关于《现代汉语词典》词汇计量研究的思考

苏新春

Abstract *A Dictionary of Modern Chinese* is standardized dictionary of words in Chinese Putonghua. The establishment of a database of *DMC* for a quantitative study of its entries as a closed, exhaustive and specific body will change significantly the situation of lacking quantitative analysis in traditional lexical study, so that to help the systemization and accuracy of the study of Chinese lexicology and lexicography. This paper expounds the point of departure, ways and methodology and perspective in theory and use of this study. It also makes some suggestions on the quantitative study of Chinese words.

《现代汉语词典》(以下简称《现汉》)词汇计量研究是在数据库基础上对该词典的所有词目、字形、释义、注音进行专题、封闭、量化的统计性研究。下面就这一课题研究的开展谈几点认识。

一 开展《现汉》词汇计量研究的出发点

1.1 对传统词汇研究的变革

半个多世纪以来的现代汉语词汇研究,在词汇的诸多方面取得了长足的进步,但审视过去,也发现它在大观上存在若干不足,主要表现为:定性式的研究方法、非整体的词汇研究观及取材的非充足性。

所谓定性式研究方法,即研究主要凭借的是研究家对材料的主观感受与判断。研究中个人的识断起着主要作用,所依据的主要是典型性、富于个性的语料。这种以识断选例、从个案窥全局的特点,不可避免地会带来个别结论与普遍规律、个人见解与普遍材料之间的矛盾。在词汇学史上,各宥所见、见仁见智的现象屡屡可见,使得词汇研究长期处于"人治"阶段,难以走上科学化的道路。如对是否存在词汇系统,看法长期得不到统一,直至有人把词汇系统明确地,尽管不很清晰、不很完整地描述了出来,意见才趋向一致。[①]又如普通话词汇系统的来源与状况,尽管人们普遍认为有五大来源、六大来源,但每一种来源词汇与普通话词汇在进与退、量与质、渗透交融与沉淀同化等方面的关系如何,有着什么样的演化规律,有无富于操作性的量化标准,至今都还是不甚明了。在这样的基础上再来说普通话词汇的整体状况,当然会有偏颇。

又如词典,无论是专科术语类还是断代的语文类词典,都有着属于自己的核心词汇部分,也都存在着分布于边缘地带的过渡成分,可现状却是:对词目的收录、保留、删除都缺乏对过渡成分严格的区分标准,成为又一个典型的"定性"领域,处处可见"吾辈数人,定则定

矣”的痕迹。新词新语研究由于缺乏量的分析与说明,以致于旧词当新词者有之,偶用词作常用词者有之,误用词当定型词者有之。在对现代汉语词汇整体面貌还缺乏清晰、完整描述的情况下,所建立起来的词汇理论也就难免粗疏、缺漏。

1.2 频率研究在词汇研究中具有直观、可靠的作用,是词汇内在规律的体现

语言是一种符号系统,是人类使用的交际工具,它所具有的价值上的重要性,结构上的通用性,使用上的常见性等“质”的内涵都会在语言要素的“量”上反映出来。词汇的定量研究正是立足于语言的这种“质”与“量”的关系特征之上。词汇的量化研究主要体现为词的频率研究,如结构频率、分布频率、使用频率等。在语言的各种要素中,词汇是体现频率特征最突出,相对来说也是较易进行量化研究的;可惜长期以来人们却宥于词汇是一盘散沙的观点,在研究中一直缺乏系统、整体、量化的观点。近 10 多年来,在汉语史的研究中人们开始重视了专书专人的研究,但对断代词汇的计量研究,特别是基于断代词汇计量研究之上的词汇理论研究,却迟迟难以进入状态。其实,这项工作在理论上确立起了正确而明晰的认识,并与当代成熟的计算机数据库技术结合起来,它就变得自然而必然了。

