语文词典标注词性的基本原则

樊立三[1] 亢世勇[2] 王兴隆[2] 马永腾[2]

（1. 山东工商学院, 山东烟台;2. 鲁东大学 2, 山东烟台 264025）

摘　要：从目前标注词性的现代汉语语文辞书中选取了较有代表性的五部语文词典，我们建成了《汉语语文词典词性标注数据库》，基于该真实语料库我们对这五部词典的词性标注差异进行了详细地考察和比较。针对标注现状，本文指出了具备一定客观性和真实性的词性标注的几项基本原则，以及词典词性标注还要保持动态的开放性。

关键词：汉语词典；词性标注；语料库

The principle of labeling the part-of-speech on the Chinese dictionary

Fan lisan[1]　Kang shiyong[2] Wang xinglong[2] Ma yongteng[2]

（Shandong Institute of Business and Technology1,Ludong university2,The city of Yantai in shandong 264025）

Abstract: Selecting from the present five representative modern Chinese dictionaries which labeled the part-of-speech , we completed *Database about the part-of-speech of Modern Chinese language*, we carried on the inspection and comparison which based on corpus for language materials to the five dictionaries' difference of the lexical category. In view of the present situation of the marking, this article has pointed out several basic principles which are more objective and authentic about the marking of the Lexical category: and the dictionary of the lexical category also should maintain the dynamic openness.

Key words: Chinese dictionary; lexical category tagging ; corpus

引言

汉语语文辞书的词性标注问题一直为大家所关注，也一直是汉语语文辞书编纂未能很好解决的难题。近年来，多部词典作了标注词性的尝试，使这一问题的解决在理论和实践两方面都有了相当大的进展。目前需要继续探索，并及时总结经验，把这一方面的研究引向深入。我们拟选取几本有代表性的词典，在此基础上进行词性标注的比较，总结出了语文辞书词性标注应当遵循的基本原则。

1《汉语语文词典词性标注数据库》（简称《数据库》）的实现

1.1《数据库》介绍

要真正掌握一个语言系统中的词汇，不仅要理解其词汇意义，而且要了解其用法。而要了解其用法，就不能不了解其语法方面的表现与特性。词性是一个词在与其他词组合时所显示出来的语法性质，词典提供尽可能详细的语法信息，可以更好地帮助读者使用。正是鉴于这样的思考，我们以和清华大学合作完成的包含 11 万词语

作者简介：樊立三（1979—），男，山东临沂人，文学硕士，山东工商学院学生处. E-mail:fls0043@163.com

的《现代汉语电子词典》（简称《电子》）（孙茂松、亢世勇主编）为基础，选取了目前标注词性的现代汉语语文辞书中较有代表性的《现代汉语规范词典》（简称《规范》）（李行健主编，外语教学与研究出版社，2004）、《应用汉语词典》（简称《应用》）（郭良夫主编，商务印书馆，2000）、《新世纪汉英大词典》（简称《新世纪》）（惠宇，外语教学与研究出版社，2004）、《多功能学生语文词典》（简称《多功能》）（肖懋燕、陈杰编著，上海辞书出版社，2001）等四本词典，录入了四本词典所收字、词、语全部条目和所标注的词性，共有28337条，从而建成五本词典的词性对比词表，这就是我们考察的封闭域。

