现代汉语合成词结构数据库

刘 云　俞士汶　朱学锋

摘　要　本文主要介绍了北京大学计算语言学研究所开发的现代汉语合成词结构数据库，并从未登录词的识别和为对外汉语教学研究提供依据这两个方面探讨了现代汉语合成词结构数据库的应用。

关键词　现代汉语　合成词　结构　数据库

Construction of the Contemporary Chinese Compound Words Database and its Application

Abstract　In this paper the authors offer an introduction to the contents and use of the compound words database of contemporary Chinese constructed at Peking University.The paper also includes an introduction to the application of the word formation database.

KEYWORD　Contemporary Chinese,Compound Words, Construction,Database

1．引言

为研究现代汉语复合词的构造规律和未登录词的识别，北京大学计算语言学研究所针对《现代汉语语法信息词典》中的所有双音节和三音节词，开发了一个现代汉语合成词结构数据库。这项工作与《现代汉语语素库》都是对《现代汉语语法信息词典》的补充。自 1986 年以来，北京大学计算语言学研究所和中文系合作，历时十余载，于 1995 年底研制出了《现代汉语语法信息词典》，其规格说明书全文发表在 1996 年第 2 期《中文信息学报》上，更详细介绍这部词典的专著《现代汉语语法信息词典详解》于 1998 年 4 月由清华大学出版社出版。（俞士汶等 1998）并于 1998 年研制出《现代汉语语素库》，这些都在汉语信息处理领域发挥了重要的作用。（俞士汶等 1999）

《现代汉语语法信息词典》共收入 51696 个词语，其中单音节词 3803 个，双音节词 32711 个，三音节词 7926 个，四音节词 7220 个。由于单音节词没有内部结构，四音节词暂时还没有考虑，所以数据库的登录项实际上是在《现代汉语语法信息词典》中的 40637 个双音节词和三音节词中选择的。由于单纯词（460 个）是一个整体无法再进行结构分析，人名（75 个）、地名（462 个）对研究现代汉语汉语的构词规律作用不大，所以该数据库把这三类词排除在外。这样现代汉语合成词结构数据库实际登录的词共有 39370 个。（为称说方便，以下简称构词库）本文将介绍这个构词库的主要内容及其应用。

2． 合成词构造的属性描述

对于一个词语可从两个方面考察，如果从这个词语的整体而言，则可考察其外部功能；如果从这个词语的局部而言，则可考察其内部构造。就现有的研究成果来看，学者们把目光更多地投向了外部功能，诸如词语的词性考察，句法功能的考察等等。对于词语的内部构成的考察则显得门庭冷落。难怪张旺熹、崔永华两位先生在总结了对外汉语教学语法研究所取得的成就之后，指出了尚存的不足，其中的第一个方面就是“ 语素、构词及语段研究严重缺乏 ”，并指出“ 加强汉语语素和构词研究是推动对外汉语教学发展的需要 ”。（ 张旺熹、崔永华 1999）由此可见，不仅中文信息处理界对语素和构词研究极为重视，对外汉语教学界对语素和构词研究也极为关注。但自从陆志韦等著的《汉语的构词法》出版后，语言学界对词语的构造始终未能给予足够的关注，尚没有更大规模、更深入的研究成果问世。值得欣慰的是，清华大学计算机系建立了一个大规模的汉语语素数据库，并且在汉语语素库的基础上建

立了汉语构词知识库。（苑春法、黄昌宁 1998）这是一种从下往上的研究方式，我们采取的是一种从上往下的研究方式。我们是在已有的《现代汉语语法信息词典》的基础上，采用相应的标记集对汉语的构词方式进行研究，而且还对构词的层次进行了分析。

