Journal of Chinese Language and Computing 15 (2): (103-112)

现代汉语语义构词规则初探*

亢世勇[1]　许小星[1]　孙茂松[2]

1.烟台师范学院中文系　kangsy46@sohu.com

2.清华大学智能技术与系统国家重点实验室

Submitted on March 1 2004，Revised and Accepted on October 12 2005

摘 要：

本文介绍了汉语语义构词研究的总体思路、《汉语语义构词数据库》的实现，在包含 5 万多个双音合成词的数据库基础上经过分类统计得出的字位在汉语构词中的分布状况、字义与词义的关系类型、汉语语义构词的具体规则，最后简单总结了语义构词规则的特点。

关键词：

现代汉语，语义，构词规则、词汇语义学，语料库，中文信息处理

1 引言

汉语构词法是汉语词汇学、语法学长期以来关注的问题，取得了丰硕成果。近年来随着汉语信息处理的发展，汉语构词法的研究又有了更加实际的意义和需求，又取得了很多有价值的研究成果。综观这些研究成果，主要是从语法形式入手研究构词法，关注并揭示词的结构类型、构词的语素类型与词性之间的关系等。这些成果丰富和发展了词汇学、语法学理论，同时，也为汉语信息处理词语的识别与理解提供了基本依据。基于语义的汉语构词法研究才刚刚起步，目前所能见到的比较有影响的成果有周荐、葛本仪、鲁川、傅爱平等先生的有关研究，这些成果有的提出了研究的思路、有的粗略地分析了双音复合词两个语素之间的语义关系、有的从理论上总结“意合”构词研究与实际应用上的不足。总的来说，这些成果是比较粗略的，限制了其在计算机语言信息处理当中的应用。面向信息处理的基于大规模标注语料库的汉语语义构词规则的深入细致的研究总结势在必行。

* 本项研究得到中国国家社科规划项目（01CYY002）和中国国家 973 项目（G1998030507）的资助

2 汉语语义构词研究的总体思路

语素义与词义的关系一直是词汇语义学关心的理论问题，进入信息时代，其实际意义尤为重要。搞清楚语素（字）义经过整合转化为词义的规则，不仅对人（尤其是留学生）望文（字）生（词）义识读新词语具有重要的指导作用，而且是计算机语言信息处理当中未登录词语的识别以及语义理解的重要依据。随着信息时代的到来，现代汉语词汇迅速发展，“新词”激增，而“汉字”却未增，“这就证明了：汉族人既有用旧字造新词的创造能力，又有看旧字懂新词的领悟力。”“我国人工智能学者和语言学者要通力合作，让电脑模拟汉族人看旧字懂新词的智力。这就应促使‘汉语基因工程’上马，即把字符当作汉语的‘基因’，构建各级‘意序模式库’，阐明造字、造词和造句的‘意合规则’。”这样，计算机就可以利用这些规则去识别那些越来越多的未登录词语，同时“也有助于对外汉语教学，要教外国学生也像汉族人一样，有看旧字懂新词的领悟力和把新知识‘意译’为汉语时有用旧字造新词的创造力”[1]。

如何研究由字义整合转化为词义的规律，即汉语语义构词规则？我们认为拟分三步走。第一步按照一个统一的语义分类体系，分别建立现代汉语字、词的语义分类信息库，尽可能获得全面、系统的字、词的语义分布信息。正是在这种思想指导下，我们本着人机两用的研究理念，引入“字位”的观念（所谓“字位”就是最小的语义构词单位，即形音义一体化的字，每个字位一形、一音、一义），遵循“一字一条、一义一条、意义与语法功能结合、非语素字单独立条”等原则将“国标 GB2312”所定义的 6763 个汉字衍生为 17430 个字位，按照《同义词词林》的语义分类体系给每个字位归了类，录入数据库，建成了大型的《汉字义类信息库》[2]。第二步，在字、词语义分类信息库的基础上，通过统计比较说明字、词语义分布的实际情况以及二者之间的对应关系，为进一步进行语义构词规律的研究提供一个理论基础。经过比较研究，我们发现：（1）字的义类体系和词的义类体系基本一致。（2）字、词在各个义类中的分布比例基本一致。（3）从大类到每个小类，除了个别的类外，字、词的绝对数量多少是一致的，即除了个别类外，绝大部分类字最多、词也是最多的，相反，如果字最少、词也是最少的。可见，字与词在义类上有对应关系，大部分词的意义是在字义的基础上整合而成的。第三步，进行语义构词规律的研究。选取一定数量的双音合成词，利用“汉字义类信息库”对构成双音合成词的每个字进行语义标注[3]，建成大型的《汉语语义构词数据库》，在此基础上进行统计归纳，总结出由字义整合成词义的具体规律[4]。前面两步工作已经完成，本文介绍的是第三步工作的一个初步结果。

