汉字属性字典及其软件系统的研试

一九八五年年底，北京图书馆研制的“汉字属性字典及其软件系统”通过鉴定。有关计算机和汉字方面的专家认为，这一研究成果填补了国际上计算机在处理我国文字信息支撑软件方面的空白，系统功能较为齐全，具有重要的实用价值，并可作为国家制订有关汉字属性标准的重要依据之一，它的推广、应用必将加速汉字处理应用系统的建立。

为什么要研究这一系统？人们知道，电子计算机处理西文以及日文在软硬件基本功能方面已不存在问题，因而一些使用要求十分复杂的信息处理应用系统能够陆续在许多国家建成，在国民经济等许多领域发挥重要作用。而计算机处理中国汉字取得迅速进展还是最近几年的事。从汉字的输入、到大型字库的建立和管理、支持汉字处理的操作系统的建立、满足各种使用要求的支撑软件系统的配备以及高精度的输出技术等，仍处在完善或研制过程中。值得注意的是，不仅我国，美国、日本一些有名的计算机公司也在投入这方面的研究和开发工作。只有这些软硬件基本功能完备了，一些复杂的涉及汉字信息处理的计算机应用系统如大型图书情报信息中心和网络才能建立起来。其中满足我国汉字信息处理要求的支撑软件系统只能由我国来完成，因为它必须以我国传统的复杂的文字处理需求为依据，别的国家很难解决。汉字属性字典及其软件系统的研制正是为了填补这一空白。

这一系统有哪些功能？第一，为国家已经公布的信息交换用汉字编码字符集中的汉字和非汉字编码提供对照的其它国家和地区的汉字和非汉字编码、电报码，以解决我国与不同国家和地区之间机读信息的相互交换。如我国的GB2312—80同日本JIS6226、日立字符集、台湾CCCII、“通用汉字标准交换码”、大陆与台湾公用的电报码、以及非汉字字符集ASCII、EBCDIC等字符集之间，实现相互转换。第二，提供按照国家有关部门规范的汉语拼音、部首、笔画笔形、四角号码等排序方法对汉字进行排序所需要的序值信息，以满足计算机系统建立文件、显示、打印汉字信息所需的排序要求。第三，提供汉字与汉语拼音和威氏拼音的转换信息，以满足国内外一些计算机应用系统对汉字罗马化方面的要求。第四，提供国家正形的通用汉字与异体汉字（包括繁体字）相互连接的信息，以满足汉字信息检索和文字规范方面的使用要求。第五，提供国内经过考核效率较高而又易于掌握的小键盘汉字输入编码，供录入人员选用。如为国内以形为主的方案，以音为主或音形结合的字词输入方案留下空间，待国家有关部门考核后补入。第六，可为文字研究和汉字信息处理部门提供汉字音、调、偏旁部首、笔画、笔顺、笔形等各方面的属性元件的信息，以便这些部门利用现代化手段开展各种统计、分析研究。上述这些信息均可以以磁带、软盘和纸张打印的形式提供给使用者。

关于汉字属性字典。数据是软件的基础，根据系统设计的要求，首先需要编制汉字属性字典。而字典又要根据系统功能的要求，确定下列属性元素(约五十个数据项)：

1、国标字符集编码，集属标志，区位号。

2、其它字符集对照编码：①日立码，②CCCII码，③JIS码，④ASCII码，⑤EBCDIC码，⑥标准电报码。

3、国内优秀的输入方法编码：①以音

为主输入编码，②以形为主输入编码，③其它类型输入编码。

4、字符类型区分：

①标志1：

A—Z：汉字

a—z：国标中非汉字部分符号、数字、外文字母、音符等。

②标志2：

O：正形汉字

1：非正形汉字

5、汉字音、形特征：

①注音（异音数量最多收7个）：

（1）汉语拼音（包括音调标志）

（2）注音字母

（3）威氏拼音

（4）国语罗马

②整字笔画数

③整字起笔至末笔笔顺笔形及其编号

④部首及其顺序号

⑤汉字部首外部分笔画数

⑥汉字部首外部分起笔至末笔笔顺笔形及其编号

⑦四角号码

6、异体汉字连接：

①异体字个数

②各异体字字形及其国标集属和编码或区位号

③其它字形、结构参见信息

7、备用栏位。

该系统的特点。第一，既满足图书馆情报界对汉字信息处理方面的要求，也兼顾一般事务管理系统、办公室自动化以及文字研究、出版部门处理汉字的需要，系统及其产品具有通用性。做到了一家一次辛劳，众家长久分享、受益。如汉字四种排序方法的实现、国际编码与多种字符集编码、电报码及拼音的转换，与不同输入方法编码的连接、汉字各种属性的拆分等，都体现这一点。第二，凡是国家一级、部一级、图书馆情报界、工具书出版界等领域关于处理汉字、目录组织等方面的标准、规则，只要在计算机功能允许的条件下，系统都予以遵循。从而保证软件系统具有实用性和规范性，为国家今后制定有关汉字属性方面的标准提供依据。如以国家标准GB2312—80的汉字及其编码为依据编制属性字典，采用1964年教育部、文改会、语言所、文化部四单位确认的四种汉字排序方法，以及图书馆情报界、国家政府机关、工具书编纂部门对书目、人名、团体名称、产品名称汉字的排序规则；编制汉字属性字典时又依据了国家有关部门历年制订的有关注音、笔画、笔顺、笔形、文字正形、正字、标准电码、四角号码等一系列规范。第三，软件系统具有可扩性。由于目前国家公布的汉字交换码基本集只有6763个，而我馆用字量远远超过这个数字，因此整个属性字典的主文件不局限于基本集，已把国际基本集（G0），以及将来陆续公布的各辅助集（简体字系列G2，G4，繁体字系列G1、G3、G5等）全部考虑在内，加以统一的标识和处理。不仅对字符集的增加有可扩性，而且对将来根据需要增加新的汉字属性也留有扩充空间。第四，系统的维护既可靠又方便。在软件整体结构上，规定数据的增改删只有一个入口，并在使用权上加密，防止多渠道更新可能造成的数据混乱。而在数据更新时，只要改动某一文字的有关属性值，相关联的功能软件能及时调整系统整体，确保系统整体的可靠性和可恢复性。这对于我国汉字量大、多字符集，而且持续相当一段时间才能完成的情况下，十分必要。而且对一般用户在使用基本字符集的基础上，可能还需增加一些内部用字，当需要增加属性值时，可容易地调整。第五，系统具有可移植性。为了保证软件的通用性，系统既可以整体移植，又可以对部分软件和软件的处理结果进行移植。这样除满足了少数大用户对汉字属性有全面使用要求