1.3《现汉》在反映现代汉语词汇面貌上的代表性

在确定了词汇计量研究的认识后,选取有足够容量、富于代表性的、系统自足的断代词汇材料,就成为关键问题。词汇研究与词典有着天然的联系。词典是词汇材料的聚合体,它反映的是具有普遍的社会性、定型成熟、并经过人们整理的系统的词汇材料。对现代汉语词汇研究来说,《现汉》有着难以替代的特殊价值。这不仅仅是因为它是一部语文词典,收录的主要是人们日常使用的生活词语;也不仅仅因为它是中型词典,收录了 5 万多条词语,现代汉语的基本词、常用词都见于其中;更不仅仅由于它有着极广的流传面和极高的权威性。最重要的是因为:它是致力于以反映现代汉民族共同语词汇系统为己任的词典。

规范型词典全面反映语言的词汇体系,就要对词语作全面收录,不因某些词语无需查检而不收。……规范型词典如果把数以万计的常用词排除在外,它将是一部残缺不全的词典,也就谈不上为民族共同语规范化服务。而单纯以释疑解难为目的的词典,在收词上就不一定照顾到词汇系统的全面,一些很常用而不需索解的词可以不收。

规范型词典对民族共同语词汇的记录是全面的,但不是穷尽的(在理论上和实践上都是不可能的)。《现汉》是一部中型词典,它在收词上既是全面的,又有较强的选择性。选词的依据,主要不是看查考的需要,而是看词语在语言使用中出现的频率。[②]

以上是《现汉》编纂者对词典功能、选目依据的说明。正是这种说明成为最终影响本课题把《现汉》作为分析材料的最重要因素。

《现汉》的编纂者多是造诣精深,学有专攻的行家里手,他们从上百万张资料卡片中反复斟酌,层层筛选,最后确定收录的五万多条词语,无疑是对现代汉语词汇的一次全面整理和规范。[③]

这表明,《现汉》所收纳的词目很大程度上反映着现代汉语词汇的构成与概貌。正因为如此,后来以反映现代汉语词汇系统为己任的《同义词词林》[④]、《简明汉语义类词典》[⑤]等,都把《现汉》作为不可或缺的收录对象。

二 《现汉》词汇计量研究的思路与方法

《现汉》为现代汉语词汇研究提供了一份很有价值的材料。对词汇的来源与分布、词汇

成分与系统、词汇单位与结构、词义成分与色彩、词汇演化与词义诠释、常用词与非常用词、常用字与难僻字等等,对规范词典的选字与收词、立目与诠释、标音与词汇属性标注、释义内容与释义方法等等,可以说凡是与词汇和词典有关的理论与实践问题,都可以通过对这份语料的封闭、穷尽、定量的研究来作出有说服力的分析。

课题研究的基本作法是把《现汉》所有的内容都输入电脑,建立一个专题数据库。一个词语为一条记录,将词目、注音、释义、词频、结构、义类、词语来源、版本、页码等分别设立字段。字段的设立很灵活,可以根据不同的研究需要随时进行标注。为了方便对比,还将前后相隔 13 年的第二版与第三版同时输入,既可以透视词汇词义在历时状态的演变,也可以清楚地再现后版对前版的改进、修订,在辞典编纂学上提供非常有意义的对比材料。

《现汉》数据库内容丰富,计量研究以专题的形式进行。专题的计量研究有三个基本要求:

第一,语料的封闭与穷尽。进行专题研究时,对该专题范围内的语料要做到准确、封闭与穷尽。准确是必须真实地反映《现汉》的本来面貌,不能有讹误,把人为的差错带入语料中。封闭是使得专题研究做到纯化,不与无关的问题相搀杂。穷尽是保证语料不出现缺损、遗漏,使计量研究反映出来的频率、比例等数据真实可靠。这三点是计量研究的基础。当然,计量研究的本质是归纳研究,在使用有相当数量的语料时,个别数字的增减不会影响到语料的量与质,但作为严格的计量研究来说,数据的准确应该是计量研究的第一位要求。

