现代汉语述补结构用法词典工作阶段总结
（2008 下半年-2009 上半年）

詹卫东 李超

1 概述

“现代汉语述补结构用法数据库”的编排方式是以述语为纲列出词典的条目，对每个述语（动词或形容词），按补语的不同语义类型来分项描写它能搭配的补语。包括结果补语、趋向补语、可能补语、程度补语、介词补语。述语与补语的一个搭配构成词典的一个词条，即数据库的一条记录，每条记录包含若干属性字段，以描述该词条相关的详细信息，主要包括：述语词和补语词的读音、义项、词性；述语词的来源、HSK等级（收录部分非HSK动词）、语义角色、所带不同语义类型补语的特点；述补结构整体的释义、备注和例句。

2 工作量完成情况总结

北京大学中文系十余位语言学相关专业的研究生近一年来对数据库的条目进行了增补和编校，目前已完成条目的基本情况如下：

述语1642	分义项2097		
		结果补语	7750
		趋向补语	6947
补语500	分义项583	可能补语	3174
		程度补语	1242
		介词补语	963
总条目			20076

词条的各主要属性字段均已填写完成，包括**注音、词性、释义、例句**（至少三句）、**频率、来源**等；同一述语的语义角色（论元成分）是它所带各类补语的共性特征。目前**语义角色**字段均已标注，并且每一角色后都有说明和举例。不同补语类型的共性特征概括尚未完成，有待继续补充完善。

3 近期词典修订工作总结

3.1 标注不一致的查找和修改

由于数据库的编校工作是多人同时进行，部分属性字段的处理存在不一致，如注音错误、词性标注不一致、补语类型标注不一致等。对于这些不一致的错误，都可利用数据库的统一查询功能发现并修正。以补语类型标注不一致为例，修改步骤为：

[1] 用以下 SQL 查询抽取有效记录的“补语-补语类型”对，得到 622 对，大大多于数据库的补语总数，说明其中有不一致的标注。

```
SELECT tuples.buyu, tuples .bytype, count (*) INTO complement
FROM tuples
GROUP BY tuples.buyu, tuples.bytype;
```

[2] 查看以上抽取结果，找出其中类型标注不一致的补语，如图所示：

	过来	结果	2
	过来	趋向	267
	过去	趋向	215
	过头	程度	6
	过头	结果	32

其中“过来”的补语类型是“趋向”，但有 2 处错标成了“结果”；“过去”则没有标错的。“过头”的补语类型是“结果”，但有 6 处错标成“程度”。

[3] 在编辑界面中用复杂查询功能定位到补语类型标注有误的条目，并进行修改。如在查询条件中输入以下条件：

即可查找到补语动词“过来”的补语类型误标为“结果”的 2 个词条，然后将逐一将其补语类型修改为“趋向”。

按照这样的方式修改了大约 1000 多个存在错误标注的词条，基本消除了补语类型标注的不一致，使得“补语-补语类型”对的数量减少到 507，与补语动词的数量基本一致。

以上三个步骤同样适用于修改补语拼音和补语词性的标注不一致，修改的效果如下：

修改前抽取“补语-补语注音”对数量为 798，根据其中出现不一致的补语找出并修改了 400 多个补语注音错误的词条，使得“补语-补语类型”对的数量减少到 509，与补语动词的数量基本一致，有少数多音字，如“倒”作补语的读音有时为 dao3，有时为 dao4。

修改前抽取“补语-补语词性”对数量为 776，根据其中出现不一致的补语找出并修改了 700 多个补语词性标注错误的词条，使得“补语-补语类型”对的数

量减少到 516，与补语动词的数量基本一致，有少数兼类词，如“到”有时是动词，有时是介词。

3.2 补充填写少数词条的某些不完整的属性值

少数词条存在属性值为空或例句不全的情况，发现后均加以补充完全。如：

a) 通过数据库查询发现 130 个词条的补语词性标记为空，如下图所示：

2928	调查	不看
263	吃	得消
264	吃	不消
980	够	得着
981	估计	得到
957	高	很多
958	高	许多
4481	挤	不着
680	调查	得下去
684	调查	得着
392	敌	得过
1554	酸	到
1011	规定	好
1018	过	得来
1145	混	进来
4824	低	得可怜

