文章编号：1003-0077（2011）00-0000-00

面向普通未登录词理解的二字词语义构词研究 *

吉志薇，冯敏萱

（1.南京师范大学，江苏省 南京市 210097；2.南京师范大学，江苏省 南京市 210097）

摘要：把词素作为基本资源，从语义上寻找他们组合成词的规律，可以辅助自然语言理解。该文首先参照《现代汉语词典》和知网标注了二字词的词素意义，继而从意合结构、意根分布、意指方式、意变类型四个角度标注了词素间的词化意义，最后综合词素意义和词化意义，在定量统计的基础上建立了一个二字词的语义描写体系。通过对论坛及《现代汉语词典》的新词进行实验，我们发现二字词的语义构词研究在普通未登录词的理解中具有一定的应用价值。

关键词：二字词；普通未登录词；语义构词

中图分类号：TP391　　**文献标识码：**A

A Study of Semantic Word-formation of Two-Character Words for Understanding of Common Unknown Words

Zhiwei Ji, Minxuan Feng

(1.Nanjing Normal University, Nanjing, Jiangsu 210097, China; 2.Nanjing Normal University, Nanjing, Jiangsu 210097, China)

Abstract: The approach of taking the morphemes as the basic resource to trace the word-formation discipline in semantics can help understand natural language .We first labeled the sense of the front and back morpheme of the two-character words by referring to the Modern Chinese Dictionary and HowNet. Then we labeled the lexicalized meaning among morphemes from the perspectives of the structure of semantic combination, the distribution of semantic root, the mode of semantic combination and the type of semantic variation. Finally, we combined the morpheme meaning with lexicalization meaning to set up a semantic descriptive system of two-character words which based on quantitative statistics. Depending on the system, we identified and understood the two-character words from BBS and the Modern Chinese Dictionary. The result shows that the study on the semantic word-formation of two-character words has a certain practical value on the understanding of common unknown words.

Key words: two-character words; semantic word-formation; common unknown word

1. 引言

根据黄昌宁的研究，未登录词中除去日期、时间、百分数、人名、地名、机构名等专名以及派生词的那一部分就是普通未登录词，也有学者称之为新词。在中文信息处理领域，未登录词是影响分词精度最主要的因素之一。未登录词中的新词数量众多，和现代汉语中基本词、常用词等在构词规律上有很大的相似性，具有重要的研究价值。在现代汉语词汇中，二字词占主体地位，对其研究有助于我们了解大多数词汇的构词规律。与词相比，词素数量相对有限，在一个相对封闭的范围内，对二字词的词素进行穷尽式考察可以帮助我们更好地发现一个字串之所以成为词的理由。

2. 确定研究对象

*收稿日期:2015 年 6 月 10 日　定稿日期：2015 年 8 月 10 日

首先利用计算机提取《现代汉语词典》[†]（第六版）中的所有二三字词[‡]和构成这些二三字词的词素，进而计算每个词素的构词量，最终选取构词能力最强的 50 个词素，在《现汉》中查找这些词素构成的二字词并将其录入 excel 表格中。

根据本文的研究目标，我们排除以下几类词汇：（1）标注有〈方〉的方言词；（2）意义虚化、读音弱化、位置固定、能产性强的典型词缀构成的词，以“子”为例，它有两种用法：有意义区别作用的自由和半自由词素，如“父子”、“男子”等；没有意义区别作用的不自由词素，如“帽子”、“旗子”等。后一类就属于典型词缀构成的词。（3）单纯词，如“卡车”。（4）简称，如“光驱”。（5）专名，如“道光”。（6）《现汉》（第六版）中新出现的二字词。

3. 构建标注体系

参照鲁川的词义方程式，本文将词素间的静态关系，即义类组合，称为词素意义；将词素间的动态关系，即词素和词素化合时产生的词素义之外的意义，称为词化意义；综合词素意义和词化意义即可得到一个二字词的释义模式。