要完整地描述词语的构造，就需标出整个词语的结构性质和词语组成部分的性质。如对于定中式的名词而言，可以是“名+名”（如“铁路”“信纸”）、“形+名”（如“温泉”“红旗”）、“动+名”（如“燃料”“刊物”）、“数+名”（如“八股”“千金”）、“代+名”（如“他人”“何处”）等，如果不标出词语组成部分的性质，那么这种描述就不够精细。同样对于“动+动”结构而言，可以是联合式（如“斗争”“转折”）、述宾式（如“罢教”“挨打”）、连动式（如“听写”“截获”），也可以是状中式（如“迁就”“捐助”）、述补式（如“推翻”“打倒”），可见如果不标明整个词语的结构性质，那么这种描述也是不够精细的。所以构词库对每个词语组成部分的性质和词语的结构性质都作了详尽的标注。

对于词语组成部分的性质，我们主要是依据组成部分的语素在合成词中的作用来标明语素的语素类。如读“song1”的“松”在“松树”中，是树的名称，起名词性作用；“松”在“疏松”、“松软”等词语中是“松散”义，起形容词性作用。语素在复合词中的作用可用“替换法”进行检测。之所以说“松”在“松树”中起名词性作用，是因为“松”可用“桃”替换，得到的“桃树”与“松树”都是树，属于同一语义范畴，“桃”是名词，故可推断这里的语素“松”也是名词。又之所以说“松”在“疏松”、“松快”等词语中起形容词性的作用，是因为“疏松”中的“松”可用“软”替换，“松快”的“松”可用“轻”替换，词义基本不变，而“软”“轻”都是形容词性的，故可推断“松”起形容词性作用。这样就可以给合成词中的语素分类，起名词性作用的就叫“名语素”，在“前字”和“后字”字段中就填为“n”；起动词性作用的就叫“动语素”，在“前字”和“后字”字段中就填为“v”，如此等等。（可参见俞士汶等 1999）

对于整个词语的结构性质，我们首先把合成词分为附加式、复合式、重叠式和简称四大类，然后在每个大类下面再细分成各种小类。附加式分为前接式、后接式两小类；复合式分为联合式、连动式、定中式、状中式、述宾式、述补式、补充式、主谓式等八小类；简称可分为缩减式和标数式两小类。为填写方便，一律使用简称。“前”代表“前接式”，“后”代表“后接式”，“联”代表“联合式”，“连”代表“连动式”，“定”代表偏正式的“定中结构”，“状”代表偏正式的“状中结构”，“述”代表“述宾式”，“补”代表“述补式”，“充”代表“补充式”，“主”代表“主谓式”，“重”代表“重叠式”，“缩”代表“缩减式”，“数”代表“标数式”。

此外，对于三音节合成词还进行构词层次上的分析。可分为三种情况。第一种是 1+2 模式，如多媒体、单音节等；第二种是 2+1 模式，如消费品、救济粮等，这种模式还包括一种比较特殊的情况，即前两个字分别修饰第三个字，如“白矮星”中的“白”和“矮”分别修饰“星”，“左右手”中的“左”和“右”分别修饰“手”，在层次上也看作“2+1”模式；第三种是 1+1+1 模式，如短平快、高精尖等，这种模式还包括三个字之间的关系不容易确定的情况，如“安理会”是“安全理事会”的简称，简称后“安”与“理”不能组合，同样“理”与“会”也不能组合（这与动词性的“理会”不同），这种情况在层次上也看作“1+1+1”模式。双音节合成词的儿化不作为三音节词，如“挨个儿”是“挨个”的儿化，所以对这类词不进行层次分析。当然，如果三字词中的“儿”不是用来儿化的，还是应当分析其层次，如名词性的“早产儿”等。同样三音节词的儿化仍作为三音节词，如“疤瘌眼儿”仍需进行层次分析。

按理说，四字词也应进行整体结构性质、组成部分的性质和层次的标注，但考虑到本项研究主要是针对中文信息处理中的未登录词的识别和复合词的构造规律，所以暂时没有对四字词进行进一步的标注。

