 如：说清楚、气不公

Vg+LP→V(共2个，占4.3%) 如：洗洁净

补充式的三音节词动词占98.7%，名词只占1.3%；结构模式以Vg+U+Vg，Vg+U+Ag，Vg+Vg为主。

（7）前缀（Q）（共 29 个，占 0.4%）

N（共 26 个，占 89.7%）如：类继父、非集团、超高温

V（共 3 个，占 10.3%）如：负反馈、负建设、洋冒进

（8）后缀（H）（共 1183 个，占 18.2%）

N（共 1087 个，占 91.9%）如：敦煌学、方法论、追星族

V（共 96 个，占 8.1%）如：正规化、国际化、轨道化

4 三音节新词语的构词规律

从上文的描述中，我们可以发现三音节新词语的构词规律：

（1）定中式最多，加前缀最少，依次是定中式、加后缀、动宾式、状中式、主谓式、联合式、补充式、加前缀。

（2）定中式构成的全部是名词，以 Ng+Ng、Ag+Ng、Vg+Ng 三种类序为主，名词词素居后的最多，占 99.5%；状中式、动宾式、主谓式、补充式构成的全部是动词，联合式构成的多数是形容词。

5 三音节新词语的识别规则

对于三音节新词语中的单一类序，按照描述中的结果确定，交叉类序的统计结果归纳如下：

构件 1	构件 2	构词方式	词性	统计结果
Ng	Vg	D(46.7%)	V(21)	V(82.2%) N(17.8%)
		W(53.3%)	V(16)	
			N(8)	

根据上表的统计结果，假设识别出的该类序列都是三音节新词语，我们可以总结出这样一条规则：Ng +Vg→V，其准确率为 82.2%；Ng +Vg→N，其准确率为 17.8%。

构件 1	构件 2	构词方式	词性	统计结果
Vg	Ng	D(59.5%)	N(845)	N 59.5% V40.5%
		B(40.5%)	V(575)	

根据上表的统计结果，我们可以教给计算机这样一条规则：Vg +Ng，优先考虑定中式，则为名词，其准确率为 59.5%；如果不是，再次考虑动宾式，则为动词，其准确率为 40.5%。

构件 1	构件 2	构词方式	词性	统计结果
Ng	Ng	D(99.6%)	N(1800)	N(100%)
		L(0.4%)	N(7)	

根据上表的统计结果，我们可以总结出这样一条规则：Ng +Ng→N，其准确率可达到 100%。

构件 1	构件 2	构词方式	词性	统计结果
Vg	BP	Z(50%)	V（2）	V(100%)
		B(50%)	V（2）	

根据上表的统计结果，我们可以总结出这样一条规则：Vg+BP→V，其准确率可达到 100%。

构件 1	构件 2	构词方式	词性	统计结果
ZP	Ng	D(96.2%)	N(101)	N(96.2%) V(3.8%)
		B(3.8%)	V(4)	

根据上表的统计结果，我们可以总结出这样一条规则：ZP +Ng，优先考虑定中式，则为名词，其准确率为 96.2%；如果不是，再次考虑动宾式，则为动词，其准确率为 3.8%。

构件 1	构件 2	构词方式	词性	统计结果
Vg	Vg	Z(46.8%)	V(22)	V(100%)
		B(25.5%)	V(12)	
		L(6.4%)	V(3)	
		C(21.3%)	V(10)	

根据上表的统计结果，我们可以总结出这样一条规则：ZP+Ng→V，其准确率可达到 100%。

构件 1	构件 2	构件 3	构词方式	词性	统计结果
Vg	U	Vg	Z(13.3%)	V(2)	V(100%)
			C(86.7%)	V(14)	

根据上表的统计结果，我们可以总结出这样一条规则：Vg+U+Vg→V，其准确率可达到 100%。

构件 1	构件 2	构词方式	词性	统计结果
Vg	Ag	B(40%)	V(4)	V (70%) A (30%)
		W(30%)	A(3)	
		C(30%)	V(3)	

根据上表的统计结果，我们可以总结出这样一条规则：Vg +Ag→V，其准确率为 70%；Vg +Ag →A，其准确率为 30%。

构件 1	构件 2	构词方式	词性	统计结果
Vg	ZP	B (50%)	V(4)	V(100%)
		C (50%)	V(4)	

根据上表的统计结果，我们可以总结出这样一条规则：Vg +ZP→V，其准确率可达到 100%。

6 结语

从上述的统计结果看，只要两个构件结合到一起的词性是单一的，我们不必考虑其构词方式，就可以很容易的将文本中的新词语加以识别并进行正确的标注；如果两个构件结合到一起的词性不是单一的，根据最大概率法，我们大致也能确定文本中的某个新词属于哪个词类。当然，我们目前主要是从静态的角度对新词语的构词方式进行了分析，在今后工作中，我们准备将静态跟动态结合起来，以期为自动分词中未登录词的识别提供更好的依据。

参 考 文 献

[1] 陈小荷. 自动分词中未登录词问题的一揽子解决方案[J]. 语言文字应用，1999，(3).

[2] 苑春法，黄昌宁. 基于语素数据库的汉语语素及构词研究[J]. 语言文字应用，1998，(3).