基于数据库的现代汉语新词语缩略语的研究*

鲍明凌　亢世勇

烟台师范学院中文系（264025）

Kang sy 46@ sohu.com

提要：我们利用成熟的关系数据库，描述现代汉语新词语中的缩略语与其相关属性的二维关系，建立了新词语缩略语信息数据库，在此基础上进行统计，对新词语缩略语的各个方面进行了研究。

关键词：新词语；缩略语；数据库

1．新词语缩略语数据库的研究和实现

1．1 缩略语的有关问题

缩略语这个术语，学术界有不同的提法。如“简称”、“缩略词语”、“简称词”、“简缩语”、“略语”、“省称”、“省文”、“省语”等等，与之相应的，学术界对其定义也是说法不一。比较之后，我们认为王立廷在他的《缩略语》一书中的定义是比较好的：“普通话中由固定说法经过压缩和简略而形成的词语。”这个定义能够比较全面的概括出缩略语的特点。这个定义一是圈定了缩略语的存在范围——普通话中；二是明示和暗含了缩略语成立的两个条件：①原词语必须是固定说法；②必须先存在固定说法的原词语，才使缩略语的产生成为可能；三是说明了原词语形成缩略语的两条主要途径——压缩和简略；四是指出了缩略语的性质——词或者语。

如何界定新词语中的缩略语，说法不一。张志毅和张庆云先生在《词和词典——汉语缩略语的特点》一文中提出了比较好的标准。在此基础上，我们总结了以下几条界定标准。

（1）、缩略语必须有原词语，而且原词语要“具有某种程度的固定性、复呈性”（张志毅，张庆云，1992，p. 73）。缩略语是某些语言的简化形式，所以，只要是缩略语就应该具有自己的原词语。

（2）原词语同缩略语之间必须“具有先正后逆的二向性”（张志毅，张庆云，1992，p74）。从原词语可以缩略出缩略语，而且从缩略语我们也可以找到它的原词语。比如“彩电”——“彩色电视”、“彩色电视机”，“残运”、“残运会”——“残疾人运动会”。

（3）在形式上，缩略语中必须具有从原词语这个“全体提取部分的关联性”（张志毅、张庆云，1992，p75）。就是说原词语和缩略语之间必须在音节形式上保持一定的、必要的联系。作为现代汉语的缩略语，这个部分就是原词语中的代表性的音节、语素、词语等。如“湖南—湘”，这个“湘”与“湖南”不存在部分与整体或任何音节形式上的联系，所以，“湘”绝对不可以看作是“湖南”的缩略语。这里有必要指出，英语缩略方式与汉语不同，较多的是从原词语中提取字母，如IC、CPU、CD-ROM等，有的学者将它们叫做“字母词”。这类词语目前已有好多进入到现代汉语当中。而汉语缩略语提取的大多是语素或词语。由于外来词的大量引入，现代汉语新词语中的缩略语也存在不少的英语缩略语和汉语语素或词或短语结合的汉英混用的缩略语，这种情况符合此项标准，我们在数据库中将它们称为“准缩略语”。

（4）从语义信息量上看，原词语和缩略语之间的“语义信息量必须具有守恒性”（张志毅、张庆云，

* 本项目研究得到国家社科规划项目（01CYY002）的支持。

1992，p75)。就是说二者在语义上应该完全相同或基本相同。这种情况，仅指词汇意义，不包括色彩意义和语法意义。

1．2 缩略语数据库的属性信息的设立与描述

1．2．1 原词语。 是由缩略语逆向复原而确定的缩略语的原词语。

1．2．2 类型。对缩略语类型进行归纳，将缩略语大致分成以下四种类型：

① 简称：就是原词语是较长或较复杂名词或名词性短语提取语素后的简化形式。如“峨影”、“儿捧”、“二轻局”。在数据库中用“简”来表示。

② 略语：就是原词语主要是非名词或非名词性短语的简缩形式。它简缩的方式只有缩合和节略两种。如“结扎”、“代谢”等。在数据库中，用“略”表示。

③ 缩语：缩语指的是与原词语并列结构项数相等的数词+并列项的共同成分（共同项）构成的缩略语。有时在数词的后面提取上合适的量词，原词语并列项的共同成分的位置可能在并列项的尾部、中间或前面，或不确定，在类型字段中，属于这种缩略语的我们用“缩”来表示。