第二,开阔观察视野,多方设立参照点,增加对比度。有比较才有鉴别,在对比中更能凸显语料的特点。对比的角度可以多样而灵活。例如在研究《现汉》同形词词目的设立时,就将同形词之间的意义差别与多义词义项之间的差异、单义词义项之间的差异、先为同形词后来为多义词,及先为多义词后为同形词等四种材料进行了对比,结果清晰显示出词典在同形词的设立中表现偏重词形差异、轻视词义关系、且贯彻不太一致的现象。⑥

第三,从理论上深入准确地阐释,揭示其内在特点与规律。专题研究的选定本身就是在一定理论认识下的产物,但专题语料经过封闭、穷尽的调查统计出来后,并不就等于解决了问题。选择观察语料的角度,确定分析语料的理论和方法,明确分析语料的目的,乃是词汇计量研究中至关重要的问题。否则,一堆语料放在面前将毫无生气。材料并不具有自动显示语言规律的作用,只有在理论的观照下语料才能将它的内在价值显示出来。当然,没有理论指导和明确研究,也无从发现真正有价值的语料。之所以强调这一点,就是要克服以为计量研究只是材料统计的偏颇观点。计量研究只是一种手段,只是对语料的一种处理方法,重要的是通过大量、准确的计量分析来发现语言的本质属性与规律。例如,对同形词历来只把它当作词汇的书写形式来研究,在口语的研究中它还根本引不起人们的重视,因为口语中是无所谓同形不同形的,要讲的也只是同音词。到了书面语才有了文字表达形式的是否同形的问题。这个问题当然在词典编纂对词目的设立时首当其冲,但从词汇理论的高度来看,它却是关涉到词汇学中最重要的基本单位“词”的意义范围到底如何确定这一核心问题。对它的处理直接与“词”这一基本单位的确立、汉语单位的层级性、词汇系统的数量、口语中的词与书面语中的词是否一致、人们对词语的认知能力都联系在一起了。

以上三点缺一不可,互为前提。没有语料的准确,计量研究将失去基础;没有多角度的对比,难以深入到语料的内部世界;没有理论上的深入挖掘,将只是材料的堆砌,计量研究将

失去灵魂,语料的内在特点与规律将难以显现。

三 《现汉》词汇计量研究的理论与应用前景

3.1 对现代汉语词汇的整体情况作出全面、系统、量化的调查与说明

由于《现汉》是以努力反映现代汉语词汇系统为目的的,因此,将《现汉》的词汇来源、结构、义类、属性、词性、色彩等问题调查清楚,也就可以说对现代汉语词汇的整体面貌有了一个较为清晰的认识。现以《现汉》第二版(1983 年)的一些基本情况为例作些说明:

共收词目 56147 条,其中单字词目 10540 条(如再分出单音词、单音构词素,或表音汉字,则还有着另外层面上的意义),复音词目 45607。

共有义项 68344,两个以上义项的 9996 词,义项最多的达 24 个,平均每词 1.27 个义项。

复音词中双音节词 35056,三音节词 5703,四音节词 4365,五音节词 260,六音节词 114(百闻不如一见),七音节词 27(一朝天子一朝臣),八音节词 41(一言既出驷马难追),九音节词 4(司马昭之心路人皆知),十音节词 2(只要功夫深铁杵磨成针),十二音节词 1(只许州官放火,不许百姓点灯),固定结构 33(半 …半 …)。

把《现汉》作为现代汉语共同语的语文类通用词汇的一个载现物,进行精心的整理与爬梳,对清晰地了解现代汉语词汇的分布概貌与规律,是很有意义的。譬如汉语复音词的音节分布情况,就将最有构词能力的双音节形式一览无遗地展示了出来。这个数据与《现代汉语常用词词频词典》(音序部分)[⑦]在 2500 万字的语料中统计出来的数字绝对数上有所不同,但所占据的比率高低却比较一致,如:

	总数	一字词	二字词	三字词	四字词	五字词	六字词	七字词
《词频词典》	77482	7611	46729	11213	9633	1414	675	207
		10 %	60 %	15 %	12 %	1.8 %	0.9 %	0.3 %
《现汉》	56147	10540	35056	5703	4365	260	114	27
		19 %	62 %	10 %	8 %	0.5 %	0.20 %	0.04 %

当然二者之间也有不同。如《现汉》的一字词的比例就显得高出不少,这与《现汉》将字义分得过细有关,如只以单字为计算的话,《现汉》是 8600,所占的比例只有 15.3 %。另如《现代汉语常用词词频词典》最长的词是七字词,"统计语料 2500 万字。分词词典有词条 130691,实际统计 77482 条,最长词条有 7 个字。"[⑧]而《现汉》最长的词却达 12 个字,八字以上的达 48 例。尽管这个数字不算多,但它却给人们留下这样的思考:语文类词典对长音节的谚语、歇后语、俗语、惯用语等长词形的熟语该不该收?收到什么程度合适?

又如:历来人们都有这样的说法,现代的词语绝大多数都是多义词。可是通过调查却发现,只有一个义项的词有 42828 条,高达 76 %。全部词条平均下来每词的义项才 1.27 个。看来习常的看法离事实相去甚远。至于说《现汉》"全书单字复词的义项总计有几十万个",[⑨]有点像是信口开河了。

3.2 汉语词汇理论的深入研究与建构

有了这样一份现代汉语系统、充足、自足的词汇材料,并在数据库技术上进行多角度多层面的计量分析,再来探讨汉语词汇的诸多理论问题,将会大大有助于视野的拓展,使许多似是而非、见仁见智,或蒙胧感知、语焉不详的重点、难点变得清晰明了。

如《现汉》收录的是常用的语文类词语,可是在《现代汉语频率词典》[10]按使用度排列最常用的 8548 条词中却有 645 条不见于《现汉》,即最常用的词语中有近 8 %不见于《现汉》。这是不是因《现汉》漏收而造成的弊端呢?其实不是。"符合国家标准 GB13715《信息处理现代汉语分词规范》的词或短语一般都是语法词典的收录对象。"[11]这一分词规范中对"分词单位"作了这样的说明:"汉语信息处理使用的,具有确定的语义或语法功能的基本单位。它包括本规范的规则限定的词和词组"。它最大的特点就是收有一些结合紧密、使用稳定的词组,甚至只是一种不能独立使用的语法结构。[12]用这样的观点来看,《现代汉语频率词典》是用"机器"分出来、属于信息处理用的词典,与《现汉》有着很不相同的性质,在它里面有着许多在"人"看来难以理解的词语,如:"为的是"、"老是"、"较为"、"越来越"、"极为"。而《现汉》则应该算是为"人"服务的词典,它与为信息处理用的词典在词汇单位上就有着明显的区别。除了要遵照结合紧密、使用稳定的标准外,它还得考虑意义是否完整,是否具有独立使用的功能。由此再生发开去,就不难理解,在词汇研究中对最基本单位"词"的认识与确立中,除了多义词与同音词的划界、词与词组的划界以外,还面临着一个"人"的分词与"机器"的分词如何划界的问题。[13]再把思考的范围延伸开去,就是研究现状给人们提出了这样一个不容回避的问题:词汇研究需要根据不同的研究目的、功能,分出不同类型的研究范式,即为"人"服务的词汇研究与为"机"服务的词汇研究。

又如在断代词汇研究中,共同语词汇与各种不同来源的词汇之间如何分清处于过渡状态的成分,如何区分二者的性质,确定其身份一直是一个难点,也是词汇理论说明得最为含混的部分。现在则可以利用这份语料作出相当清楚的论述。例如,先调查第二版中的方言词,统计它们的来源方言及所占比重,再与第三版对比,看看有哪些方言词退出了共同语,哪些仍作为方言词保留下来,哪些被共同语所同化,又新增了多少方言词,新增方言词的来源如何。通过这样封闭、穷尽的专题研究,将可以清楚地观察到方言词汇与普通话词汇之间的关系及演化过程。[14]