3.1 词素意义的标注说明

本文首先依据《现汉》标注词义，又参照知网对前字和后字的义项进行归类，最后根据词义为前字和后字选择相应的义类。以“滚水”为例，由《现汉》可知“滚水”的词义是“正在开着的或刚开过的水”。“滚”字的义项有：

序号	词素性质	义项	义类
①	v	滚动；翻转	变空间位置
②	v	走开；离开(含斥责意)	变空间位置
③	v	(液体)翻腾，特指受热沸腾	外观变
④	d	表示程度深，特别	特性值
⑤	v	使滚动；使在滚动中沾上(东西)	变空间位置
⑥	v	同“绲③”	使消失
⑦	n	姓	特性

观察上表可得“滚”字 7 个义项分属 5 个义类，义项①、②和⑤均属于“变空间位置”这一义类，需要进行合并。根据词义“正在开着的或刚开过的水”可知，“滚”字在“滚水”一词中使用的是义项③，所属义类为“外观变”。同理，“水”字在“滚水”一词中使用的义项是“最简单的氢氧化合物”，所属义类为“液”。因此，“滚水”的词素意义应为“外观变+液”。8984 个二字词共包含 2268 个不同的词素，通过标注，我们构建了基于这 2268 个词素的词素-义类数据库。

3.2 词化意义的标注说明

词化意义主要是从意合结构、意根分布、意指方式和意变类型四个方面进行界定：意合

[†]下文简称《现汉》。

[‡]尽管本文的研究目标是二字词，但考虑到三字词的意义以及进一步研究的需要，我们选择高频词素时也兼顾到了三字词。

结构说明词素和词素间的语法关系；意根分布是指二字词意义核心所在的位置；意指方式说明词素义和词义之间的关系；意变类型立足于历时发展，说明词义变化的类型。具体分类如表 1：

表 1：词化意义标注体系

意合结构（YH）	并合（B）	词素间是并列、选择等关系，如“国家”。
	加合（J）	词素间是修饰、限定关系，两个词素可用“的”来连接，如“疑点”。
	摹合（M）	词素间是修饰、限定关系，两个词素不可用“的”来连接，如“初学”。
	配合（P）	有两种类型：一种词素间是支配、关涉的关系如 “司机”；一种词素间是陈述的关系，如“心烦”。
	续合（X）	有两种类型：一种词素间表示连续的几个动作，如“取法”；一种词素间表示补充的关系，如“漂白”。
	衬合（C）	有两种类型：一是“实字”与“虚字”的组合，如“以上”；一是“物体”与“量词”的组合，如“车辆”。
	叠合（D）	重叠词，如“上上”。
意根分布（YG）	前根（Q）	前字为根，如“服老”。
	后根（H）	后字为根，如“步行”。
	二根（E）	二字为根，如“伟大”。
意指方式（YZ）	惯指（G）	词义基本等于词素义的简单化合，“惯”指某物。
	加指（J）	词义中新“加”了其他内容。
	失指（S）	词义中“失”落了某个词素义。
	另指（L）	词义“另”有所指，类似于词义引申中的换喻，强调其相关性。
	仿指（F）	词义运用了相“仿”的比喻，类似于词义引申中的隐喻，强调其相似性。
	专指（Z）	词义具有“专”门意义，一般用于专业术语。
意变类型（YB）	特指（T）	词义在变化过程中缩小，出现“特”指义。
	泛指（F）	词义在变化过程中扩大，出现“泛”指义。

另外，在标注失指（S）、另指（L）和仿指（F）时还需标出产生失落、换喻或隐喻的词素的位置（YZWZ），具体有三种：前字（Q）、后字（H）以及整词（Z）。

3.3 标注示例

词型	词义	词素意义	词化意义	释义模式
矮小	又矮又小	量度值+量度值	BEG	BEG+(量度值+量度值)
案发	案件发生	事情+存现	PHG	PHG+(事情+存现)
车貌	车辆的外观	器具+外观	JHG	JHG+(器具+外观)