1.2 五本词典的词类体系对照表、词性标注符号及标注规范（以《电子》的词类体系为序）

电子	应用	新世纪	多功能	规范
动词 v	动词	动词	动词	动词（助动词和趋向动词）
形容词（不包括区别词、状态词）a	形容词（形容词或形容词性短语）	形容词(包括形容词或形容词性短语)	形容词	形容词
名词(不包括人名、地名、机构名、）n	名词（名词或名词性短语）	名词(包括名词或名词性短语)	名词	名词(包括处所词、时间词、方位词)
人名 nr				
地名 ns				
时间词 t				
副词 d	副词	副词	副词	副词
区别词 b				
状态词 z				
叹词 e	叹词	叹词	叹词	叹词
连词 c	连词	连词	连词	连词
介词 p	介词	介词	介词	介词
语气词 y				
拟声词 o	象声词	拟声词	象声词	拟声词
机构名 nt				
其他专名 nz				
方位词 f				
处所词 s				
数词 m	数词	数词	数词	数词
量词 q	量词	量词	量词	量词
时间量词 qt				
数量词 mq				
代词 r	代词（包括指示词）	代词	代词	代词
助词 u	助动词	助词	助词	助词（包括语气词）
	前缀	前缀		前缀
	后缀	后缀		后缀
成语 i	成语	1（包括成语、谚语、俗语、惯用语、歇后语等在内的熟语）	成语	2（收录成语、惯用语、和其他不标注词性的词组或固定语）
习用语（包括熟语和自由短语）l	1（俗语、惯用语、谚语、歇后语、熟语）			

通过上表，我们不难发现，各家词典的词性标注体系存在的分歧主要表现在：（1）名词要不要再细分出时间词、方位词等下位词类？（2）形容词要不要再细分出状态词、区别词等词类？（3）熟语、惯用语、成语等语类的分合以及划界应该如何处理？（4）要不要标注前缀、后缀，是否另立一类？（5）助词是否要细分，以及语气词是否也单立出来？这些问题，都需要我们在理论上达成一致的共识，才能得到更好的解决。

词性差异标注规范是：如“(ga\gn)\(a\n)\(a\n)\(a\n)\(a\n)”表示：同一个词条，《电子》的词性标注为(ga\gn)，《应用》标注为(a\n)，《新世纪》标注为(a\n)，《多功能》标注为(a\n)，《规范》标注为(a\n)；另外，在对比标注过程中，为了对比考察的方便，根据《电子》所收词条，其他词典中收录了却未标注词性的任何词语（包括字、词、短语）都一律标注为“2”，其他词典中根本未收入该词的项标注为“3”，如“in\i\I\3\2：舌剑唇枪”。

2 词性标注的基本原则

2.1 语言单位的划定要明确

（一）词与语素

根据《数据库》，我们可以看出无论是词性标注相同还是差异，词与语素的划界都是较为棘手的，也是现今没有能够很好解决的问题。根据统计，《数据库》中所涉及语素方面的标注词条约有 270 个标注类型，共 4284 个词。如：

ga\(a\v)\a\2\a　惺、恢　　gn\(n\q)\n\(n\q)\n　腔、匝

有些词汇在古汉语中可单用，现代汉语中却不再单用，而多用在某些固定词组中，解决好这样的问题并不容易，“主要原因是汉语中词和语素的界限本来就是模糊的，要想来个泾渭分明，是不可能的。问题的复杂性还远不止这一点，许多字往往在这种场合作词，而在那种场合作语素，其中的规律性也不太明显。”（莫彭玲，2000）。如：

民：在很多情况下可以单用，“爱民如子、民以食为天”，我们多归为“词”而非“语素”。

再者，就是有些词汇有时单用，如“点了一炷香”，但有时候却只有和别的词组合后才能表示完整的意义，如“麝香、檀香”，对于这类情况，我们多归为词而不是语素。如：

(a\v)\(a\n)\(a\d\n\v)\(a\n)\(a\n)　香

有些单字在某些组合中单用，而在其他组合中却不能单用；如：(qt\gd\gn)\(d\n)\(d\a\n)\(d\n)\(a\d\n) 时

小时候不懂事，天天嚷着父母买电视。

青春年少时，总是羡慕那些成熟的人；年近不惑了，又对韶颜稚齿者钦羡不已。

处理方法可以是，在“时”条中，可以标词性，但要有一个用法说明“用于书面语，前面不能加‘的’”。造成这种状况的主要原因是现代汉语本身是一个“不同质”（郭锐，2002）的系统，包含一定量的文言成分，这些成分在特定的上下文中可以单用。我们认为这种情况下可以看作词，标出词性，但要指明是用于现代汉语书面语，并用真正的现代汉语句子作例句。