又如词汇研究的核心是词义问题。通过观察词典的释义可以了解这个时代的词义状况,通过同一辞书前后两个不同时期释义的对比可以逼真地了解到词义的历时演变情况。在对照《现汉》三版对二版的修订中,最突出的一点就是克服了"阶级斗争"时代人们自觉与不自觉地加载在词义理解与运用上的那种"阶级斗争"意识。[15]词义诠释中阶级意识过强不仅仅是词典释义的问题,其实也是词义自身内涵的再现。通过它可以考察一个词语在不同时代表现出的升浮沉降、广狭宽窄的变化。这是时代变迁最实在的写照。

3.3 现代汉语规范词典的编纂与完善

以上所有关于词汇状况与理论的研究,都将为现代汉语规范词典的编纂和修订提供有力的理论、方法和材料。《现汉》数据库的计量研究将给词典编纂带来大量新的课题、材料与数据,对规范词典编纂的完整、精确、严密化将起到重要作用。

首先,通过数据库的计量研究,可以发现、归纳、总结《现汉》在词典编纂上成熟的、经验性的、带有规律性的东西。如《现汉》收了不少成语,而全书的四字词共有 4365,如何区分成语与非成语,一直是颇令人挠头的问题。以往人们多从结构的稳定性、可替换性、来源、字面义、词里义等角度来区分,现在通过数据库调查,发现成语类词语的释义表现出一个显著的释义特征,就是有释义"专用语",如"比喻""引申""指"等。使用了这类专用语的约占成语类

词语的 70 % ,而在非成语类的四字词中则极少出现 ,它们绝大多数采取的是对语素义直接说明的释义方式。[16]其实释义“专用语”正是立足于成语的词义内涵与词义特征基础之上的 ,是《现汉》编纂者们在大量释义实践中使用并趋于定型的释义方式。现在通过对所有四字词的计量分析 ,发现并总结出这一释义特点 ,使得对成语的区别与认识又多了一个内在的、易于把握的认知标准。

又如 96 年的修订版对 83 年版到底作了哪些修订工作 ,除了词目的增删外在释义方面还作了哪些变更。通过数据库“不匹配功能”的查询 ,发现所作的修订远远超出人们的估计。试以“面”字同语素词为例 ,两个版本共收 324 条“面”字词 ,其中 83 年版收 255 条 ,96 年版收 300 条 ,两版共有的词语是 231 条。也就是说 96 年修订时删了 24 条 ,增了 69 条。在继承下来的 231 条中 ,作了修订的有 87 条 ,高达 40 % ,其中绝大部分就是对 83 版的匡误、订正、补充、完善。[17]其中 ,除了增减例句、参见条例的完善等 ,最值得注意的是对释义的修改。如【面纱】,83 版是“妇女蒙在脸上的纱”,96 版是“❶妇女蒙在脸上的纱。❷比喻掩盖真实面目的东西 :揭开宫廷的神秘～”。看起来这是义项数量的增减 ,其实它牵涉到这个词的存在与否。如只有前一个义项 ,它只是一个单纯的指物名词 ,大可不必收入语文词典 ,正因为有了后一个义项 ,它脱离了具体的指物性 ,成为具有普遍使用价值的派生义 ,才使它显得与其它单纯的指物名词不一样 ,才成为一个社会性的语文词语。又如【面黄肌瘦】,83 版释为“形容人脸色发黄、肌肤消瘦”,96 版改为“脸色发黄、肌肤消瘦 ,形容营养不良或有病的样子”。“脸色发黄 ,肌肤消瘦”是对“面黄肌瘦”的逐字解义 ,诠释的是它的字面义 ,而“营养不良或有病的样子”才是它的形容对象。这里的修改显然是后出转精。96 版所作的修订幅度是相当大的 ,在“花”字、“白”字、“人”字等同语素词调查中 ,发现所作的修订都在 2540 %之间。目前 96 版的修订工作只有不多的简要概述 ,尚无系统的总结说明 ,[18]而现在数据库技术却使这一切昭然若揭。词典的修订不单单是一个技术问题 ,这是弥足珍贵的一份词汇词义与词典理论研究的材料。全面总结 96 版的修订工作 ,对规范词典的编纂理论和编写实践 ,都有着很好的启迪与借鉴作用。