4. 二字词语义描写体系的构建

对 8984 个二字词的词素意义和词素间的词化意义逐一进行标注和统计，可得词素意义分布表（见表 2）、词化意义分布表（见表 3）和释义模式分布表（见表 4）。综合词素-义类数据库，我们构建了二字词的语义描写体系。

表 2：二字词词素意义分布表部分示例

词素意义	出现频率	实例
特性值+人	1.60%[§]	老媪、骄子、大敌
变空间位置+变空间位置	1.45%	出游、发射、行进
人+人	1.29%	儿孙、父老、妻小
变空间位置+生物部件	1.20%	闭口、点头、撒手
特性值+特性值	1.07%	工巧、机灵、老诚
生物部件+生物部件	1.07%	额头、口齿、毛发
特性值+特性	1.01%	凶气、正风、肥力
关系值+人	0.88%	外宾、国人、敌手
特性+特性	0.76%	才力、操行、风华
变空间位置+器具	0.75%	上镜、跑车、出槽

表 3：二字词词素间词化意义分布表部分示例

词化意义	YH	YG	YZ	YB	频率	实例
JHG	J	H	G		40.95%	傲然、口惠、爱情
PQG	P	Q	G		14.08%	乘机、点题、罢工
MHG	M	H	G		6.86%	不济、白描、伴生
BEG	B	E	G		6.16%	矮小、茶水、出没
XEG	X	E	G		3.55%	颁行、编发、拆分
JHLZ	J	H	LZ		2.11%	白领、陛下、寡人
JHFZ	J	H	FZ		2.08%	鳌头、儿戏、柴门
JHZ	J	H	Z		1.89%	平光、单眼、心皮
PQFZ	P	Q	FZ		1.73%	到家、冲天、打鼓
JHJ	J	H	J		1.68%	手稿、白眼、密电

[§]本文所有数据均四舍五入精确到小数点后两位。

表 4：二字词释义模式部分示例

常用释义模式	出现频率	实例
JHG+(特性值+人)	1.30%	差生、副手、坏人
JHG+(特性值+特性)	0.91%	大胆、中级、狂气
MHG+(变空间位置+变空间位置)	0.71%	环行、晃动、迸发
JHG+(人+人)	0.71%	帝子、妇人、东家
JHG+(关系值+人)	0.67%	敌人、国手、生客
BEG+(特性值+特性值)	0.57%	工巧、机灵、老练
PQG+(变空间位置+生物部件)	0.56%	抄手、点头、倒头
JHG+(特性值+精神)	0.51%	粗心、骄人、勇气
BEG+(变感知+变感知)	0.41%	查点、建白、区分
PQG+(变空间位置+器具)	0.40%	扒车、发球、装机

5. 二字词语义描写体系的应用

（1）实验对象

根据研究目标，本文从天涯论坛一则名为“你好，陌生人！日记接龙，献给八卦的筒子们”的帖子[**]中选取 2014 年 4 月至 2015 年 4 月的所有留言，经过简单的人工处理，得到共计 3128 个字的实验语料。

（2）实验过程

分别利用陈小荷的中文信息处理实验平台和中科院的 ICTCLAS 对实验语料进行分词。选取两种分词软件均切分有误的二字词，可将其分成两类：一是专名，如“倒 春寒、回 南 天、汪 峰、徐 静 蕾、齐 秦、星 某 克”等；一是普通未登录词，如“舍 友 、前 路、自 处、煎 蛋、水 煮、微 博、发 帖、命 格、妹 纸、脑 抽、驴 饮、扎 口”等。应用二字词的语义描写体系对分词有误的 22 个普通未登录词进行识别和理解。