（二）词与短语

短语是词和词的语法组合，一般来说，短语通常包括大于单词而小于整句的各类形式，其中既有搭配固定的惯用语、成语和谚语，也有搭配不那么固定的词组。

1v\3\n\3\n　绿色革命、保护关税　　1v\v\n\3\2　信息爆炸、宏观调控

通过上述举例，我们可以发现有些词语尤其四音节的，在究竟是标注词性还是确定为固定词组上存在很大差异，误标或者不统一的情况很多，像“信息爆炸(2)”、“绿色革命（n）”，或统一标注为“n”，或即使分辨不清也应该都不予以标注。

因此，要在现代汉语词典中标注词性，应当弄清一个个作为条目的词语究竟充作语言的什么单位的代表。如何区别单独成词的语素和不单独成词的语素，决定什么样的语素组合只是词，而什么样的语素组合构成一个短语。只有弄清这些问题，把现代汉语中真正的单音词一一遴选出来作为条目，才能在语文辞书中准确地标注词性。

2.2 词类体系的采用和词性的鉴别标准要有操作性

词性标注必须建立在对词汇科学分类的基础之上，然而词类问题一直是汉语研究中的老大难问题，在到底该分哪些类、具体词的词类归属等问题上仍存在很大争议。本文所采用的《数据库》中收录的五本词典词性标注体系，存在些许差异，如：

A. 区别词、状态词以及语气词是否分立出来而单独另立词类的问题

b\a\a\3\3　　无线、精装　　z\a\3\3\a　　满当当、黑黢黢

B. 词性标注体系的设立恰当与否

在《新世纪》词性标注体系中，把熟语（包括成语、谚语、俗语、惯用语、歇后语等）作为一个条目标注；《现汉》则采取对成语、熟语、一般词组等不予以标注的原则；而其他四本词典则对熟语内部进行了分类标注，如《规范》分立出成语、惯用语和常用固定语但未标注词性，《电子》则分立出了成语、习用语并予以细化，《应用》则把成语、俗语、惯用语、谚语、歇后语、熟语分别单独列出，《多功能》则仅分立出了成语并予以词性标注。

ln\I\l\3\2　　妻儿老小、郎才女貌、左道旁门、獐头鼠目、过街老鼠

目前，语文词典标注词性基本上是沿用暂拟系统的 12 个词类，在这个词类体系中，很多词类之间没有明显的区别性特征，“暂拟系统划分词类的标准是词汇语法范畴，缺乏明确的、可观察的特征，操作起来有一定的困难，标注词性时容易出现漏洞”（唐健雄，2002）。

划分词类的依据，“最佳的出路是依据词的语法功能为划分的依据”（陆俭明，2005：34）。而怎样依据词的语法功能给词归类，陆俭明（2005：35）、邢福义（2003：5）等都对划分词类提出过可行的语法功能标准。

而我们划分词类，也是根据词占据语法位置的能力，即词的语法功能。所谓词的语法功能主要是指：（1）在句法结构中充当句法成分的能力；（2）和某个或某类词语组合的能力。例如，考察具有如下语法功能的词：