另一方面 ,通过数据库的计量研究 ,也会发现《现汉》还存在着许多不足。数据库技术的应用使得《现汉》中许多藏而不露或若隐若现的毛病都显露出来。如收录词目是词典编纂中的第一大问题 ,收词稳当、妥贴、均衡一直是规范词典编纂家们追求的目标。无论是在《现汉》的编纂之初 ,还是修订之后 ,编纂者们在这方面都花费了大量的精力 ,也有许多经验之谈。然而在数据库的查询中仍时时可见收录中的不妥 ,该删的未删 ,该收的未收 ,收与未收之间的失衡 ,类与类之间不对称 ,不在个别。如 96 版对 83 版删复音词 4785 条 ,新增复音词 9611 条。说它这么多全删去了 ,并不尽然 ,有的原来单字下只有一个复音词的 ,现在归入单字的释文 ,只是变换了一种存在形式。说新增的都是新词语 ,做到词典的与时俱进 ,也并不尽然 ,因为不少属于以前该收而未收的漏收词语。如新增词语中有四字词 1059 条 ,其中属旧有成语的不在少数 ,如“笔走龙蛇”、“匕鬯不惊”、“拔刀相助”、“白璧无瑕”、“斑驳陆离”、“饱以老拳”、“杯盘狼藉”、“辅车相依”、“覆水难收”。再把新增收复音词与《辞海》(1979 年版) 相对照 ,发现竟然有 1700 多条词语见于后者。可见说它们为“新增词语”可以 ,说它们皆为“新词语”则否。假如能对这些或删或增的词语作详细调查 ,并参之以其它有关词频词典、专科术语词典、方俗语词类 ,相信规范词典在收录词语时将会有更扎实的理据 ,并通过这一

增一删的语料,可以窥伺到规范词典的语文性与稳定性是如何体现的。

四 余论

在《现汉》词汇计量研究的进行中,面对不断出现的新语料、新方法、新成果、新参数,会时时琢磨着汉语词汇研究的过去、现状与未来。跳出具体、大量、琐细又细致、缜密、严格的词汇计量研究,会深深地感到传统的汉语词汇研究走到今天,正面临着新的发展与抉择。

4.1 "人"≠"机"分立的词汇研究范式

当在操作数据库语料中第一次兀然发现《现代汉语频率词典》的 8548 条常用词中有 645 条不见于《现汉》时,最初生出的感觉是后者漏收。随着分析的深入,才认识到这其实是在两种不同学术规范下,用了不同的方法和标准处理"词"的结果。简言之,《频率》是为"机"服务的,《现汉》是为"人"服务的。中文信息处理的崛起对传统的汉语词汇研究是一个极大的推动。传统的汉语词汇研究在服务于信息处理的同时也促成了自己的进步,旧有范式受到冲击、面临分化就成为不可避免的事了。在这个抉择过程中,为"人"服务的词汇学与词典学研究者,应该保持清醒的头脑:

首先是明确为"机"与为"人"是两种不同范式的词汇研究,运用的方法不同,依据的理论不同,服务的对象不同,以此框彼,大可不必。二者的差异集中体现在"词"的研究上:前者是在大规模语料中完成的,它要求词库是海量的,词语多多益善,后者讲求词量的适中与适用;前者要求词结构的稳定、凝合,后者除此之外还要求词义的完整、有着较强的独立性;前者对字形和语音的统一性要求高,而对词义内容的差异程度则较忽略,后者则重在意义的同一性或差异性,并以此来驾驭词形的分与合。这些根本性的分歧必定会影响许多已有问题的解决,甚至会影响到问题存在的必要性。