利用词素-义类数据库自动标注二字词前后字的义类组合，以“安监”为例，首先从词素-义类数据库中分别提取“安”和“监”的所有义类，可知“安”有 6 种义类，“监”有 2 种义类；然后将“安”的所有义类逐一与“监”的所有义类进行组合，最终共得 12 种义类组合类型（见表 5）。依据词素意义分布表，计算机会对所有义类组合进行自动排序，同时返回排名最高的义类组合作为该词最有可能的词素意义。仍然以“安监”为例，观察表 5 可得，“安监”的义类组合中，排名最高的是“变空间位置+变感知”。

表 5：“安监”的义类组合类型

新词	义类组合	出现频率
安监	变空间位置+变感知	**0.21%**
安监	变领属+变感知	0.07%

[**] http://bbs.tianya.cn/post-funinfo-3189865-1.shtml

安监	变良态+变感知	0.06%
安监	变空间位置+处所	0.04%
安监	变领属+处所	0.02%
安监	状况值+处所	0.03%
安监	使存现+变感知	0.02%
安监	使存现+处所	0.02%
安监	变良态+处所	0%
安监	情绪+变感知	0%
安监	情绪+处所	0%
安监	状况值+变感知	0%

依据释义模式分布表，计算机会为已经确定词素意义的新词标注释义模式并进行排序，同时返回排名最高的释义模式，据此推测新词的词义。观察表 6 可得，词素意义为“变空间位置+变感知”的释义模式共有 5 种，其中“XEG+(变空间位置+变感知)”的排名最高，因此“安监”最有可能的释义模式就是“XEG+(变空间位置+变感知)”。

表 6：“安监”的释义模式排序

新词	释义模式	出现频率
安监	XEG+(变空间位置+变感知)	**0.11%**
安监	MHG+(变空间位置+变感知)	0.06%
安监	BELQ+(变空间位置+变感知)	0.02%
安监	PQG+(变空间位置+变感知)	0.01%
安监	PQLZ+(变空间位置+变感知)	0.01%

（3）实验结果

观察表 7 可得，除了“自处”一词，其他 21 个普通未登录词的词素意义均在词素意义分布表中出现过，即这 21 个词含有辅助计算机自动识别的词素意义类型，可被计算机识别，识别率为 95.45%。

表 7：22 个普通未登录词的识别结果

二字词	词素意义	频率
逗比	特性值+变感知	0.35%
发帖	变感知+读物	0.19%
乖萌	特性值+存现	0.10%
煎蛋	使存现+生物部件	0.03%
裸辞	存现+外观	0.04%
妹纸	人+器具	0.08%
萌屎	存现+排泄物	0.01%

命格	特性+特性	0.76%
脑抽	生物部件+变空间位置	0.07%
弃楼	变领属+机构	0.02%
前路	变空间位置+建筑物	0.36%
舍友	关系值+人	0.88%
微博	特性值+量度值	0.06%
微博	特性值+变莠态	0.06%
微信	特性值+特性值	1.07%
雾霾	液+现象	0.03%
心塞	生物部件+无生物部件	0.17%
艳遇	特性值+时间	0.13%
扎口	变空间位置+生物部件	1.20%
装 X	变空间位置+信息	0.10%
自虐	{firstPerson}+特性值	0.01%
作死	做+做	0.22%
自处	…	0%

我们将词素意义分布表的构词量百分比[††]作为标准，结合构词量，在降序排列的词素意义分布表中以 20%左右的梯度进行分类，设定了五个参照集（见表 8）。在这个表格中，处于第 1 参照集的词素意义构词量最多，处于第 5 参照集的词素意义构词数量最少。构词数量越多，证明此类词素意义构词能力越强，因此五个参照集中，第 1 参照集的构词能力最强，剩下四个的构词能力依次降低。

表 8：五个词素意义参照集

参照集	构词量百分比	词素意义	构词量	频率
1	20.11%	特性值+人	144	1.60%
		…	…	…
		变属性+变属性	30	0.33%
2	21.04%	变属性+生物部件	29	0.32%
		…	…	…
		无生物部件+器具	12	0.13%
3	21.72%	态度+精神	11	0.12%
		…	…	…
		精神+无生物部件	5	0.06%