(1)从充当句法成分能力看。

a. 可以用作主谓结构中的谓语，但不能带真宾语。如“人很多”中的“多”是谓语。但象“多一点”中的数量短语“一点”是准宾语，而不是真宾语。

b. 可以作述补结构中的补语，如“洗干净、捆得结实”中的“干净、结实”是补语。

c. 直接或加“地”后作状中结构中的状语，如“安全地转移”中的“安全”是状语。

d. 直接或加“的”后作定中结构的定语，如“挺拔的山峰”中的“挺拔”是定语。

(2)从和某个或某类词搭配组合的能力看。

a. 可以受“很”类程度副词修饰，如“很高、特别雄伟”，但在受“很”修饰的同时，不能再带宾语，即不符合“很+形+宾”（符合此条件的词语，我们处理为动形兼类）。

b. 可以用“a +不+a”的形式提问，如“硬不硬、痛苦不痛苦”。（表心理活动的动词和部分判断动词也符合这种条件，我们处理为动形兼类）

上列所举(1)、(2)所体现的语法形式特征，我们把具有这类特征的词语归为形容词。我们认为从充当句法成分的能力与从和某个或某类词搭配组合的能力两个角度来分析语法特征，并予以可见的形式，这样也许可操作性大一些。

词典标注词性，大量的工作主要是进行词的归类，即从具体的词出发，一个个地分析、判定它们的词类归属。词典不是仅就少量的典型用例进行分析，而是要对整部词典中所收录的几万个词作穷尽处理，因为词类问题不仅有一个定性的问题，还有一个定量的问题，其中包括正处于语法特性的渐变过程中而不容易判断其所属词类的词。因而，确定词性的鉴别标准便尤其重要。在为词典确定鉴别词性的标准时，应当对词的语法功能作全面地考察，在分析研究的基础上提出比较明确、完备的可操作的形式特征，使之能够有效地说明语言现象。

2.3 历时层面和共时层面的标准要明确

语言是发展的，现代词汇与古代词汇有很大的不同。“古今汉语一脉相承，编纂辞书必须考虑古今汉语的关系问题”（程荣，1999）。因此，编写现代汉语辞书要经常面对和研究某些古代汉语的问题，在描写共时层面汉

语现象的同时也要从不同角度说明与之密切相关的汉语现象，如：

安　(v\ga\gnr)\(a\n\r\v\d)\(r\d\q\a\v)\(a\r\v)\(r\v\a\q)

上例中“安”是否要增立古汉语义“表示疑问，询问处所，相当于哪里：沛公安在？”

其次，文白层面上是否成词的把握问题，如：

(qt\gd\gn)\(d\n)\(d\a\n)\(d\n)\(a\d\n)　时

时：①季节、时令：四时八节②时间、岁月：历时十载③指某一段时间：此一时，彼一时④指规定的时间：过时不候⑤时辰：子时⑥时机、机会：时来运转⑦有时候，时而：⑧当时，这时：时大风雪，旌旗烈⑨时常、常常

再者，“任何语言共时平面上的词，都实际存在着两种不同的历史层次和领域层次”（陆俭明，2005：1）。比如口语、方言义增立与否的问题：

(d\a)\(a\v\d)\(a\d\n)\(d\v\a\a 古)\(n\v\a)　　好

(d\h\v)\(v\n\d\u)\(d\v\n)\(v\d\n\v 古)\(n\d\u)　过

上例中的“好”<名>在《规范》词典中释义为“指赞扬的话或问候的话（口语中多儿化）”，如“叫好儿”；“过”<量>在《规范》词典中释义为“用于动作的次数”，如“衣服已经洗了三过了”。

专业文献词汇一般不单用，多在列举时使用，但也可在科学术语组合元素（如：一氧化氢）中使用。这包括155个化学名称用字，320个动植物名称用字，如：gn\n\3\3\2（2）　铅、邛

还有天干地支、星宿名、卦名等都易处于中间状态，使用范围比较狭窄，一般处理为词。

在给词条确定义项诠释词义时，或者标注词性时，要兼顾到历时层面上古今的变化，处理好古今词义关系，也要注意到共时层面上领域的区别，比如《多功能》在标示词性时就明确标注了<方>标志。