3． 构词库的各个字段

现在的构词库共有 40637 个记录，每个记录有如下几个字段（其中的词语、读音、词类、同形、义项、备注这几个字段是利用的以前的成果）：

词语：《现代汉语语法信息词典》中的两字词和三字词。

读音：词语的拼音。用 1，2，3，4，5 分别表示阴平、阳平、上声、去声和轻声。

词类：词语所属词类的代码。

同形：词类相同的同形词中，拼音不同后者词项不同的，分别标上 A、B、C；词项相同而义项不同的，则填 1，2，3；字母与数字同时存在时，则将字母置于数字之前，如 A1、A2、A3、B1、B2、B3 等。

构词：词语的结构类别。

义项：简单释义。如对于同形字段中填 1 的“板栗”为“植物”，同形字段中填 2 的“板栗”为“果实”。

备注：用语举例或作其他说明。如对于“板眼”，此字段填了“他~多/很有~”。

层次：用于描写三音节词的构造层次。

前字：用于描写双音节词的前字的所属的语素类。语素类主要有名语素 n，形容语素 a，动语素 v，副语素 d，助语素 u，前接成分 h 等等。

后字：用于描写双音节词的后字的所属的语素类。语素类主要有名语素 n，形容语素 a，动语素 v，副语素 d，助语素 u，后接成分 k 等等。

4． 构词库的应用

现代汉语构词库的开发无论是对本体研究，还是对应用研究都将起到推动作用。限于篇幅，本文仅谈构词库对中文信息处理中的未登录词的处理和在对外汉语教学中的应用这两个方面。

4． 1 未登录词的识别

中文信息处理中无可避免的未登录词是自动切分和词性标注中的瓶颈之一，不妥善地解决未登录词问题而进行机器翻译和信息提取都不会取得较理想的效果。有了构词库后，就可以把构词库与以前做的语素库结合起来识别未登录词了。语素库中的每个语素都标明了语法属性，构词库中的每个词都标注了构成的方式。这样对于一个未登录词就可以从两个方面确定其语法属性。首先，从整个两字词的构成模式入手。可以利用构词库提取出汉语的所有的构词方式和频率，也就是把词类、构词、前字和后字这四个字段链结起来进行匹配，经初步检索共得到构词模式 200 多种，当然这其中有的数量比较多，有的数量比较少。这里的构词模式是综合了词类、构词、前字和后字四个字段之后得到的，如“a 联 aa”（昂贵、矮小）就是指词类字段为 a，构词字段为“联”，前字字段为 a，后字字段为 a 的构词模式；又如“v 述 vn”（昂首、拔河）是指词类字段为 v，构词字段为“述”，前字字段为 v，后字字段为 n 的构词模式。最常用的几种构词模式有如下几种：a 联 aa（1175 个），n 定 nn（9570 个），n 定 an（3324 个），n 定 vn（1534 个），n 联 nn（1559 个），n 后 nk（594 个），n 联 vv（307 个），v 联 vv（3114 个），v 述 vn（2300 个），v 状 vv（619 个），v 述 vv（540 个），v 补 vv（561 个）。从统计数字可以看出各种词类的主要构词模式，如双音节的形容词共有 2147 个，“a 联 aa”这种模式就占了 1175 个，约占 54.73%；双音节的名词共有 19367 个，其中“n 定 nn”这种模式占了 9570 个，约占整个名词的 49.41%，“n 定 an”模式占了 3324 个，约占 17.16%，“n 联 nn”模式占了 1559 个，约占 8.05%，“n 定 vn”占了 1534 个，约占 7.92%，“n 后 nk”模式占了 594 个，约占 3.07%“n 联 vv”模式占了 307 个，约占 1.59%，这五种模式合起来占整个名词构成模式的 87.20%；双音节的动词共有 8894 个，其中“v 联 vv”模式 3114 个，约占 35.01%，“v 述 vn”模式 2300 个，约占 25.86%，“v 状 vv”模式 619 个，约占 6.96%，这三种模式合起来占整个动词构成模式的 67.83%。这样在识别未登录词时如