其次是要充分利用为"机"服务的研究成果。如词汇研究与规范词典都要求面对的是常用词、通用词,也希望能根据频率来选词,这样的工作就完全可以利用信息处理用的词汇研究成果。现在词频数据的来源早已超过了百万级语料的规模,而是在千万级,甚至亿万级的语料规模之上获取的。对这样的统计结果,只要稍加人工干预,现代汉语通用词汇的确定将成为易事。又如汉字的使用频率、使用度、构词频率等也都有现成的成果,"现代汉字"的确立完全可以在频率的基础上来确立,而词典中主观成分极浓的"难字"≠"僻字"≠"古字"≠"生僻字"的认定,可以已矣。

4.2 强化基于计量分析基础上的词汇理论研究意识

之所以提出这一点就是因为,以往的研究中主观色彩太浓,众说纷纭的争论太繁,了无结局的问题太多。其实,许多词汇理论问题在大规模的语料计量研究中都会一目了然。如笔者最近发表的一篇关于同形词研究的论文是就 83 年版的材料而发论,[19]其中列举了数条 96 年版的例子,后来又对 96 年版的所有语料进行了"重复项查询",发现其中的矛盾凸显得更为清晰。如果脱离计量分析的基础来谈这个问题,其结论很难为人信服,怕又会陷入无休止的纷争之中。又如对如何鉴别普通话中的古词语,向来难有定论,现在通过对词典中所有相关语料进行封闭的分析,从释义用词与释义方式等形式特征入手,离析词义成分,再参之以前后时代同一语料的对比,相信要确定其"古"的身份并非不可能,甚至可以细致地发现词义成分与色彩在历时状态下的嬗变过程。

4.3 词汇学应大规模地利用词典材料

把词汇学研究与词典学结合起来,在中国语言学历史中有着良好的传统,古代的字书、词书历来也都是词汇研究的对象。当代学者中也有在这方面作出突出贡献的学者,如刘叔新先生当年的《词汇学与词典学问题研究》,就以横跨两大领域而分外醒目,符淮青、张志毅、苏宝荣等先生也都取得了大量成果。然而,之所以现在仍要提出这个问题,一是仍有人认为,只有研究"活"的口语才是正宗,而词典材料是死的材料。殊不知,能进入词典的语言材料都是经过整理、稳定了的、并具有全民性的普通词汇,是"语言"系统的词汇词义。要研究共时状态下的共同语的词汇系统,词典材料是不可替代的宝贵材料。二是对词典材料不应只是摘取式、例句式、个案式的利用,越是具有高度概括性的词汇理论研究,越是需要大规模、穷尽式、以计量分析的方式来利用语料,这样才能在更扎实的基础上总结词汇规律。超千万字、集大成、穷尽式的大型辞书,如 13 卷的《汉语大词典》,8 卷的《汉语大字典》、41 卷的《现代汉语方言大词典》、5 卷的《汉语方言大词典》都已出现。它们都是从事词汇理论计量研究极有价值的分析材料。充分利用数据库技术,大规模地利用词典材料,应成为当代词汇研究者必须具有的意识和技能。

4.4 词典编纂对数据库的更广泛利用

辞书界已经开始注意了数据库技术在词典编纂中的利用。从《辞书研究》上的两篇文章,可以看到,在短短的几年中,对数据库的利用迈开了相当大的步伐。1996 年对数据库还只是输入、编排、转换、检索、查询等低层次的利用,[20]到 2000 年已出现了"词典编辑系统"的创制与试用,表现出了迅速跟上世界词典编纂自动化、电脑化的趋势。[21]我在这里想提出的,一是对数据库的利用不要仅停留在"编"语料的过程中,而要深入到对语料的处理如采集、统计、归类、对比、分析上。二是要使数据库的使用成为"百姓"手中的寻常之物。作为词典编纂的专业人员,应做到凡是能使用电脑的人都应学会数据库的使用,像使用 Word 或 WPS 那样自如。像单音语素的义项切分与同语素词族意义之间的覆盖与呼应,是编写释义中很值得注意的一件事,可是以前只能根据顺序或倒序来查词。而在数据库中可以根据语素查询,很轻松地穷尽包括处于词中位置的所有派生词,使同语素词成为一个全封闭的系统呈现在编写人员面前。又如词典修订中的增删,也是编写过程中需时时留意的,而数据库对此也能自动进行排比对照。只有做到数据库的普及使用,才能更好地把科学、准确处理语料的精神贯彻到所有编纂过程之中。要防止词典编辑系统那样全功能的数据库软件成为工程家手中的专利产品或只限于个别大单位使用的"阳春白雪"。