[††]构词量百分比是指在 8984 个二字词中，一定范围的词素意义能构成二字词的比例。

4	24.30%	数量值+思想	4	0.04%
		…	…	…
		人+钱财	2	0.02%
5	12.83%	情绪+关系	1	0.01%
		…	…	…
		使存现+思想	1	0.01%

在 22 个普通未登录词中，有 6 个二字组处于第 1 参照集中，成词可能性非常大；有 4 个二字组处于第 2 参照集中，成词可能性比较大；有 5 个二字组处于第 3 参照集中，成词可能性一般；有 4 个二字组和 2 个二字组分别处于第 4 和第 5 参照集中，成词可能性比较小。

表 9：21 个二字组的成词可能性分布表

参照集	二字组
1	扎口、微信、舍友、命格、前路、逗比
2	作死、发帖、心塞、艳遇
3	乖萌、装 X、妹纸、脑抽、微博
4	裸辞、煎蛋、雾霾、弃楼
5	萌屎、自虐

利用释义模式分布表标注各词，结果如表 10。观察可得，22 个词中，只有“发帖、命格、舍友、雾霾、作死”5 个词的释义模式可以大致推测出正确的词义，理解正确率为 22.73%。

表 10：22 个普通未登录词的释义模式

二字词	词素意义	频率
逗比	MHG+(特性值+变感知)	0.22%
发帖	PQG+(变感知+读物)	0.12%
乖萌	MHG+(特性值+存现)	0.08%
煎蛋	PQG+(使存现+生物部件)	0.02%
裸辞	PQG+(存现+外观)	0.02%
妹纸	JHG+(人+器具)	0.07%
萌屎	PQG+(存现+排泄物)	0.01%
命格	BEG+(特性+特性)	0.35%
脑抽	PHG+(生物部件+变空间位置)	0.03%
弃楼	PQG+(变领属+机构)	0.01%
弃楼	PQLH+(变领属+机构)	0.01%
前路	JHG+(变空间位置+建筑物)	0.13%
舍友	JHG+(关系值+人)	0.67%
微博	MHG+(特性值+量度值)	0.03%

微博	JHLH+(特性值+变莠态)	0.02%
微信	BEG+(特性值+特性值)	0.57%
雾霾	JHG+(液+现象)	0.03%
心塞	JHG+(无生物部件+无生物部件)	0.20%
艳遇	JHG+(特性值+时间)	0.07%
扎口	PQG+(变空间位置+生物部件)	0.56%
装 X	PQG+(变空间位置+信息)	0.06%
自虐	PHG+({firstPerson}+特性值)	0.01%
作死	BEG+(做+做)	0.10%
自处	…	0

本文的实验语料来自论坛，所以这些分词有误的普通未登录词大多为网络语言。这些词有些为原创，难以寻找构词理据，如“心塞”；有些为谐音，难以还原词素意义，如“妹纸”；有些为借用，往往产生了引申义或比喻义，如“扎口”等。因此，尽管大多数词都含有可辅助计算机自动识别的词素意义，但计算机还是很难准确地推测出它们的词义。

鉴于上述实验的局限性，作为补充，本文又在《现汉》（第六版）新出现的 2400 多个二字词中选取了新的实验对象。本文构建的二字词语义描写体系只对 8984 个二字词中出现过的词素所构成的新词有应用价值。经过筛选，我们共得到 1419 个有效新词，删掉 6 个同形词，最终确定了 1413 个实验对象。经过实验，我们发现有 1367 个新词含有至少出现一次的义类组合形式，约占新词总数的 96.74%。基于五个词素意义参照集，这 1367 个二字组的成词可能性如表 11：