果从语素库中判断未登录词的前字和后字的语法属性就可以根据频率推出其词性和构词方式；如果根据上下文可推测出未登录词的词性，那么也可以根据构词频率推出构词方式，并且确定前字和后字的语法属性；如果词性、前字和后字都无法确定，那么可以根据构词库中每个语素的构词模式和频率来确定未登录词的构词模式。

其次，可从每个语素的构词模式入手。语素库中对每个语素都标注了属性，对于那些由单一属性的语素构成的未登录词而言，很容易就可根据构词模式确定其词性和构词方式，如构词库中共收录了以“ 险 ”为前字的两字词 17 个，以“ 险 ”为后字的两字词 19 个，在“ 险 ”为前字的两字词中有六种构词模式，表中的“ 构词模式数量 ”指这种构词模式的双音节词语的个数，下同。具体构成情况如下：

例词	词类	构词	前字	后字	构词模式数
险恶	a	联	a	a	4
险乎	d	后	a	k	1
险些	d	后	a	u	1
险隘	n	定	a	n	9
险阻	n	联	a	v	1
险胜	v	状	a	v	1

在“ 险 ”为后字的两字词中有六种构词模式，具体构成情况如下：

例词	词类	构词	前字	后字	构词模式数
奸险	a	联	a	a	5
惊险	a	联	v	a	1
保险	a	述	v	n	1
火险	n	定	n	n	5
风险	n	主	n	a	1
救险	v	述	v	n	6

“ 险 ”这个语素在语素库中标注了两个语法属性，一个是“ a ”，另一个是“ n ”， 以“ 险 ”为前字构成两字词时其语法属性都是“ a ”，且后字是“ a ”时构成联合式的形容词，后字是“ n ”时构成定中式的名词。以“ 险 ”为后字构成两字词时其语法属性可以是“ a ”，也可以是“ n ”，如果前字是“ a ”则“ 险 ”是“ a ”；如果前字是“ v ”则“ 险 ”是“ n ”的可能性是 87.50%，是“ a ”的可能性是 12.50%；如果前字是“ n ”则“ 险 ”是“ n ”的可能性是 83.33，是“ a ”的可能性是 16.67%。通过前字与后字匹配模式的频率可以推知未登录词的词类、构词、前字和后字，如构词库中没有收录《现代汉语词典》和《倒序现代汉语词典》中以“ 险 ”为前字的“ 险地 ”和以“ 险 ”为后字的“ 出险 ”，但是可以根据“ 险 ”的构词模式推出“ 险地 ”和“ 出现 ”的词类、构词、前字和后字。对于“ 险地 ”，由于后字“ 地 ”只有助语素“ u ”和名语素“ n ”两种属性，而“ 地 ”作为助语素“ u ”时只出现在“ 猛地 ”“ 突地 ”两个词中，构成“ d 后 du ”模式，“ 险 ”没有“ d ”这种属性，所以“ 地 ”只可能是“ n ”，“ 险 ”作为前字时通常是“ a ”，这样就可以推出“ 险地 ”的构词模式是“ n 定 an ”。对于“ 出险 ”，由于前字“ 出 ”只有量语素“ q ”和动语素“ v ”两种属性，而属性“ q ”没有用来构词，从而可以确定前字“ 出 ”是“ v ”，当前字为“ v ”时，“ 险 ”是“ n ”的可能性是 87.50%，是“ a ”的可能性是 12.50%，这样就可推出“ 险 ”可能是“ n ”，从而进一步得到“ 出险 ”的构词模式“ v 述 vn ”。由此可见，在确定未登录词的构词模式时，要综合考虑前字和后字的各种匹配情况，要把规则和统计的方法结合起来。

