现代汉语虚词用法知识库简介

昝红英[1]　张坤丽[1]　朱学锋[2]　俞士汶[2]

[1]中国郑州大学信息工程学院　iehyzan*qq.com

(*を@に替え送信)

[2]北京大学计算语言学研究所

虚词对于现代汉语文本的内容理解有着举足轻重的作用，但是同实词相比，面向语言信息处理的针对汉语虚词用法的研究相对薄弱。郑州大学信息工程学院与北京大学计算语言学研究所合作，积多年之努力，初步建成**现代汉语虚词用法知识库**（**C**hinese **F**unction word usage **K**nowledge **B**ase，简称 **CFKB**）。**CFKB** 包括**现代汉语虚词用法词典**、**现代汉语虚词用法规则库**以及**现代汉语虚词用法标注语料库，**是一个三位一体的语言知识库。汉语的每个虚词可能区分为若干义项，每个义项又可能区分为若干用法。表 1 为现代汉语虚词用法词典中虚词义项分布情况；表 2 为现代汉语虚词用法词典中虚词用法分布情况。累计起来，目前的现代汉语虚词用法词典中收录的虚词词语总数为 2401 个，共涉及 2982 个义项及 4337 个用法。

表 1 现代汉语虚词用法词典中虚词义项分布

义项 / 词类 / 词数	1	2	3	4	5	6-10	词语共计	语义共计
副词	1375	134	37	12	4	4	1566	1848
介词	97	31	6	4	1	2	141	213
连词	273	31	8	3	0	0	315	371
助词	32	5	5	2	0	11	45	71
语气词	30	11	6	4	3	4	58	131
方位词	224	39	8	4	0	1	276	348
总计							**2401**	**2982**

表 2 现代汉语虚词用法词典中虚词用法分布

用法 / 词类 / 词数	1	2	3	4	5	6	7	8	9	10	10以上	词语共计	用法共计
副词	1214	179	84	38	21	12	4	3	2	1	8	1566	2356
介词	66	30	23	7	4	5	7	0	1	1	2	141	331
连词	156	50	55	24	16	7	4	0	1	2	0	315	696
助词	30	4	3	1	0	1	2	0	0	0	1	45	144
语气词	30	7	7	4	2	0	1	4	0	0	2	58	169
方位词	164	34	11	24	19	6	6	4	6	1	1	276	641
总计												**2401**	**4337**

现代汉语虚词用法词典对每个虚词的每个用法赋以一个的用法编码（**ID**）。ID 具有唯一

性，编码规律为：

POS_全拼音[_tn] [_m] [x] [y]

其中，

POS：为虚词词性（d， f， c，p，u，y）

tn：n 为数字，标明序号，用于同音不同形的虚词的编码。

m：为数字，不同义项的编号。

x：（a，b，c，d…）为用法编号。

y：（a，b，c，d…）对用法 x 的进一步细化编号。

[ ]：表示内容可缺省。

例如，用法编码 u_de5_t2_1ba 表示助词（“u”）中全拼音为 de5 的第二个虚词 t2 （即“的”）的第一个义项（“1”）的第二个用法（“b”）的第一个细类（“a”）（同音助词“得”的编码为“t1” ）。

对虚词每一个用法编码 **ID**，**现代汉语虚词用法词典**均有一条记录描述它所对应的虚词用法的详细属性信息。属性描述字段包括“词语”、“词性”、“全拼音”、“释义”、“用法”、“例句”、“小类”、“句首”、“左搭配”、“左紧邻”、“右紧邻”、“右搭配”、“搭配”、“句末”等，其中“句首”、“左搭配”、“左紧邻”、“右紧邻”、“右搭配”、“搭配”、“句末”属于用法特征描述。现代汉语虚词用法词典中给出了用法搭配涉及的词语信息，现代汉语虚词用法规则库中给出了用法搭配涉及的词语及词性信息。例如，**现代汉语虚词用法词典**对副词“就”区分了 7 个语义，涉及 21 个用法。详见附录数据库文件“就.xls”。作为样例，这里只给出“就”的两个用法的部分信息描述。

ID：d_jiu4_1a
词语：就
词性：d
全拼音：jiu4
释义：表示很短时间以内即将发生。<b>
用法：~ + 动词
例句：我~去<b>|这~走<b>|你等会儿，他马上~回来<b>|足球联赛明天~开始<b>|我们很快~把工作做完了<z>|不到一支烟的工夫，大家~纷纷提出告辞<r>
小类：时间
句首：
左搭配：表示未来时间的词语或事件描述，如：“很快|一会|马上|立刻|不到|眼看”。
左紧邻：
右搭配：
右紧邻：
搭配：
句末：

ID：d_jiu4_1b
词语：就
词性：d
全拼音：jiu4
释义：表示很短时间以内即将发生。<b>
用法：~ + 形容词

例句：我这头痛病一会儿~好<b>|麦子眼看~熟了，赶紧准备收割吧<b>
小类：时间
句首：
左搭配：表示未来时间的词语或事件描述，如：“很快|一会|马上|立刻|不到|眼看”。
左紧邻：
右搭配：
右紧邻：
搭配：
句末：

注 1：~ 代表“就”
注 2：<b>代表来源于《现代汉语八百词》，<r>代表来源于《人民日报》语料库，<z>代表自选例句。
注 3：“右紧邻”等用法特征相关的词性信息见现代汉语虚词用法规则库。