附注

①有关论文可见:黄景欣(1962)《试论词汇学中的几个问题》,刊《中国语文》第 3 期。刘叔新(1964)《论词汇体系问题——与黄景欣同志商榷》,刊《中国语文》第 3 期。周国光《概念体系和词汇体系》,刊《安徽师大学报》1986 年第 1 期。刘叔新(1990)《汉语描写词汇学》,商务印书馆。

②晁继周、单耀海、韩敬体(1996)《关于规范型词典的收词问题》,见《〈现代汉语词典〉学术研讨会论文集》,商务印书馆,第 70、72 页。

③李建国(1996)《〈现代汉语词典〉与词汇规范》,见《〈现代汉语词典〉学术研讨会论文集》,商务印书馆,第 83 页。

④梅千驹等(1983)《同义词词林》,上海辞书出版社。

⑤林杏光、菲白 (1987)《简明汉语义类词典》,商务印书馆。

⑥⑲苏新春 (2000)《同形词与"词"的意义范围 ——析〈现代汉语词典〉的同形词词目》,《辞书研究》第 5 期。

⑦⑧刘源 (1990)《现代汉语常用词词频词典》(音序部分),宇航出版社。

⑨鲍克怡 (1996)《现代汉语工具书的代表作》,《〈现代汉语词典〉学术研讨会论文集》,商务印书馆,第 22 页。

⑩北京语言学院语言教学研究所编 (1986)《现代汉语频率词典》,北京语言学院出版社。

⑪俞士汶等 (1998)《现代汉语语法信息词典详解》,清华大学出版社、广西科学技术出版社,第 20 页。

⑫任海波、范开泰 (2000)《现代汉语真实文本短语标注的若干问题》,《语言文字应用》第 1 期。

⑬苏新春、顾江萍 (2000)《"人"、"机"分词差异及规范词典的收词依据 ——对 645 条常用词未见于〈现汉〉的思考》,刊《辞书研究》第 5 期。

⑭苏新春 (2001)《普通话词汇系统对方言词的吸收与更新 ——〈现汉〉方言词研究》,刊《语言》,总第 2 期,首都师范大学出版社。

⑮苏新春 (2000)《当代汉语变化与词义历时属性的释义原则 ——析〈现代汉语词典〉二、三版中的"旧词语"》,刊《中国语文》第 2 期。

⑯余桂林 (2001)《成语的语义特征与释义特点 ——〈现汉〉(第二版)四字词研究》,刊《汉语词汇理论与实践研究》(论文集),商务印书馆。

⑰赵翠阳 (2000)《从"面"字语素词看〈现汉〉96 年版的修订》,第三届现代汉语词汇学术研讨会会议论文,厦门。

⑱韩敬体 (1997)《〈现代汉语词典〉修订工作概述》,《辞书研究》第 1 期。

⑳王伟 (1997)《从〈现汉〉修订谈词典编纂中的应用及展望》,《辞书研究》第 1 期。

㉑陆汝占 (2000)《汉语词典编纂一体化环境》(上)(下),《辞书研究》第 2、3 期。

作者简介

苏新春,男,1953 年出生,江西南昌人。1985 年毕业于华南师范大学汉语史专业,现为厦门大学中文系教授,主要从事现代汉语词汇研究。在词汇研究中主张以下基本观点:语义为主,语形为辅;语言结构与语言文化相结合;语言阐释与语言描写相结合。