进而可在编辑界面利用复杂查询功能逐一定位到这些词条并补充补语词的词性标记。

b) 通过脚本程序对词条的例句进行统计，发现 267 个词条的例句不足三句，使用复杂查询功能定位到这些词条，补齐 3 句例句。

3.3 删除了部分误收的词条

在以上修改标注不一致或不完整的词条时，发现少数误收的词条，需要删除，如：

在修改补语词性标注不一致时发现“挖苦过”的补语词性为“助词”，这显然是将动态助词“过”误认为是补语，“挖苦过”其实并不是述补结构，应当删除。

在补充补语词性空白条目时发现“调查得下去”频率很低，而且已有相应的结果补语词条“调查下去”，所以不宜单独作为词条收录到词典数据库中，应当删除。

造成这类误收条目的原因比较复杂，无法统一查找处理，因此只能发现一个删除一个，目前删除的误收词条数量大约 50 多个。

4 对词典现有记录的统计分析

4.1 动词分级带补语情况的统计

数据库中收录了 HSK 动词表中能带补语的动词，包括甲级词 278 个，乙级词 478 个，丙级词 296 个，丁级词 177 个。另外还收录了不在 HSK 动词表中的动词 413 个。不同级别的动词所带补语数量与比例如下表所示：

动词类别	动词数量	分义项动词数量	百分比	构成述补条目数量	百分比
甲	278	453	16.9%	4897	24.4%

乙	478	627	29.1%	6725	33.5%
丙	296	361	18.0%	3086	15.4%
丁	177	209	10.8%	1678	8.4%
非 HSK	413	447	25.2%	3690	18.4%
	1642	2097		20076	

不同级别动词所带各个类型的补语数量与比例如下表所示：

动词类别	结果补语条数	百分比	趋向补语条数	百分比	可能补语条数	百分比	程度补语条数	百分比	介词补语条数	百分比
甲	2008	25.9%	1445	20.8%	876	27.6%	354	28.5%	214	22.2%
乙	2499	32.2%	2460	35.4%	1020	32.1%	414	33.3%	332	34.5%
丙	1133	14.6%	1119	16.1%	494	15.6%	194	15.6%	146	15.2%
丁	700	9.0%	610	8.8%	206	6.5%	88	7.1%	75	7.8%
非HSK	1410	18.2%	1313	18.9%	578	18.2%	192	15.5%	196	20.4%
总数	7750		6947		3174		1242		963	

不同级别动词所带补语数量的平均值如下表所示：

动词类别	平均补语数	平均结果补语数	平均趋向补语数	平均可能补语数	平均程度补语数	平均介词补语数
甲	17.62	7.22	5.20	3.15	1.27	0.77
乙	14.07	5.23	5.15	2.13	0.87	0.69
丙	10.43	3.83	3.78	1.67	0.66	0.49
丁	9.48	3.95	3.45	1.16	0.50	0.42
非 HSK	8.93	3.41	3.18	1.40	0.46	0.47

上述统计表明，在目前数据库中，HSK 表中级别越高的动词，所能带的补语平均数量越多。而非 HSK 动词所带补语的平均数量要少于 HSK 表中的动词。

4.2 关于例句句长的统计分析

述补用法词典数据库是以实例为基础来解释和辨析述补结构用法的。因此，所选例句应尽可能来自真实文本的使用，而且结构简单清晰，便于理解和翻译。通过编写程序统计，数据库中 20076 个述补词条下共有例句 61824 句，按字节统计的平均句长为 32.966 字节/句；对例句进行自动分词后，按词数统计的平均句长为

12.015。不同长度句子的频率分布如下表：

句长（词数）	频率
10	6791
11	6613
9	6526
12	5932
8	5523
13	5165
14	4258
7	3810
15	3507
16	2749
17	2119
6	1680
18	1630
19	1264
20	920
21	734
22	548
5	482
23	434
24	320
25	217
26	167
27	113
28	91
29	60
4	53
30	44
31	23
32	15
34	11
33	9
35	7
36	5
3	2
38	2

从平均句长和句长分布来看，数据库中词条的例句多数是长度适中、结构简单的句子，基本适合用来对述补结构的用法进行示例。不过可以结合下面 4.3 所描述的减少例句用字的工作，进一步进行简化。