④ 准缩略语：这种缩略语之所以被称做准缩略语是因为这种缩略语是由缩略部分和未缩略部分组成。比如“工资级差——工资等级差别”，实际进行缩略处理的部分只有“等级差别”；“五岛气候”——热岛气候、雨岛气候、干岛气候、湿岛气候、浑浊岛气候，当中的“气候“是个未缩略的词，也可以说是从原词语中提取的共同项，这显然不同于“德智体/全面发展”中的“全面发展”（我们把它看作是保留项），所以，它是一个典型的缩略语。

1．2．3 缩略方式。 四种类型的缩略语，存在不同的缩略的方式，总体看来，大致存在以下几种缩略方式：

① 缩合。这种缩略方式就是把分成几节的原词语，先缩略掉某些节（被整节删除的节，我们在节缩项中填写），或者缩略掉某些节的某些部分，剩余的部分再结合而构成缩略语。缩略掉的节或节的部分的位置多种多样。如军人家属——军属、四川大学——川大、奥林匹克运动会——奥运会、农业、林业、畜牧业、副业、渔业——农、林、牧、副、渔、政治协商会议——政协等。这是缩略语中用的最多的一种缩略方式。

② 节缩。 就是直接缩略掉分成几节的原词语的某个节或某些节，剩余的部分就是代表原词语的缩略语。如中国人民政治协商会议共同纲领——共同纲领、电视连续剧——连续剧、政治教导员——教导员、大庆油田——大庆、九三学社——九三等。

③ 提取。就是从原词语中先提取某个或某些成分，使缩略语在形式上保留原词语的痕迹，然后再提取与原词语密切相关的内容，以最大限度的体现原词语的语义信息量，二者结合构成缩略语。如国际货币制度改革及有关问题委员会——二十国委员会（这个委员会共有二十个国家组成）、关于加强电影艺术片创作和生产的领导的意见（草案）——电影三十二条（该文件的内容共计三十二条）、中国高技术研究发展计划纲要——863 计划（该计划的提出时间是 1986 年 3 月）；有的是从原词语中直接提取某个或某些成分作为缩略语，而不需要再提取与原词语相关联的内容，这种提取与节缩可以说殊途同归。如果提取项与原词语中的共同项一致，就是提取的是原词语中的共同项，我们就按照共同项处理，在构成方式中用 g 表示，如三废——废气、废水、废渣当中的“废”是提取项，也是共同项，我们描写成 ng；而三废处理——废气、废水、废渣的处理，当中的“处理”不宜看作是共同项，所以，就作为原词语中的提取项，描写成 nge。

这种缩略方式，在较多的资料中，都叫做是“缀加”。我们参考王立廷先生的《缩略语》一书，归纳的缩略语的提取方式的类型主要有以下几种：提取时间（t）、提取地点（d）、提取量词（q）、提取范围（f）、提取性质（x）、提取方位（o）、提取人物（r）、提取内容（y）、提取共同项（g），共计九种。提取了什么类型，就用该类型值的代码填写。如“东北九省”，在提取字段中填上“o”，表示提取方位。

以上三种缩略方式基本可以涵盖缩略语的缩略方式，但是，有的缩略语在进行缩略处理时也会有自

己独特的缩略方式，在我们收集的缩略语数据库中，大致存在以下几种特殊的缩略方式：

④特殊的缩略方式：

A、近义词语替换。如澳洲抗原——澳大利亚/抗原、浮吊——浮式/起重机等。对于当中的同义代替部分，我们用 m 表示，其余部分仍然按照位置号和区号合并描写，以上的两个例子，我们可以用 mb, alm 来描写。

B、上位词语代替下位词语。如国家教委——中华人民共和国/教育/委员会在描写它的构成方式时，我们采取的方法是，用“s”来表示某词语是用上位的词语来替换的，其余仍然按照位置号+区号的方法描写，所以，该例子就可以描写为 sb1c1。

C、英文中的字母进行缩略。有时是单纯的字母缩略，有时是用数字+英语单词的共同字母，或者是合并的英语单词的字母与汉语某个或某些词语的合并，对于此类的缩略语，我们把一个单词看作一节，若提取的字母是几个单词中共同存在的，按照共同项处理，否则，也采用位置号+区号的描写手段。