3 《汉语语义构词数据库》的实现

以《同义词词林》为基础，结合《现代汉语词典》《新词语大词典》[5]选取了 52366 个双音合成词，然后将《汉字义类信息库》信息用计算机给这些合成词中的每个字标注义类标记和简单释义，经过人工校对，建成大型《汉语语义构词数据库》。数据库中所用的语义类标记大类有：A 人、B 物、C 时间与空间、D 抽象事物、E 特征、F 动作、G 心理活动、H 活动、I 现象与状态、J 关联、K 助语、L 敬语。数据库样例如下：

ID	合成词	合成词的语义类	前字	后字	字、词语义关系类型
2	力争	Je12	Ka19,尽力,努力	Ha02,争夺	6
3	联邦	Di02	Ie09,连接,联合	DI02,国	6
4	联播	Hh03	Ie09,连接,联合	Hh03,传播	6
5	联电	Hi11	Ie09,连接,联合	Bg04,有电荷存在和电荷变化现象	6
6	联合	Ie08	Ie09,连接,联合	Hj30,合并	6
8	联结	Ie08	Ie09,连接,联合	Ie02,发生某种关系,结合	6
9	联军	Di11	Ie09,连接,联合	DI11,军队	6
10	联盟	Di02	Ie09,连接,联合	Ed60,结拜的	6
11	联盟	Hi63	Ie09,连接,联合	Ed60,结拜的	6
12	联赛	Hh07	Ie09,连接,联合	Hh07,比赛	5
13	联系	Ie02	Ie09,连接,联合	Je01,联结，联系（多用于抽象的事物）	6
14	联想	Gb01	Ie09,连接,联合	Gb03,推测	6
15	联姻	Hj51	Ie09,连接,联合	Da01,婚姻	6
16	联翩	Ka11	Ie09,连接,联合	Fd01,很快地飞	4

4 字位在构词中的总体分布

经过对《汉语语义构词数据库》的统计，17430 个字位约有 13972 个字位在双音合成词中出现，占 80.17%。这些字位对 5 万多个双音合成词的覆盖范围如下：

字位频度序列	100	500	1000	2000	3000	4000	5000	6000	7000	8000	9250
覆盖范围(%)	11.3	29.8	43.4	60.4	71.2	78.6	83.9	87.9	90.9	92.9	95.3

前 100 个字位是：子 Kd06、大 Ea03、人 Aa01、不 Ka18、心（心思）Df02、车 Bo21、事 Da01、水 Bg01、军 DI11、白 Ec04、然 Kd06、小 Ea03、手 Bk08、酒 Br12、门 Bn04、身（身体）Bk01、体（身体）Bk01、火 Bg03、风 Bf02、家（家庭、家族）DI05、电 Bg04、女 Eb35、长 Ea01、头（名词后缀、方位词后缀）Kd06、内 Cb05、眼 Bk03、口 Bk04、山 Be04、出 Hj64、地 Bn12、田 Bn12、草 Bh03、民 Aa01、无 Ka18、书 Dk20、道 Bn11、路 Bn11、儿（名词后缀,少数动词后缀）Kd06、春 Ca19、鱼 BI14、国 DI02、房 Bn01、船 Bo22、金（金属）Bm01、开（开始、开拔）Ig01、分 Hj30、场 Cb28、红 Ec01、兵 Ae10、冷 Eb26、文（文章）Dk19、老 Eb36、音 Bg07、物 Ba01、意（意思）Df12、初 Dn04、美 Eb30、处（地方）Cb08、色 Bg06、待 HI07、数 Dn03、话 Dk11、