4.3 关于例句用字的统计分析

通过程序对全部例句进行字频统计，例句共计 919,882 字（不含标点符号和非汉字字符），使用了 4,041 个不同的汉字。其中有许多不常用的汉字，需要进

一步调整例句，对非常用字进行筛选排除，改用相对常用、简单的汉字来组成例句。

具体修改方案如下：

（1）确定一个目前例句中怀疑不当用字列表（记作 BadHZList）。产生 BadHZList 的方法如下：

扫描目前例句中所用的 4041 字总表，对一个汉字 Hi，满足以下两个条件中的任何一个，都加入到 BadHZlist 中。

（a） Hi 是 GB 码二级汉字，即 Hi 的内码第一个字节在 D8H～F7H 之间。（目前例句中这样的字有 554 个）

（b） Hi 不在 CCL 语料库字频统计表中排位前 2000 的汉字以内。

（c） Hi 不在国家语委 2500 常用字表内。（例句中符合 b，c 两个条件的字有 1439 个）

将上述 554 字和 1439 字求并集，得到 BdaHZlist 共 1445 个汉字。

（2）确定一个例句可用字表（记作 GoodHZlist）。产生 GoodHZlist 的方法如下：

将国家语委 2500 常用字和 CCL 前 2000 高频字求交集，然后再跟现有例句用字经过上面第 1 步筛选后剩下的全部字求并集。共得到 2596 字。

（3）将 BadHZlist 中的字分配给课题组各位同学，在 web 页面上用高级查找方式，定位到包含这些字的记录，修改相应的例句。修改后的例句中如果用字不在 GoodHZlist 中，则在编辑过程中由程序实时报告提示。填写者遇到提示时，需要人工判断例句是否合适。如果程序提示的字填写者主观感觉并不罕用，可以选择提交，否则应返回重写例句。

修改例句时应注意：

（1） 例句用字尽量在 GoodHZlist 中。如果不在该集合中，也应避免采用字义复杂、生僻，笔画多，不易读的汉字。

（2） 低频字删除后如整句仍成立，则删除之。

（3） 低频字在专有名词中，可将专名替换成非专名，如人名替换成人称代词（嬴政 → 他），地名替换成指示代词（吐鲁番 → 那里/那座城市），物名替换成所属类别（黄莺 → 那种鸟）等等。

（4） 如果低频字是在一个条目的述语或补语当中，则判断该记录是否需要删除。如不易确定，则记录下来，提交课题组大家讨论。

因为例句未经分词处理，所以上面仅仅是针对汉字的频率高低来考虑例句是否合适，但例句中实际上还包含字频不低，但相应词频很低的情况。比如“擦白”这个条目的例句：

银饰在多次变黑之后就很难擦白了。

她把自己的脸擦白了。

战士们一遍一遍擦拭武器，有的把烤蓝都擦白了。

上面第三个例句中“擦拭”中的“拭”是低频字，根据上面所说修改例句的要求，应该去掉，而后半句中的“烤蓝”，虽然两个字都不是低频字，但“烤蓝”是一个非常低频的词，这种情况也不应该出现在例句中。一般来说，人名（含外国人名）、地名、专业领域名词，等等，其中的用字往往不是低频字，但整个词在日常用语中都是低频的。这些情况都需要做简化处理。

关于述补词典修改例句的工作安排（from 詹卫东）0917
各位好，

前一段时间李超对述补词典的例句句长及用字进行了考察，为我们进一步简化例句提供了依据。（具体情况见附件中 word 文件的说明）。下一个阶段，请大家根据要求对现有例句进行修改。统计所得低频字 1445 个。每个人平均需要处理的字在 200 以内。具体到一个字及其所在例句是否需要修改，还需要人的主观判断。

具体任务请李超分配一下。注意跟各位同学协商一下时间，根据各自时间多少来分配具体的工作量。同时大家在工作过程中也可再相互协调。要求在 10 月 18 日前完成。

请大家在 http://ccl.pku.edu.cn/sunaokadict/ 网页下操作。该网页下的程序已经根据修改例句的需要做了修改。http://ccl.pku.edu.cn/vc 网页下程序未做修改。不要在这个目录下操作。