1．2．4 构成方式。在建设缩略语数据库时，对简称型的原词语的描述，我们采用的描写方法是区号+位置号。我们首先将原词语划分成几节，这个“节”可以是语素、可以是词、可以是短语，甚至是小句，划分的依据主要是在保留各节意义完整性的基础上，从构成方式易于描写的角度。不同的节之间用‘/’表示。每一节算做一个区，从第一节开始，用 a 表示，按照原词语的线性排列，后面的每一节，依次用 b\c\d\e\f\g……表示，这个用于表示节的字母，就是区号，在每一个区中，不同的语素(或者是不同的音节)，分别用不同的数字来表示，第一个语素的位置号是 1，第二个语素的位置号是 2，依次类推。举一个例子。北京大学——北大，原词语北京大学，首先可以分成两节——北京/大学，当中的“北京”就是 a 区，“大学”就是 b 区，再看他们的位置号，当中的“北”是语素 1，“京”是语素 2，“大”是语素 1，“学”是语素 2，所以，“北京大学”用区号+位置号就可以表示为 a1a2b1b2，这样，在描写北大这个缩略语的构造方式时，就可以用 a1b1 来表示了。如果提取的不是某一节的某个语素，而是一节，那么，就用该节的区号直接表示，而不用位置号。如国际/乒乓球/联合会——国际乒联，它的构成方式的字段就可以用 ab1c1。对于具有共同项的原词语，当中的共同项，我们用 g 表示，其余部分仍然按照划节后的位置号和区号来表示。如中学、小学——中小学可以表示为 a1b1g，如果在缩略语中存在几个共同项，如“三学四评——学政治、学文化、学技术、评政治、评文化、评技术、评团结”，可以在共同项的后面加上 1、2、3 等数字表示。上述例子可以描写成“ng1ng2”．如果采用节缩的方式进行缩略，缩略语中删除省略的某一节，那么，在构成方式字段中用该节的代码与“0”组合来表示该节是整节删除。如东航——东方/航空/公司这个缩略语。当中的“公司”是整节删除，所以在构成方式的字段中就填写‘a1b1c0’

对于没有进行划节的缩略语，构成规律性并不强，不符合我们的规划，所以不对他们进行划节处理。

1．2．5 结构。就是缩略语的语法结构，汉语缩略语不同于印欧缩略语，缩略后仍然存在语法结构关系，填写该字段时，分别用“并”表示并列结构，用“偏”表示偏正结构，用“支”表示支配结构，用“陈”表示陈述结构。

2．新词语缩略语统计研究

2．1 新词语缩略词的总体情况

在缩略语数据库的基础上，我们分别对缩略语的四种类型进行了全方位的信息统计，在我们采集的新词语数据库中，共有新词语 38306 条，我们从中提取的缩略语 2957 个，可见缩略语在整个的现代汉语新词语中所占的比例是 7.72%。 商务印书馆 1985 年正式出版的《现代汉语词典》用缩略法构成的就有 1008 个，上海辞书出版社 1987 年出版的《汉语新词词典》共收新词 1854 条，缩略语的有 891 条。在现代英语中，据统计，缩略词（abbreviation）在我们看到和听到的英语词汇中，占 25%以上。现有的各种缩略词词典的版本数比起普通英语词典版本数差不了多少，公认的标准缩略词词典《Abbreviations Dictionary》1986 年版本收入的词条达 23 万余条，与收入英语词目最多的最新版《牛津英语大词典》

的616，500条词目相比，占1/3还要多得多，比1985年版《新英汉词典》的词目要多1.9倍。商务印书馆1994年出版的《最新高级英汉词典》的第一页，共收入17个词目，其中缩略词竟占了10个：a、@、AA、AAA、AAAL、AAAS、AAM、AAS、A.B、ABA。另外张志毅先生在《汉语缩略语特点》一文中所统计的数量大约是24%，通过前后数据的比较，我们不难发现，缩略语在近几年的发展速度是较快，但并不证明缩略语是新词语产生的主要途径。

2．2缩略语的音节

类型	总数	一	二	三	四	五	六	七	八	九
简称	数量	11	1120	382	43	18	2	3		
1580	比例	0.69%	70.89%	24.06%	2.71%	1.13%	0.13%	0.19%		
缩语	数量		270	86	204	22	17	8	4	
612	比例		44.19%	14.08%	33.39%	3.60%	2.78%	1.31%	0.65%	
略语	数量		444	32	17		2			
495	比例		89.70%	6.46%	3.43%		0. 40%			
准缩略	数量		1	10	193	45	13	3	3	2
语270	比例		0.37%	3.70%	71.48%	16.67%	4.81%	1..11%	1.11%	0.74%

