现代汉语中二字复合词的构词格式研究*

---基于汉语语素数据库的研究之一

苑春法　黄昌宁　周思宇
清华大学计算机系

摘　要：研究汉语复合词的构词规律，对汉语分析中的未登录词的处理有着重要意义。清华大学建立的汉语语素数据库〔5〕对汉语复合词进行了大规模祥尽的描述。在对汉语语素数据库作了初步统计和分析之后发现，名词、动词、形容词在汉语二字复合词词汇中占了很大的比重；由汉语语素组合构成某一种二字词(如名词或动词或形容词)时构词方式和语素类序列的种类比较集中，且有规律可循；大多数汉语二字词的意义就是组成它的两个语素意义的组合。

The Word—Formation of The Two Character Compound Words in Contemporary Chinese

___One of the studies based on Chinese morpheme data-base

Yuan Chunfa　Huang Changning　Zhou Siyu
Dept. of computer　Tsinghua University　Beijing　China

ABSTRACT: To study how the compound words are constructed with morphemes is very importent for unknown words processing in Chinese analysis. In the Chinese morpheme data-base〔5〕constructed by Tsinghua Uniersity,the word formation characteristic of compound words has been detailed described. After some statistic analysis, some importent roles were found. First, the noun, verb, adjective are the majority of the two character compound words. Second, every one of the three kinds of compound words only have a few major kinds of the composite patterns and combining order of morpheme classes. They have high frequency. Third, the meaning of the compound word is usuly combined with the meaning of the composite morphemes.

1. 引言

研究汉语复合词的构词规律，在汉语分析中有着重要意义。这些规律是识别未登录词（指机器可读中词典中未收入的词）并给出其词性判别的重要知识。

---

* 国家自然科学基金支持项目 No.69375017

研究复合词的构词规律从汉语语素的研究着手，是一条有效的途径〔1〕、〔2〕、〔4〕。而建立大规模汉语语素数据库，对汉语语素及复合词进行祥尽描述，为汉语复合词研究以及汉语分析提供了一个强有力的工具。

清华大学对汉语语素数据库的建设投入了大量的人力、物力，共填写张工作单25000张，对汉语语素及汉语语素构成的复合词进行了祥尽地描写，并已全部输入计算机，在Foxbase环境下形成一个汉语语素数据库。

汉语语素数据库对复盖汉语6763个常用汉字的语素的描述包括：一个汉字形成几个语素，每个语素由该汉字的哪些义项组成，汉字的读音，释义，语素的素性分类，语素的构词能力（语素单独成词，还是和其它语素相结合才能构成词）以及组成词时该语素在词中的位置。在语素数据库中对应一个语素的一个义项为一个记录，它有以下数据段：语素名、义项号、读音、释义、素性类别、成词度、成词时语素的位置、备注等。

对汉语语素构成的复合词的描述包括：该词由那些语素组成及其读音是什么；该词的词性，构词方式，（例如述宾，述补，定中，主谓等）以及构成词的语素的素性排列顺序（如：nn表示由二个名词性语素构成）；该词的词义由二个语素的意义合成的情况，该词是单义还是多义等；以及词的释义。一个复合词在数据库中对应一个记录，它有以下数据段：词形、语素、义项号、词类、构词方式、类序、字义组合、多义、释义、备注。

在对汉语语素数据库的复合词进行初步统计的基础上，本文将对其结果进行分析。

2. 汉语中二字复合词的结构

本文主要对复合二字词进行研究。在汉语语素数据库中由语素构成的二字词共计有43097 个，其中名词有22016 个占 51.1%，动词有 15666个占 36.4%，形容词有3276个占7.6%，三类词合起来占总二字词的 95%，也就是说占了绝大部分。研究这三类词的构词规律有着决定性的意义。

2.1 复合词的构词方式统计

复合词的结构基本上和词组、短语、句子的结构一样，也存在着主谓，偏正、联合、述宾，述补等结构。通过对名词，动词，形容词构词方式的统计结果如表1所示。

从表1 中可以看出有以下三个明显特征：

(1) 名词的构词方式以体素联合和定中偏正为主，其中定中偏正占80.6，体素联合占9.3%。二者共占名词二字词总量的约90%。

(2) 动词以述宾、谓素联合和状中偏正三种构词方式为主，它们各占39.7%、27.0%、23.3%。共占动词二字词总量的90.0%。

(3) 形容词以谓素联合为主，占形容词二字词总量的62.5%。

表1:　　二字词的构词方式统计

构词方式	名词	动词	形容词
体素联合	2058	5	10
谓素联合	299	4252	2046
定中偏正	17752	0	164
壮中偏正	242	3647	460
述补	11	927	25
量补	34	0	0
述宾	290	7134	165
主谓	74	243	93
述介	0	23	4
前缀	38	5	0
后缀	776	115	126
重叠	54	13	126
简称	29	13	0
数词缩语	8	0	0
固定词组	230	38	41
未注标记	121	172	16
合计	22016	15666	3276