4．2 为对外汉语教学研究提供依据

现行对外汉语教学中关于词的教学只涉及词类的划分、词在句中的语法功能及词语的意义，很少涉及构词法的教学。这不是说大家没有意识到构词法在词汇教学中的作用，而是

构词法的研究还没有充分地展开，能用在对外汉语教学中的研究成果还十分有限。因此吕文华先生从建立对外汉语教学中的语素教学的角度呼吁要加强构词法的研究。（吕文华 1999）

要想把构词法引入对外汉语教学必须对汉语构词法作深入的研究。对外汉语教学讲究循序渐进，因此要把常用的而且又简单易学的构词方式先学，把非常用的而且又复杂难学的构词方式后学，这就需要对汉语的构词方式作定量考察。在笔者建立的构词库中共有双音节词 32711 个，其中单纯词 426 个，人名 35 个，地名 292 个；在其余的 31958 个合成词中有前加式 44 个，后附式 1098 个，重叠式 118 个，简缩式 112 个；这样余下的 30586 个复合词中有定中式 16255 个，占复合词的 53.16%；联合式 7178 个，占复合词的 23.47%；述宾式 3813 个，占复合词的 12.47%；状中式 2036 个，占复合词的 6.66%；述补式 698 个，占复合词的 2.28%；主谓式 304 个，占复合词的 0.99%；连动式 256 个，占复合词的 0.84%；名量式 46 个，占复合词的 0.15%。可见在双音节复合词中，按频率排列依次是定中式—联合式—述宾式—状中式—述补式—主谓式—连动式—名量式。这样根据统计结果可以认为定中式、联合式和述宾式是常用构词法，因为三者都占了整个复合词的 10%以上，而且三者合起来占了复合法构词的 89.10%，占整个双音节词的 83.23%；状中式和述补式是非常用构词法，前者占 6.66%，后者占 2.28%；主谓式、连动式和名量式是罕用构词法。对于常用构词法可以作为教学的重点，使学生在初级阶段就能掌握常用构词法，这样学生在学习词汇时，自觉地分析其内部构造，使学生从长期被动的学习状态和死记硬背的苦恼中解脱出来，有效提高词汇的记忆、理解和运用的能力，为进一步地学习打下坚实的基础。

5． 结语

现有的汉语构词库是一个单独的数据库，如果需要可以把它同以前做的语法信息词典或语素库综合起来。

由于社会的发展、时代的进步，新词不断涌现，旧词也不断产生新义，这样无论怎么扩大机器可读词典的规模，也解决不了无可避免的未登录词。但汉语的语素基本上是一个封闭的集合，每个语素的语素项基本固定，这样通过有限的语素来处理无限的新词新义，可以收到以不变应万变的效果。当然，现有的构词库由于暂时没有引入语义知识，在识别语素项，判断构成方式等时还会有些困难。下一步的工作就是要把每个语素构成的词语按语素项分别排列，也就是字词轮排。（可参见俞士汶等 1999）这样每个语素在其组成的词语中的意义就体现出来了，这样就可以像通过已知构词推知未登录词的构词方式一样，可以通过已知的语素义去推知未登录词的意义。到那时，计算机处理未登录词的能力就进一步加强了。

参考文献

陆志韦 1975 《汉语的构词法》（修订本），中华书局。

吕文华 1999 建立语素教学的构想，《对外汉语教学语法体系研究》，北京语言文化大学出版社。

俞士汶、朱学锋、李 峰 1999 现代汉语语素库的开发及应用，《世界汉语教学》第 2 期。

俞士汶、朱学锋、王 惠、张芸芸 1998 《现代汉语语法信息词典详解》，清华大学出版社。

苑春法、黄昌宁 1998 基于语素数据库的汉语语素及构词研究，《世界汉语教学》第 2 期。

张旺熹、崔永华 1999 对外汉语教学语法问题研究的基本态势，载陈章太等编《世纪之交的中国应用语言学研究》，华语教学出版社。

作者通讯地址：刘云 北京大学计算语言学研究所 100871
武汉华中师范大学语言学系 430079
俞士汶 朱学锋 北京大学计算语言学研究所 100871