光 Bg03、力 De04、自（自己）Aa05、刀 Bo09、头 Bk02、别（分离）Ie09、级（等级）DI16、病 Dl01、情（感情）Df04、灯 Bp01、衣 Bq03、地（地面）Bn05、后（未来的）Ca12、油 Br08、查 Hc18、黄 Ec01、加 Ih05、江 Be05、水 Be05、大（程度深）Ka01、实 Ed01、价 Dj02、气（人的精神状态）De03、声 Bg07、年 Ca18、动 Ih01、工（工人、工程）Ae02、称（名字、名称）Dd15、定（确定）Ie06、花 Bh11、入 Hj64、传 Ie01、木 Bm03、石 Bm04、法（法律）DI25、死 Ib03、评 Hc20、天（天空）Cb07。

这些字位在构成 5 万多个双音合成词中出现的次数、数量与所占比例如下：

出现次数	627	458	318	213-268	100-193	90-99	80-89	70-79
字位数量	1	1	1	3	36	14	20	42
比例（%）	0.007	0.007	0.007	0.021	0.25	0.098	0.14	0.29
出现次数	60-69	50-59	40-49	30-39	20-29	10-19	9	8
字位数量	45	101	136	183	608	1641	296	365
比例（%）	0.32	0.71	0.95	1.28	4.26	11.51	2.08	2.56
出现次数	7	6	5	4	3	2	1	
字位数量	419	552	639	851	1135	2049	5010	
比例（%）	2.93	3.87	4.48	5.97	7.96	14.37	35.13	

出现在双音合成词前面的字位有 8931 个，出现在后面的字位 10647 个，前后两个位置上都有的字位有 5606 个，只出现在前面的有 3325 个，只出现在后面的有 5041 个。可见大部分字位在构词时位置是比较固定的。这也可以作为未登录词识别的一个有利条件。

5 字义与词义关系类型

经过对《汉语语义构词数据库》中 5 万多个合成词的意义与构成合成词的两个字位的意义之间的关系的考察，我们把字义与词义的关系归纳为以下八种类型（此处 A、B 代表构成合成词中的前后两个字位）。

（1）A+B=A=B　（2）A+B=A　（3）A+B=B　（4）A+B=C

（5）A+B=A+B　（6）A+B=A+B+D　（7）A+B=A+D　（8）A+B=D+B

第一种方式是指 A、B 是同义的，词义就是其中的一个字位义；第二种方式是指词义只保留了字位 A 的意义，B 的意义已经不存在了，即带有后缀的词以及一些偏义复词；第三种方式是指词义是字位 B 的意义，而字位 A 已经不存在了，即带有前缀的

词以及一些偏义复词；第四种是指词义和字位义之间没有任何明显的联系，AB 组合后产生了新的意义，词的引申义和比喻义也属于此类；第五种是指词义是由 A、B 两个字位义相加而成。第六种是指词义包含了 A、B 两个字位义，但是又加上了其他的意义（D），主要包括改变词性、前一个字位义与后一个字位义有领属关系、某个字位改变词性、带有某种陪义；第七种是指字位 B 的意义已经变成了其他意义（D），词义由 A、D 两个字位义构成，有的又加上了其他的意义；第八种是指字位 A 的意义已经变成其他意义（D），词义由 D、B 两个字位义构成，有的又加上了其他的意义。

各种类型包含的合成词的数量与所占比例如下：

类型	1	2	3	4	5	6	7	8
合成词数量	4035	1031	297	4201	14455	23562	2780	1886
比例（%）	7.71	1.97	0.57	8.02	27.60	44.99	5.31	3.60

在这八种类型中只有第四种（A+B=C）看不出字义与词义的关系，其他 7 种字义与词义都有明显的关系，第四种只占 8.02%，而其他七种加起来占 91.98%，数据表明，字义与词义有密切的关系，可以由字义推知词义。造成每类当中 A、B 两个字位与词义关系的具体情况，我们将进一步研究。

6 双音合成词语义构词的具体规则

通过对《汉语语义构词数据库》的分类、归纳、统计，从语义大类着眼，初步归纳了汉语双音合成词语义构词的具体规则，并将这些规则进一步归纳为四个大的类型。下列规则中“A、B、C、D、E、F、G、H、I、J、K、L”为语义类大类的标记，具体规则中“AB”表示双音合成词中前一个字位的语义类为 A 类、后一个字位的语义类为 B 类，其他类推。