2.2　复合词构词类序的统计

在现代汉语中，词根＋词根的复合式合成词在整个词汇系统中占有很大的比重。汉语没有形态变化，名、动、形容词性语素交错排列，组成各种类型。可以构成"名＋动"，"动＋名"，"名＋形"，"形＋名"，"动＋形"，"形＋动"，"名＋名"，"动＋动"，"形＋形"共九种素性排列类型。表 2给出了二字词的素性排列统计。

表2:　　二字词的素性排列统计

类＋序	名词	动词	形容词
名＋动	255	631	20
名＋形	90	20	160
名＋名	12583	8	32
动＋名	2559	5338	112
动＋形	23	584	70
动＋动	218	7010	60
形＋名	4630	43	129
形＋动	93	1127	127
形＋形	151	34	2205
其它类序总合	1414	871	361
总计	22016	15666	3276

从表2中可以看出

(1) 名词中绝大多数都是由名词性的语素参加构成，而且这些名词性的语素多数位于后面。例如“名＋名”占 57.2%，“形＋名”占21%和“动＋名”占 11.6%。词中的第二个语素多是名词性的。

(2) 复合动词绝大多数都是由表示动作行为的动词性语素参与构成的，而且多数动词都是由动词性语素按“动＋动”占44.7%，“动＋名”占34.1% 和“形＋动”占7.2%构成。词中的第一个语素是动词性的占多数。

(3) 形容词的素性排列类型很集中，大多数是“形＋形”占67.3%。

2.3 名词的构词规律

二字复合名词的主要构词方式为定中偏正和体素联合两种形式，合计约占二字复合名词的90%。在二字复合名词中数量最多的是以 “名 ＋ 名”为类序的定中偏正结构，有10280个占总数46.7%。其次是“形+名”为类序的定中偏正结构个占总数的20.6%，再其次是“动+名”为类序的定中偏正结构和“名+名”为类序的体素联合结构。二字复合名词主要的构词特点是由两个名词性语素构成一个名词的情况为大多数，即类序“名＋名”12583个，占57.2%， 再一个特点是两个语素构成一个名词而后一个语素是名词性语素的情况是绝对大多数占89.8%。在少数情况下动词性语素和形容词性语素相互组合也可形成名词，如“捕快” “动乱” “跳高”(动+形)，“白描” “大选” “奇遇”(形+动)，“冲突” “差使” “打扰”(动+动)等。

2.4 动词的构词规律

二字复合动词的主要构词方式为谓素联合、述宾和状中偏正，占总数的90.1%。主要的类序为: “动+动”(7010,占44.7%)，“动+名”(5338,占34.1%)，“形+动”(1127,占7.2%)，合计占86.0%。二字复合动词主要的构词特点为两个语素构成一个动词时两个语素中至少有一个动词性语素的情况占大多数，在其中，动词处于在前一位置的又属多数，即"动+动"的谓素联合结构和"动+名"的述宾结构。在少数情况下名词性语素和名词性语素相互组合也可形成动词，如“针砭” “砥砺”等，这种由两个名词性语素组成一个动词的情形是很少见的，约为总数的万分之五。还有名词性语素和形容词性语素组成一个动词情况如“远足” “安心”(形+名)，“客满” “病危”(名+形)，“珍重” “错怪”(形+形)等。 这惺的组成反映语素性在组词时的相互演化。

2.5 形容词的构词规律

二字复合形容词主要的构词方式为谓素联合(2046,占62.5%)，其主要的类序:"形+形"(2205,占67.3%)。其他类序如: “名+动” “名+形” “名+名” “动+名” “动+形” “动+动” “形+名” “形+动”均可组成形容词，但数量较少。尤其是“名+名”情况最少仅有32个。

3. 语素在构成二字复合词时其意义的转化

为了研究语素在构词时它的意义发生变化的情况，在汉语语素数据库中对于每一个二字词我们用“字义组合”数据段来描写这一特性。“2 ”表示二字词的意义是两个语素意义的组合，“0 ”表示词的意义已经发生了转化，不再是两个语素意义的组“1 ”是介于“2 ”与“0 ”之间的一种情况，即词的意义和两个语素的意义有关系但又不完全是两个语素意义的组合。为找到二字复合词构词时意义发生变化的规律，对其字义组合特性进行了统计。

3.1 字义组合特性统计

表3: 字义组合统计表

字义组合	0	1	2	未注标记
名 词	220 (1.0%)	2294 (10.4%)	19328 (87.8%)	174 (0.8%)
动 词	31 (0.2%)	964 (6.2%)	14596 (93.2%)	75 (0.5%)
形 容 词	22 (0.7%)	369 (11.3%)	2850 (87.0%)	35 (1.1%)