可以看出，二音节在简称缩略语、缩语、略语中占绝对的优势，其次是三音节和四音节，而准缩略语的音节优势是四音节。这与准缩略语直接从原词语中移植某些词语的特点是密不可分的。新词语的音节数有增多的趋势。

2．3缩略语的类型

类型	简称	缩语	略语	准缩略语
数量	1580	612	495	270
比例	53.43%	20.70%	16.74%	9.13%

从上表的统计来看，简称缩略语所占的比例是最高的，达到了53.43%，是缩略语的主要类型，这证明在缩略语中，原词语是名词或名词性短语的占大多数；其次是缩语，占缩略语总数的20.70%这与缩语本身简短、信息量大的特点密不可分；略语所占的比例相对较低；准缩略语所占的比例是最低的，这个比例是准缩略语自身语法特征不够稳定的特点所决定，也从另外一个角度说明语言所具有的稳定性。

2．4缩略语的缩略方式

由表可见，简称缩略语中，缩合是它最主要的缩略方式。而提取和同义代替，大多数的语法学家认为，不是缩略语。这两种方式在我们的收集的缩略语中的比例是很低的。

缩语的类型决定它基本的构成形式就是并列项项数+并列项中的共同项，当然，为了使缩语的表义更加明确、更加清楚、更加富有区别性、更加符合汉语的语法特点，所以，在它基本的构成形式上，可以有从原词语中提取的要素，如性质、范围、量词、时间、地点等，或者采用特殊的处理方式，如同义代替、上位词语代替等。当然，采用提取时间、提取地点、提取范围等方法的缩语的比例是很低的。只有5.00%，所以，缩语从形式上也尽可能通过提取共同项的方式使之表义明确、直观。考察缩语的缩略方式，在下文构成方式一节有详细的介绍。

总数 1580（简称）	缩略方式	缩合	节缩	提取	同义代替
	数量	1503	62	8	6

（缩略方式表示原词语演变成缩略语的方法，构成方式则是原词语中提取的音节或语素、短语在缩略语中的明确表示。）

总数	构成方式										
1580	构成	Ab0	A1b1	A1b2	A1b1b3	A2b1	A2b2	A1a3	A0b	A1b1g	A1b1c1	A1b1c3
	数量	30	566	161	91	66	45	41	23	38	41	23
	比例	1.90%	35.82%	10.19%	5.76%	4.18%	2.85%	2.59%	1.46%	2.41%	2.59%	1.46%

从缩略方式考察，略语中采用缩合方式的共计 489 个，占 98.79%，采用节缩方式的 4 个，占 0.81%，采用同义代替的 2 个，占 0.40%， 所以，缩合是略语最主要的缩略方式。

类型	结构类型	偏正	并列	陈述	/	支配
总数 1580 简称	数量	1372	128	13	65	0
	比例	86.84%	8.10%	0.82%	4.11%	0
总数 612 缩语	数量	471	138	1	0	1
	比例	77.09%	22.59%	0.16%	0	0.16%
总数 270 准缩略语	数量	268	0	2	0	0
	比例	99.25%	0	0.75%	0	0
总数 495 略语	数量	134	120	42	0	199
	比例	27.07%	24.24%	8.48%	0	40.20%

因为准缩略语本身的特殊类型，所以，它的用到的缩略方式多种多样。排除综合运用多种缩略方式，它的缩略方式大致有两种情况：或者它是通过缩合一部分的方式；或者是通过缩语与原词语中的某个或某些词语结合。所以，我们从这两个大的角度来考察准缩略语的缩略方式。在 270 个准缩略语中，用到缩合方式的 128 个，占 47.94%，这种情况主要是由于当中的未缩略词语从汉语自身的特点或语法特点等方面不便进行缩略，否则便会破坏缩略语的表义的明确性；通过缩语与原词语中的某个提取部分结合的有 142 个，占 52.06%。这种现象很显然也是从表义明确的角度考虑的。

2．5 语法结构

缩略语的语法结构类型如下表：

由此可见，简称缩略语中的最主要的语法结构关系是偏正结构。其次是并列结构。这种情况与简称缩略语是较长或较复杂的名词或名词性短语提取语素后的简化形式的语法性质有密切的关系。