表 11：1367 个二字组的成词可能性分布表

参照集	个数	百分比 ‡‡	实例
1	671	49.09%	座驾、坐台、足坛、总监、主唱、震区、掌门、约谈、选秀、小资
2	308	22.53%	走光、纸媒、艳遇、血拼、型男、试婚、圈钱、碰瓷、脑残、卖家
3	201	14.70%	醉驾、宅男、意淫、歇笔、限行、戏骨、微博、路痴、机洗、汉化
4	157	11.49%	撞衫、死磕、烧钱、猛料、裸婚、官瘾、茶歇、壁葬、北漂、爆粗
5	30	2.19%	助产、医患、星途、庭辩、岁尾、司勤、蜡疗、闺密、懂眼、菜鸟

我们选取了词素意义排名最高的“特性值+人”作为考察对象，由释义模式分布表可知，“特性值+人”最常和“JHG”连用，其次为“JHZ、JHJ、JHGT”等。在 1413 个新词中，共有 71 个词的义类组合中有“特性值+人”这一类，由于此类排名最高，所以计算机自动将

‡‡百分比是指二字组个数在 1367 个总数中的百分比。

“JHG+(特性值+人)” 认定为这些词最有可能的释义模式。依据“JHG+(特性值+人)”进行推测，词义应为“具有某种特性的人”。参照《现汉》（第六版）的释义，我们可以发现共有31个词，如“坐台、主厨、杂役、淫妇、新兵”等可以表示这种词义，其余40个词如“坐驾、坐大、重器、中号”等均不含这种词义，理解正确率为 43.67%。由此可见，基于《现汉》（第六版）1413个二字新词的实验效果更好，本文的研究成果对较为规范的普通未登录词的应用价值更大。

6. 结语

通过面向自然语料的实验，我们发现在规模较小的语料中，普通未登录词对分词精度的影响非常之大。现有的基于词表的分词方法、基于统计的分词方法以及基于隐马尔科夫模型的分词方法对普通未登录词的识别都有点儿束手无策，而二字词的语义描写体系能够有效地辅助识别普通未登录词。现有问题是究竟频率多大的词素意义可以被基本认定为词，还有待进一步验证。通过进一步的对比实验，我们还发现，二字词的语义描写体系对较为规范的二字词的理解效果更好。从实验结果来看，“从语义上寻找词素和词素组合成词的规律，进而指导普通未登录词的识别和理解”这一思路对中文自动分词存在着较高的应用价值，对这一专题深入研究，看似是一条提高自动分词精度的可行之路。

参考文献：

[1] 李行健.汉语构词法研究中的一个问题—关于“养病”“救火”“打抱不平”等词语的结构[J].语文研究,1982,(2):61-68.

[2] 符淮青.现代汉语词汇[M].北京:北京大学出版社,1985.

[3] 王树斋.汉语复合词词素义和词义的关系[J].汉语学习,1993,(3):17-22.

[4] 苑春法,黄昌宁.基于语素数据库的汉语语素及构词研究[J].世界汉语教学,1998,(2):7-12.

[5] 朱彦.复合词的语义结构与词素义的提示机制[D].广西师范大学硕士学位论文,2000.

[6] 冯海霞,张志毅.《现代汉语词典》释义体系的创建与完善[J].中国语文,2006,(5):455-480.

[7] 鲁川,王玉菊.汉语信息语法学[M].济南:山东教育出版社,2008.

[8] 中国社会科学院语言所词典编辑室.现代汉语词典（第六版）[Z].北京：商务印书馆,2012.

作者简介：作者一吉志薇（1988——），女，硕士研究生，主要研究领域为计算语言学、词汇语义学。Email:sichenfeimengli@163.com；作者二冯敏萱（1978——），女，副教授，中文信息处理、平行语料库建设。Email: fengminxuan@njnu.edu.cn。通讯作者：吉志薇，联系方式：13040086073。

 吉志薇

 冯敏萱