从表中明显可以看出:
(1) 不管是名词，动词还是形容词，字义组合是“2 ”的都占绝大部分。也就是说语素在构词时，一般总是保持原来的意义不变，这也是语素的一个特点。
(2) 从统计上可以看到，只有很少一部分的语素在构词时意义发生了变化。

3.2 汉语语素在构词时意义发生转化的规律

二字名词的意义与构成它的语素义完全不同的有220个词，其中190个是事物名字。其中有中草药名如: 柴胡、丹参、当归、地黄、麦冬; 动物名如: 猫熊、蕲蛇、章鱼; 植物名如: 牛膝、三七、大蓟; 物名如: 麻将、扑满、条几; 官职名如: 尚书、太宰、秘书; 地名如: 澳门、内江、蓬莱; 译名如: 便士、基督、拷贝、拉美。还有一些固定用法表示一些事物的名称，如: 回禄(火灾)、陵迟(酷刑)、东床(女婿)。二字动词的意义与构成它的语素义完全不同的有31个词，它们多是一些固定用法。是源远流长的中华民族文化的产物，如: 耳食、姑息、落草、买帐、涂炭、挖苦、洗练、张罗等。

二字形容词的意义与构成它的语素义完全不同的有22个词，也多是一些固定用法，是社会俗成的产物。如: 狼籍、雷同、糟糕、道地等。

语素在构词时意义绝大多数保持不变，少数变化情况又是有规律可循。这使语素可以在未登录词处理的研究方面起很大的作用。自然语言的词汇随着人们的实践和社会的需要不断地变化发展，旧词的转义，新词的产生，使得不论机器可读词典的规模如何扩大，也终究不能覆盖输入文本中出现的全部单词。汉语中语素基本上是一个封闭集，具有长时间的稳定性。对汉语语素的大规模描写，完全有可能建立一种有效处理汉语未登录词的独特方法，而且这项工作对汉语词法学和语素学的研究，对汉语的计算语言学研究不无助益。

4. 汉语未登录词处理的设想

当一个字串在机器可读词典中找不到相匹配词时，这里存在着两种可能。(1) 字串是未登录的人名、地名或译名。(2) 字串是由一个未登录词。这里注重研究第二种情况(人名、地名、译名的识别与处理这里不作研究)。

基于汉语语素数据库建立一个汉语未登录词处理系统，这里存在着四个待解决的问题。一个字有多个义项，在汉语语素数据库中，一些相近的义项(引申义)归并为一个语素。意义相近但素类不同，如动词性语素的“锁”和名词性语素的“锁”应为同一个语素。一个语素的一个义项(一个语素项)为一个记录。对于一个未登录词中的每个字首先要确定其义项，这是待解决的问题之一。这个问题可以设想用概率统计的办法或借助文献〔6〕基于例子的思想来解决。这样未登录词中的每个字就可以和汉语语素数据库中的一个语素项相对应了。利用数据库中的有关语素项的描述可以形成未登录词的每一个字元素的复杂特征集。这样组成未登录词的语素素性类的序列就形成了。第二个待解决的问题是如何确定未登录词的构词方式。可以设想借助《同义词词林》利用基于例子的方法推断未登录词的构词方式。第三个待解决的问题是，在未登录词的构词方式和组成它的语素素性类序列已知的情况下如何确定未登录词的词性。可以设想用两种知识结合起来确定未登录词的词性。(1)使用概率统计知识。如“名+形”以“主谓”构词方式构成名词的概率和以“定中偏正”构成名词的概率可以从汉语语素数据库中都可以统计出来加以利用。(2)从语言学的角度研究〔2〕。如“名+形”以“主谓”构词方式构成名词时，该形容词有些什么样的语义特征。最后一个待解决的问题是如何确定未登录词的语义。这个问题可以设想用文献〔3〕合一的算法来解决。

参考文献

〔1〕尹斌庸 《汉语语素的定量研究》 《中国语文》1984年第5期

〔2〕王政红 《名、形语素构词分析－复合词构成格式研究》《南睛大学报: 社科版》1992年 第4期

〔3〕YOSHIMURA K. AND SHUDO K. ,Towords the Intelligent Processing of unkownwords, Proc. of INSLU and AI Iizaka,Japan,July 12-15,1992.

〔4〕张登歧 《合成动词的结构及其功能》上海师范大学学报: 哲学版，1992, 2.

〔5〕苑春法、黄昌宁等《汉语语素数据库的建造与应用》 Proceedings of International Conference on Chinese Computing '94 1-4 June, 1994. Singgapore.

〔6〕Tong Xiang, Huang Changning and Guo Chengming, Example Based Sense Tagging of Runing Chinese Text, Proc. of the Workshop on Very Large Corpus, June 22, 1993, Ohio State University Columbus, Ohio USA.