缩语中属于支配结构的只有“刹三风”一个，属于陈述结构的只有“五业并举”一个。所以，在缩语中，偏正、并列结构是其最主要的语法结构。而并列结构的出现，也主要是由两个或以上的偏正结构的缩语组合而成的，这种情况与缩

语的“并列项数+共同项”的组合模式是密切相关的。

准缩略语当中属于偏正结构的 268 个，占准缩略语总数的 99.25%；其余的两个属于陈述结构，即“老中青三结合”和“德智体全面发展”，占 0.75%。

略语中属于支配结构的占主体，这是略语本身的“非体词性”的语法特点所决定的。其次是偏正结构和并列结构，所占比例最低的是陈述结构。

2．6 构成方式

对于构成方式，简称缩略语似乎没有什么规律，可以说，各种各样的构成都存在，我们统计的上表中，难以把所有的构成方式都罗列出来，但可以看出，a1b1、a1b2 这两种构成是最主要的。如果排除未能统计出的数量，在简称缩略语中，提取的语素或音节位于原词语各节的头和尾的比例是 95.29%。这一点与人们的思维习惯是密切相关的。

从上表缩语的构成方式统计，我们可以看到，“ng”这种缩语的最基本的构成方式占的比例是最高的，其他几种构成方式此上都是基础的，出于区别同形、表义明确、语音节奏等原因产生的，如提取保留项的“nge”、提取性质的“nxg”等方式。在提取性质的缩语中，提取的表示性质的词语只有“大、小、新、旧”等少数几个。

总数 611	构成方式	Ng	Ng1ng2	Nge	Nxg	Nqg	ngne
	数量	308	62	16	32	53	44
	比例	50.41%	10.15%	2.62%	5.24%	8.67%	7.20%

略语的构成方式也比较复杂，其中 a1b1 的 249 个，占 50.30%，a1b2 共计 74 个，占 14.95%，a1b1c1 共计 12 个，占 2.42%，a2b1 共计 48 个，占 9.70%，a2b2 共计 23 个，占 4.65%，其余不方便统计的 89 个，占 17.98%。所以，在略语中，a1b1 与 a1b2 是其最主要的构成方式。

准缩略语的构成方式多种多样，其构成方式是 nge 的 117 个，占 43.33%，是运用比例最高的构成方式，ab1c1 的 34 个，占 12.59%，a1b1c 的共计 19 个，占 7.04%，a1b1c1e 的 15 个，占 5.56%，a1b1e 的 12 个，占 4.44%，其余不方便统计的 73 个。准缩略语的这种构成方式的复杂性是其本身的语法特点所决定的。

以上我们对新词语缩略语各个方面进行了统计研究，重点说明了缩略语的构成方式，希望这样的研究为汉语研究以及汉语信息处理有一定的作用。

参考文献：

1、俞士汶，《现代汉语语法信息词典详解》，清华大学出版社，1998 年 4 月。

2、张志毅 张庆云，《词汇语义学》，商务印书馆，2001 年。

3、王立廷，《缩略语》，新华出版社，1997 年 1 月。

4、王魁京等，《现代汉语缩略语词典》，商务印书馆，1996 年 7 月。

5、张志毅 张庆云，《词和词典》，中国广播电视出版社，1994 年 4 月。

6、亢世勇，《汉语数据库建设及其应用》，作家出版社，2000 年 9 月。

7、李达仁，《汉语新词语词典》，商务印书馆，1993 年 7 月。

8、李行健等，《新词新语词典》，语文出版社，1989 年。

9、于根元，《现代汉语新词词典》，北京语言学院出版社，1993 年 3 月

10、陈建民，《现代汉语里的简称》，《中国语文》，1963 年 4 月。

11、刘叔新，《汉语描写词汇学》，商务印书馆，1990 年。

作者简介：鲍明凌（1975—），男，山东烟台人，硕士研究生，主要研究领域为汉语信息处理；亢世勇（1964—），男，陕西延安人，硕士，教授，硕士生导师，主要研究领域为汉语信息处理和计算语言学。

Researches on The New Clipped word ased on The Corpus

Bao Mingling Kang Shiyong

7. Yantai normal university shandong 204625 china

Email :kangsy46@sohu.com

Abstract: We use the ripely relational database to describe the relationship between the abbreviation of new words in modern Chinese and its relavant features. We establish the information database of abbreviation of new words,on the base of which we count and study all aspects of abbreviation of new words.

Key word: new words abbreviations abbrevite