6.1 同类规则

构成双音合成词的两个字位属于同一个语义类，所构成的词的语义类与其基本相同。AA 的词义 100%为 A 类，BB 的词义 88.89%为 B 类，CC 的词义 83.82%为 C 类，DD 的词义 86. 83%为 D 类,EE 的词义 84.98%为 E 类,FF 的词义 68.03%为 F 类,GG 的词义 84.5%为 G 类，HH 的词义 88.41%为 H 类，II 的词义 74.12%为 I 类，JJ 的词义 74.72%为 J 类，KK 的词义 82.08%为 K 类，LL 的词义 60%为 L 类。可见，除了 II、JJ、LL 三类外，其他类构成的词义与其同类的都在 80%以上。属于同类构成的双音合成词共有 17565 个，占 33.54%。

6.2 后向型规则

构成双音合成词的两个字位属于不同的语义类，所构成的词的语义类与后一个字位的语义类相同。属于这一类的有（后面的数字为占该类的百分比）：AB 类 63.63%，AC 类 60%，AD 类 63.46%，AH 类 49.49%，BC 类 60.73%，BD 类 69.84%，BH 类 65.84%，BI

类 51.23%，BJ 类 48%，CA 类 79.07%，CB 类 71.95%，CD 类 65.77%，CH 类 62.33%，CI 类 41.43%，DA 类 88. 22%，DB 类 62. 54%，DC 类 56.18%，DH 类 50.13%，EA 类 91.08%，EB 类 77.30%，EC 类 72.2%，ED 类 71.6 %，EF 类 52.09%，EH 类 56.05%，FA 类 74.47%，FH 类 55.93%，HA 类 59.55%，HD 类 51.16%，HE 类 53.49%，IA 类 58.2%，IH 类 57. 6%，JG 类 51.06%，KA 类 73.27%，KE 类 56.54%，KG 类 54.29%，KH 类 57.27%，KI 类 52.73%，LE 类 50%，LH 类 71.43%。可见，这些类构成的词的语义类 50%以上的都与后一字位的语义类相同，多数在 60%以上。与后一字位语义类不同的那些词语多数分布在很多类中，但也有个别的几个类比例高达 20%。属于后向型规则构成的双音合成词共有 18020 个，占 34.41%。后向型的语义重点落在双音合成词的后一个字位上，主要包括语法构词上的两种形式即偏正式结构和加前缀式。

6.3 前向型规则

构成双音合成词的两个字位属于不同的语义类，所构成的词语义类与前一个字位的语义类相同。属于这一类的有（后面的数字为占该类的百分比）：AE 类 59.15%，AF 类 75%，AG 类 56.25%， AK 类 87.5%，BA 类 86.25%，BK 类 85.46%，CK 类 72. 38%，DK 类 65.68%，EK 类 47.23%，FJ 类 51.72%，FK 类 51.95%，GJ 类 50%，GK 类 54.46%，HF 类 60.22%，HG 类 53.97%，HJ 类 55.98%，HK 类 53.85%，HL 类 100%，IK 类 45.52%，JK 类 46.97%。可见，这些类构成的词的语义类 50%以上属于前一个字位的语义类，个别类的百分比在 50%以下，该类构成的其他词的语义类分布在比较多的类里，比例都很小。属于前向型构成的双音合成词共有 1984 个，占 0.39%。前向型的语义重点落在双音合成词的前一个字位上，主要原因是后一个字位是意义比较虚灵的后缀或类后缀，整个词语的意义由前一个具有实在意义的字位决定。