2.4　词典的释义要与词性标注相协调

意义和语法功能是蕴含在词内的两个要素，二者之间存在着相当程度的依存关系，“有什么样的词义就会反映出什么样的词性；反之，有什么样的词性，也就会有什么样的词义。词性不同，释义的方法也就不同”（陈瑞国，1994）。《现汉》（2002年修订本）虽然没有明确标注出词性，但绝大多数词可以通过解释词义来体现词性，就是这个道理。所以在标注词性时，要兼顾和参考词义。

李行健先生在《规范》“前言”中讲到：按词的义项分别标注词性。“按义项给这种特定的使用单位标注词性，则可排除汉语中兼类词的干扰,使词性标注成为可能”（李建国，2004）。我们认为，这种办法是可行的，可以作为语文词典词性标注的一个原则，从表面上看，《规范》是根据意义标准来标注词性的，但细究这种词性标注方法的本质也是主要依据句法功能，因为词要在句中充当成分，必须具备相应的词汇意义。词汇意义是句法功能的物质承担者。一个词有几个意义，就可能兼几个类。就是不兼几个类，其不同意义所展示的语法功能也可能有所差异。这就比较真实地、全面地、科学地展示了每个义项的语法功能，这在词典编写上是一大进步，对词类研究乃至语法研究都有着积极的意义。比如：

a\(a\v\n)\(a\v\n)\(a\n)\(a\n)　　骄傲

“骄傲”在《规范》中有三个义项：①<形>自负；看不起人②<形>自豪：为中国女排的成就而感到骄傲③<名>指值得自豪的人或事物。同样在《应用》中也有三个义项：①<形>自高自大，看不起人②<动>自豪：我们为祖国的壮丽河山而感到骄傲③<名>值得自豪的事物和人。通过上述两本词典对“骄傲”一词释义的比较，第二个义项释义基本一致，其在句中的语法功能也相同，但词性的标注却存在差别。

因此，按照义项标注词性的尝试从总体上看是成功的，但是若单纯采用这一标准又将产生新问题，比如：难以兼顾词的同一性、义项的归并往往带有主观性、词性的特征就显得零乱等问题。总之，不论义项的多少、词性的缺立，只要确立了一个义项就要相应的标注词性，即词典的释义应与被释词在词性上相协调。

3 结语

基于大规模的语料库对词性标注进行细致地考察和分析，这不仅可以充分地说明词典不同，对同一个词的词性标注存在很大差异的现象，更重要的是可以为现代汉语语文辞书的词性标注提供充足的语料佐证，使我们的研

究具备一定的客观性和真实性。我们认为，除了要遵循上述基本原则外，还应该保证词性标注的开放性。词汇是不断发展变化的，词义也会随之消失或增添新义，因此也就要求词性标注是动态的、不断扩充的，是具有充足性和全面性的。

总之，词性标注的工作需要在实践中逐步开展和完善。通过一段时期的实践，在词性标注问题上会逐渐积累经验，在标示的体例上可以建立一套适用于各种类型的语文词典的通则。我们相信，只要努力，现代汉语语文辞书在语法信息的标注上会不断取得进步，不断走向成熟。

参考文献：

[1] 陈瑞国．词典标注词性浅谈[J]．理论学习月刊,1994(9).

[2] 程　荣．汉语辞书中词性标注引发的相关问题[J]．中国语文,1999(3).

[3] 郭　锐．现代汉语词类研究[M]．商务印书馆．2002.

[4] 李建国．但开风气不为先[J].语言文字应用，2004(3).

[5] 李行健主编．现代汉语规范词典[M]．外语教学与研究出版社．2004.

[6] 陆俭明．现代汉语语法研究教程．北京大学出版社．2005.

[7] 莫彭玲．字典标注词性的积极探索和实践—浅谈《现代汉语规范字典》的词性标注[J]．常州工业技术学院学报,2000(1).

[8] 唐健雄．语文词典标注词性及相关问题[J]．河北师范大学学报，2002(5).

[9] 邢福义．词类辨难[J]．商务印书馆．2003.