下面是词典例句用字情况以及修改例句时应注意问题的说明（即附件中 word 文件的 4.3 部分）4.3 关于例句用字的统计分析
通过程序对全部例句进行字频统计，例句共计 919,882 字（不含标点符号和非汉字字符），使用了 4,041 个不同的汉字。其中有许多不常用的汉字，需要进一步调整例句，对非常用字进行筛选排除，改用相对常用、简单的汉字来组成例句。具体修改方案如下：
（1）确定一个目前例句中怀疑不当用字列表（记作 BadHZList）。产生 BadHZList 的方法如下：
扫描目前例句中所用的 4041 字总表，对一个汉字 Hi，满足以下两个条件中的任何一个，都加入到 BadHZlist 中。
（a） Hi 是 GB 码二级汉字，即 Hi 的内码第一个字节在 D8H～F7H 之间。（目前例句中这样的字有 554 个）
（b） Hi 不在 CCL 语料库字频统计表中排位前 2000 的汉字以内。
（c） Hi 不在国家语委 2500 常用字表内。（例句中符合 b，c 两个条件的字有 1439 个）
将上述 554 字和 1439 字求并集，得到 BdaHZlist 共 1445 个汉字。
（2）确定一个例句可用字表（记作 GoodHZlist）。产生 GoodHZlist 的方法如下：将国家语委 2500 常用字和 CCL 前 2000 高频字求交集，然后再跟现有例句用字经过上面第 1 步筛选后剩下的全部字求并集。共得到 2596 字。
（3）将 BadHZlist 中的字分配给课题组各位同学，在 web 页面上用高级查找方式，定位到包含这些字的记录，修改相应的例句。修改后的例句中如果用字不在 GoodHZlist 中，则在编辑过程中由程序实时报告提示。填写者遇到提示时，需要人工判断例句是否合适。如果程序提示的字填写者主观感觉并不罕用，可以选择提交，否则应返回重写例句。
修改例句时应注意：
（1） 例句用字尽量在 GoodHZlist 中。如果不在该集合中，也应避免采用字义复杂、生僻，笔画多，不易读的汉字。

（2）　　低频字删除后如整句仍成立，则删除之。
（3）　　低频字在专有名词中，可将专名替换成非专名，如人名替换成人称代词（嬴政 à 他），地名替换成指示代词（吐鲁番 à 那里/那座城市），物名替换成所属类别（黄莺 à 那种鸟）等等。
（4）　　如果低频字是在一个条目的述语或补语当中，则判断该记录是否需要删除。如不易确定，则记录下来，提交课题组大家讨论。

因为例句未经分词处理，所以上面仅仅是针对汉字的频率高低来考虑例句是否合适，但例句中实际上还包含字频不低，但相应词频很低的情况。比如“擦白”这个条目的例句：
银饰在多次变黑之后就很难擦白了。
她把自己的脸擦白了。
战士们一遍一遍擦拭武器，有的把烤蓝都擦白了。

上面第三个例句中“擦拭”中的“拭”是低频字，根据上面所说修改例句的要求，应该去掉，而后半句中的“烤蓝”，虽然两个字都不是低频字，但“烤蓝”是一个非常低频的词，这种情况也不应该出现在例句中。一般来说，人名（含外国人名）、地名、专业领域名词，等等，其中的用字往往不是低频字，但整个词在日常用语中都是低频的。这些情况都需要做简化处理。

詹卫东
2009-09-17

各位好，090918

修改例句过程中也涉及到同时检查其他相关信息填写信息是否需要修改。当然这次工作主要以修改例句为主。

现有词条一般情况下都保留。除非该词条不符合我们之前的收录原则。

比如这次李超在查例句用字情况过程中，碰到一个述语条目是“耪”，这是一个方言用词，不在 HSK 词表中。而且明显是一个低频词，像这样的条目，应该删除。

关于修改例句工作中应遵循的原则，我在上封邮件中已经说明过了。这里再补充说明两点：

1

像“耪”这样的低频字，因为初始的述语条目中收录了，本来在编辑过程中，这个属于应当删除的，但没有删除，而且还填了 10 多个补语。在此次简化例句的工作中，应该把“耪”这样的条目删除。

现在抽取出来的“例句中不当用字表（BadHanziList)”，只是一个参考，并不意味着其中所有的字都要从例句中删除。这些字之所以被筛选出来，是依据了一定的频度统计标准，但是，频度统计永远都只是一个参考，而不是理性的“规则”。（频度统计对语料的依赖性太强，而“平衡语料”在相当程度上还只是一个美丽的传说……）
比如 BadHanzilist 表中第一个是“熬”字，按我们一般的感觉，这个字并不是很少见（各位大概经常“熬夜”吧)。但是，因为在 CCL 语料库（3 亿字规模）中统计“熬”的频度排序很低，因而被认为是一个“低频字”。而在现在的述补词典中，“熬”带补语的词条有 56 个。而且“熬”是 HSK 动词词表中的乙级词，像这种情况，就不应该删除。

大家在工作过程中，应从多个角度去衡量一个条目的合理性，一个例句选择的合理性。目前通过电脑程序给出的“当用字表”也好，“不当用字表”也好，都只是一个参考。如果完全没有参考，就会过于主观，这样填写的例句很可能不好。但如果只依赖电脑给出的参考，而不用我们自己的人脑，那就是本末倒置了。请各位务必注意拿捏。

词典的质量是在细节中体现的。希望各位付出的劳动，将来能够行之久远。

詹卫东
2009/9/18