6.4 无向型规则

构成双音合成词的两个字位的语义类不同，所构成的词的语义类比较多，但其中有一些主要的类比例比较高。这些类有：AI 类词义为 A、D、I 类的分别占 31.81%、22.73%、18.18%，BE 类词义为 B、E 类的分别占 35.60%、43.69%，BG 类词义为 A、D、G 类的分别占 10.71%、25%、39.29%，CE 类词义为 A、C、E 类的分别占 13.64%、27.28%、32.95%，CF 类词义为 C、F、H 类的分别占 11. 76%、35. 29%、35. 29%，CG 类词义为 D、G 类的分别占 21. 43%、42.86%，CJ 类词义为 C、D、E 类的分别占 25%、16.67%、16.67%，DE 类词义为 A、D、E 类的分别占 13. 21%、39. 15%、36. 79%，DF 类词义为 D、F、H 类的分别占 28%、20%、28%，DG 类词义为 D、G 类的分别占 42.59%、39.81%，DI 类词义为 D、I 类分别占 35.54%、40.50%，DJ 类词义为 D、J 类的分别占 35.71%、28. 57%，EG 类词义为 E、G 类的分别占 36.75%、43.59%，EI 类词义为 E、I 类的分别占 30.75%、48.06%，EJ 类词义为 E、J 类的 40.23%、27.59%，FB 类词义为 B、F、H 类的分别占 27.71%、33.19%、22.34%，FC 类词义为 C、D、F 类的分分别占 32%、20%、20%，FD 类词义为 D、F、H 的分别占 34.25%、19.34%、24.86%，FE 类词义为 E、F 类的分别占 26.67%、56%，FG 类词义为 F、G 类的分别占 34.69%、30.61%，FI 类词义为 F、I 类的分别占 46.56%、37.02%，GA 类词义为 A、E、G 的分别占 48.21%、21.43%、19.64%，GB 类词义为 B、G 类的分别占 37.88%、27.27%，GD 类词义为 D、G 类的分别占 37.99%、32.52%，GE 类词义为 E、G 类的分别占 42.11%、41.35%，GF 类词义为 F、G 类的分别

占 40.98%、27.87%，GH 类词义为 G、H 类的分别占 30.28%、46.79%，GI 类词义为 G、I 类的分别占 27.72%、49. 5%，HB 类词义为 B、H 类的分别占 34.27%、47%，HC 类词义为 C、H 类的分别占 28.3%、44.74%，HI 类词义为 H、I 类的分别占 48.37%、33.1%，IB 类词义为 B、I 类的分别占 37.1%、30.67%，IC 类词义为 C、I 类的分别占 32.18%、33.17%，ID 类词义为 D、H、I 类的分别占 35.82%、20.28%、31.84%，IE 类词义为 E、I 类的分别占 38.01%、41.58%，IF 类词义为 F、I 类的分别占 31.3%、41.3%，IG 类词义为 G、I 类的分别占 39.73%、23.29%，IJ 类词义为 I、J 类的分别占 42.14%、26.43%，JA 类词义为 A、J 类的分别占 38.71%、24.73%，JB 类词义为 B、H、J 类的分别占 16.49%、22.34%、23.37%，JC 类词义为 C、J 类的分别占 38.6%、18.81%，JD 类词义为 D、H、I、J 类的分别占 25.09%、15.57%、15.93%、14.84%，JE 类词义为 E、I、J 类的分别占 35.39%、21.93%、15.73%，JH 类词义为 H、J 类的分别占 49.36%、24.04%，JI 类词义为 I、J 类的分别占 45.57%、30.77%，KB 类词义为 B、K 类的分别占 37.25%、16.34%，KC 类词义为 C、K 类的分别占 37.58%、22.93%，KD 类词义为 D、K 的分别占 40.11% 、21.39%，KF 类词义为 F、H 类的分别占 44.26%、14.75%，KJ 类词义为 J、K 类的分别占 49.26%、24.40%。由此可见，这些规则构成的词义所属的语义类的确比较多，既有前向的，也有后向的，还有其他的，但仔细比较我们列出的类及其数据，不难发现，这些类中词的义类尽管比较多，但都和构成该词的前后两个字位密切相关，即和前后两个字位同类的最多，将与前后两个字位同类的加起来多数都在 60%以上，有的能够达到 80%以上，从这一点看，我们可以将这一类概括为前后向的，其中有的稍偏前向、有的稍偏后向、有的干脆是对半。属于无向型构成的双音合成词有 14797 个，占 28.27%。这一类具体规则最复杂，但构成的双音合成词相对较少。

7 汉语语义构词规则的特点

通过对具体规则的归纳统计，我们发现语义构词规则大致具有以下特点。

（1）以上四个类型的规则覆盖范围不同，大致构成如下不等式：后向型规则>同类规则>无向型规则>前向规则，后向型规则比例最高。这四个类型的规则在分布上是互补的。

（2）从这些具体规则，我们可以看到，尽管两个语义类的字位组合在一起构成的合成词语义类比较复杂，几乎每一种都可以构成多个语义类的词语，但我们也看到，其中数量最多的类还是和构成双音合成词的字位的语义类相同的语义类，即 AA 类全部为 A 类，AB 类最多的是 A 类、B 类，AD 类最多的是 A 类、D 类，等等。根据字位与词义的亲近度，四个类型的规则可以构成如下不等式：同类规则> 后向型规则> 前向规则>无向型规则。同类规则构成的词语义类和字位的语义类相同的最多。总之，四类规则共同的特点是词义都和前后两个字位有密切的关系，可以通过两个字位在一定程度上推出词的语义类，这个比例能够达到 60%以上。