基于**现代汉语虚词用法词典**，对每个虚词用法进行了类 BNF 范式的形式化描述，形成**现代汉语虚词用法规则库**，为汉语自动分析和生成提供虚词用法自动辨识的判据。目前规则中利用的用法特征包括句首（F）、左搭配（M）、左紧邻（L）、右紧邻（R）、右搭配（N）以及句末（E）。例如，关于副词“就”用法<d_jiu4_1a>的规则描述是：

<d_jiu4_1a>→[M]R
M→很快|一会|马上|立刻|不到
R→v|d*v|p*v

关于规则元语言的简单说明：尖括号“<>”中的“d_jiu4_1a”代表一个 ID，“→”表示“定义为”，“[ ]”表示可选，“|”表示多选一，“*”表示匹配任意词串；规则右部的汉字表示具体的词语，小写字母表示词性（词性编码请参考北京大学《人民日报》词语切分与词性标注语料库规范）。该规则表示如果句子中“就”字的左边“可选地”出现“很快”或“一会”或“马上”或“立刻”或“不到”，且右边出现紧邻 v 或 d*v 或 p*v，则该副词“就”的用法编码 ID 就是 d_jiu4_1a。

虚词的一个用法可用一条或多条规则描述，因此规则的数目大于用法的数目，已建立的虚词用法规则库对 4337 个用法共计 4696 条规则。目前现代汉语虚词用法规则库各个词类用法规则的具体数目为：

- 副词：共有 2456 条规则，涉及 2356 个用法；
- 介词：共有 385 条规则，涉及 331 个用法；
- 连词：共有 747 条规则，涉及 696 个用法；
- 助词：共有 165 条规则，涉及 144 个用法；
- 语气词：共有 182 条规则，涉及 169 个用法；
- 方位词：共有 761 条规则，涉及 641 个用法。

在约有 1600 万字的基本标注语料库（即 1998 年 1 月和 2000 年 1-6 月的《人民日报》语料）上我们基本完成了所有虚词用法的标注工作（助词“的”仅完成其中三个月的语料标注），即为每个虚词加上其正确的 ID（用尖括号标示），形成**现代汉语虚词用法标注语料库**，可以为每个虚词用法提供诸多实例，并在**现代汉语虚词用法标注语料库**上对各个虚词的用法分布进行了计量分析，得到了有关用法分布的统计数据。在目前的现代汉语虚词用法标注语料库中，共有副词 385,091 频次，介词 314,329 频次，连词 205,485 频次， 助词 267,773 频

次， 语气词 12,309 频次，方位词 132,385 频次。 标注的部分样例如下：

- 20000404-08-003-004/m 中 国 队 /nt **仍 然 /d<d_reng2ran2_1a>** **没 有 /d<d_mei2you3_1a>** 克服/v 心理/n 紧张/a **的/ud<u_de5_t2_1g>** 毛病/n ，/wd 开场/vi **仅/d<d_jin3_t1_1aa>** 两/m 分钟/qt **就/d<d_jiu4_1a>** **被/p<p_bei4_1a>** [中国/ns 台北队/nt]nt 攻/v 入/v 一/m 球/n 。/wj
- 20000602-12-006-003/m 6 月/t ，/wd Discovery/nx 探索/vn 频道/n **将/d<d_jiang1_1>** 播映/v 一/m 周/qt **的/ud<u_de5_t2_1bb>** 特别/a 专辑/n “/wyz **外/f<f_wai4_1g>** 星/n 人/n 入侵/vn 周/Ng ”/wyy ，/wd 优美/a **的/ud<u_de5_t2_1c>** 画面/n **与/c<c_yu3>** 镜头/n 融/vi 科学/n 事实/n 为/vl 一体/n ，/wd 使/v 观众/n 体味/v 真正/b **的/ud<u_de5_t2_1c>** 太空/s **之/u<u_zhi1_1>** 旅/n 。/wj

现代汉语虚词用法知识库（**CFKB**）的研制是在北京大学**综合型语言知识库**（**CLKB**）的基础上进行的。可以说 **CFKB** 是 **CLKB** 的衍生成果。**CFKB** 也参考了《现代汉语八百词》、《现代汉语词典》（第 5 版）以及《现代汉语虚词词典》等文献资料。

CFKB 的研制得到中国国家自然科学基金项目“规则与统计相结合的现代汉语虚词用法自动识别研究”（项目号：60970083）、中国国家 973 课题“文本内容理解的数据基础”（项目号：2004CB318100）、北京大学计算语言学教育部重点实验室开放课题“现代汉语虚词知识库研究及大规模虚词用法标注语料库的构建”（课题号：KLCL-1004）以及中国河南省科技创新人才杰出青年基金项目“面向文本内容理解的现代汉语虚词知识库研究”（项目号：104100510026）的支持。在研制过程中也得到诸多师友、同行的指教和帮助，特别是刘云博士和彭爽博士，他们为 **CFKB** 的研制做了前期的准备工作。课题组向所有支持者、贡献者表示衷心的感谢。

CFKB 对现代汉语信息处理和文本内容理解研究新增了又一类语言数据资源，也为语言教学研究提供了具有周遍性、格式规范、实例丰富的语言知识素材。期望 **CFKB** 在现代汉语文本内容自动理解研究（实用系统包括句法分析、机器翻译、信息检索、信息抽取等）和对外汉语教学实践中发挥作用。并衷心希望用户反馈对 **CFKB** 的意见、建议以及在使用中所发现的错误、瑕疵。**CFKB** 课题组将持之以恒，努力把 **CFKB** 建设成实用型语言知识库。