（3）语义构词规则从照理论上来说应该有 144 种，实际只有 130 种，其中 14 种没有。这 130 种按实际包含词语数量构成了下列不等式，括号中为合成词的数量。

BB（5004）>HH（3509）>EE（2609）>DD（2556）>EB（2530）>ED（2105）>HD（1979）>HB（1381）>BD（1041）>IB（1026）>II（966）>FB（912）>DB（842）>EA（822）>EH（794）>ID（770）>CC（719）>IH（711）>AA（670）>HI（578）>CB（573）>BK（572）>GG（560）>JD（539）>KH（470）>DA（469）>BC（466）>HE（453）>HA（446）>CD（444）>FF（442）>EI（437）>FH（395）>IE（390）>HJ（389）>DH（377）>KD

（375）>HC（365）>DC（ 344）>GD（ 331）>BE（ 309）>AD（ 301 ） >JB（289）>KE（284）>HF（279）>JK（279）>EC（272）>EK（267）>KK（264）>FI（263）>EL（262）>JJ（261）>HK（258）>BH（243）>BA（240）>HG（239）>EG（234）>IF（229）>GH（ 215）>KJ（208）>DE（208）>IC（203）>BI（203）>EF（202）>KI（199）>FD（179）>JE（175）>EJ（174）>JI（167）>KG（167）>DK（159）>KC（158）>FK（155）>KB（153）>FE（151）>CH（144）>IJ（138）>IK（134）>GE（130）>IA（121）>DI（120）>DG（108）>GK（107）>KA（102）>JC（100）>AH（99）>AB（99）>GI（95）>JG（93）>JA（91）>FJ（88）>CA（86）>CE（86）>AK（80）>FC（74）>IG（72）>AE（71）>CI（67）>GJ（66）>KF（64）>GB（64）>GA（56）>DH（55）>GF（54）>AC（50）>FG（48）>FA（46）>JF（40）>BG（28）>BJ（25）>DF（25）>CJ（24）>AI（23）>GC（19）>CF（17）>AG（16）、AJ（16）>CG（14）>LK（6）>LL（5）>LE（4）、HL（4）、AF（4）>LD（3）>KL（2）、LG（2）>GL（1）、LJ（1）、IL（1）、LI（1）。

可见，BB、HH、EE、DD 构词能力最强，构成的词最多。

（4）每个义类的字位构词能力不尽相同，按照构词频率构成下列不等式，括号里的数字是频度。B 物（21189）>H 活动（17242）>D 抽象事物（16025）>E 特征（15685）>I 现象与状态（7928）>K 助语（5381）>C 时间与空间（5281）>A 人（4604）>F 动作（4223）>J 关联（3634）>G 心理活动（3339）>L 敬语（56），可见，这 12 类中，B 物类构词能力最强，L 敬语类构词能力最差。这个不等式序列和各类字位数量多少构成的不等式序列一致，说明每类构词能力的强弱决定于该类字位的多少。其更深层的原因是物体、活动是宇宙世界的主体，词汇是反映主客观世界的，由物体、活动产生的词汇占绝对多数。

（5）每个义类字位在双音合成词前后两个位置上出现的多少也不相同。出现在双音合成词前一个位置上的义类根据频度构成下列不等式，括号里的数字是频度。E 特征（10763）>H 活动（9928）>B 物（8267）>D 抽象事物（5325）>I 现象与状态（4788）>K 助语（2808）>F 动作（2750）>C 心理活动（2296）>J 关联（2176）>G 心理活动（1735）>A 人（1433）>L 敬语（37）。这个不等式和 7．4 的不等式不同，E 特征类由原来的第四位提前到第一位，这是因为汉语偏正式构词数量最多，E 特征类往往充当偏正式构词中偏的成分，所以出现在双音合成词前一个位置上的较多。出现在双音合成词后一个位置上的义类根据频度构成下列不等式，括号里的数字是频度。B 物（12916）> D 抽象事物（10697）> H 活动（7314）>E 特征（4918）>A 人（3167）>I 现象与状态（3139）>C 心理活动（2985）>K 助语（2573）>F 动作（1473）>J 关联（1458）>G 心理活动（1604）>L 敬语（19）。这个不等式序列和 7．4 里各义类构词频度构成的不等式序列大体是一致的，和前一个位置上义类构成的不等式序列形成一个互补。B 物、D 抽象事物、H 活动、E 特征、A 人在双音合成词后一个位置上出现的频率更高。这是我们从语义类上得出的结论。这个结论可以从注重形式的语法构词中得到验证。因为汉语构词中定中式偏正结构占 53%以上、联合结构占 27%以上、动宾结构占 13%以上，这三种构词方式的总和在 92%以上，而这些结构中处在后一个位置上的大多都是事物和人一类的，因为人和事物是宇宙世界的主体，其他都是由此而生发的，在词汇发展的过程中也遵循了以人和事物为中心附加其他属性而生成新词语的规则。

（6）几乎每一条规则都可以构成属于 A 类（人）的词语，说明了多数字位都与人相关。

注释：

[1]鲁川.2003.汉语的根字和字族——面向知识处理的汉语基因工程.汉语学习,(3):1-10
[2]亢世勇等. 2002.《汉字义类信息库》的研究与实现. 汉语语言与计算学报，7(2):129-142
[3]语义类代码采用《同义词词林》的代码。
[4] 亢世勇等. 2002. 现代汉语字、词义类分布统计研究. 第三届中文词汇语义学术研讨会论文集. 台湾中研院语言所
[5]《现代汉语词典(1996 版)》，中国社会科学院语言所词典室编，商务印书馆，1996 年；《新词语大词典》，亢世勇主编，上海辞书出版社，2003 年。

参考文献：

戴昭铭. 1988. 现代汉语合成词的内部结构与外部功能的关系. 语文研究，(4)
符淮青. 1996. 词义的分析和描写. 北京：语文出版社
葛本仪. 2001. 现代汉语词汇学. 济南：山东人民出版社
刘叔新. 1985. 汉语复合词内部形式的特点与类别. 中国语文，(3)
刘叔新. 1985. 汉语描写词汇学. 北京：商务印书馆
苑春法等. 1998. 基于语素数据库的汉语语素及构词研究. 世界汉语教学，(2)
苑春法. 2000. 汉语构词研究. 语言文字应用，(1)
郑家恒. 2001. 二字词词义组合推理方法的研究. 中文信息学报，(6)
周　荐. 1991. 复合词词素间的意义结构关系. 语言研究论丛（第六辑）. 天津：天津教育出版社
周　荐. 1995. 复合词构成语素的选择. 中国语言学报，(7)
周　荐. 1999. 双字组合与词典收条. 中国语文. (4)

The Research on the Modern Chinese Semantic Word-Formation

Kang Shi-yong[1] Xu Xiao-xing[1] Sun Mao-song[2]

1. Chinese Department of Yantai Normal University
2. National Key Lab.of Intelligent Technology and Systems, Tsinghua University

Abstract

The Chinese word-forming rules are always interesting subjects in the fields of the Chinese lexicology and grammar. And with the development of the Chinese information processing, the research of Chinese word-forming rules is endowed with the more practical significance and requirements. Three steps will be taken to investigate the rules of the Chinese semantic word-forming. First of all, the semantic classifying information bases of the modern Chinese character and word must be established respectively; Second, according to the semantic classifying information bases of the modern Chinese character and word, the practical semantic distributions of character and word and the corresponding relation between character and word are investigated and illuminated by comparison and statistics, which provides a theoretic foundation for further investigation on the law of semantic word-forming; and finally, what is need to be done is to establish large-scale “the data base of the Chinese semantic word-forming”, and to summarize the particular law of integrating literal meaning into acceptation. The former two steps are accomplished, and the primary results of the third step are presented in this paper. By examining the relation between the meanings of more than fifty thousand disyllabic compound words of “the data base of the Chinese semantic word-forming” and the meanings of the two graphemes composing of a disyllabic compound word, the relation between the literal meaning and the word meaning boils down to eight types. According to classification and statistics of various types in the data base, the distributions of the Chinese word-forming about the graphemes, types of the relation between the literal meaning and the word meaning and particular rules of the Chinese semantic word-forming are given. At last, characters of semantic word-forming rules are summarized in brief.

Keywords

Modern Chinese, semantic, word-forming rules, the lexicology semantics, corpus, Chinese information processings