外，也满足了大多数用户的一般需要。为了保证这一点，整个软件源程序采用标准COBOL语言编码，转移时易于改写；属性主文件及其派生的各个分文件，结构简明，数据单元和记录长度固定，用户可直接使用的文件中的数值数据为16进制压缩数据，文件组织是顺序的，用户很容易掌握，运行能保证有较高效率，并且不受计算机环境限制，可以移植到各种大、中、小以及微型机上去。第六，系统结构模块化，程序可读性强，并注意了结构化，以便于维护。第七，数据编制符合规范，力求准确无误。措施是层层把关，相互校对，最后由有经验的文字专家总校。不只是数据本身，而且在数据结构的安排上也做了比较周密的考虑，如我国国标码与台湾省码的连接，采取一对多的结构，从而保证了计算机识读台湾码时不会有遗漏。

使用效果。该软件系统已经装载在M—150H电子计算机系统上，并利用排序软件处理北京图书馆的有关业务，运行效果有以下几点：

1.提高工效极为显著。

对GB2312—80中6763个汉字进行排序。偏旁部首法排序：所用CPU时间37秒，I／0时间4分09秒。汉语拼音方法排序：所用CPU时间58秒，I／0时间11分03秒。笔画法排序：所用CPU时间37秒，I／0时间4分14秒。四角号码方法排序：所用CPU时间59秒，I／0时间15分08秒。若用手工方法对6763个汉字排序，据我馆有经验的排片人员估计，无论用上述哪种方法，一个人至少也需要10天左右时间才能完成，即使把慢速的计算机终端打印机打印时间（平均30分钟）也包括在内，提高工效在240倍以上。

以姓名为键，对全馆1349名职工基本档案排序，使用四种方法排序，每种平均用CPU时间17秒，I／0时间3分50秒。把终端打印机打印时间（平均20分钟）计算在内，仍比手工处理提高工效在150倍以上。

我馆每年都要编制大量的专题目录、联合目录、新书通报，这些目录，少则几千条、多则数万条款目，仅在目录排序上就花去不少时间，由于手工效率低费工时，许多应该编辑的书目都不能按时完成。预计采用这一软件排序将大大缩短编排周期，减轻排序人员细而繁的劳动负荷，加快新入藏文献资料的报道工作，为科学工作者更好地提供服务。

2.处理结果准确性高。

由于建立每个汉字排序键值的程序是按照文字的各属性序值，严格地以不变的规则进行比较计算，因而利用键值排出的结果自始至终体现了一致性，绝没有手工排序时常出现的疏忽、或因人而异、因时而异造成的混乱和差错。三个多月的运行表明，排序中如果出现差错，不是程序本身错误造成的，只是由于个别文字某个属性值不对，一旦纠正则永远有效。最近计算机运行的结果反过来还纠正了校对人员疏忽或掌握规则前后不一致的一些错误，充分显示了计算机利用该软件排序的优越性。

3.处理结果科学性强。

由于对汉字属性从各个角度进行了科学的拆分与赋值，而且这种拆分深度大大超过手工传统作法，如所有汉字整个字以及部首外部分都从第一笔至最末一笔将笔形都做了分类和赋值，而手工排序时仅取前五笔（图书馆目录规则规定），软件排序时比较了上述的赋值，因而排出的结果（四种方法都利用了笔形属性）远比手工精确、科学。还有在201部首序列表中，我们对一些附加的归入部首也赋了序值，并利用它对汉字排序，排的结果层次精细分明，富有规律性，易于使用者查索。

4.可多途径应用

由于属性字典综合了汉字的音、形多方面特征和各种对应的编码，这一系列基本元素形成的属性主文件，为各个领域应用电子

北京图书馆举办科组长研讨班

〔**本刊讯**〕为了加强科组建设，提高科组基层干部的思想认识和管理水平，北京图书馆于上年举办了两期科组长研讨班，有120人参加。

研讨班上，由谢道渊等四位馆领导分别作了《科组长的基本职责与应具备的素质》、《自觉进行职业道德的修养，做一个称职的图书馆工作人员》、《实现科学管理的几个认识问题》和《关于新馆规划和五年发展纲要的设想》的讲话（全文另发），并进行了答疑和讨论。在此基础上，科组长进行个人小结，经研讨小组评议后写成文字材料，交人事处存入干部业务档案，备作考核干部的参考依据。研讨班还要求科组长在学习结束后一个月内，针对本科组执行岗位责任制的具体情况，写出存在的问题、改进的设想及具体落实措施，由所在部、处领导审查并写出评语后送交人事处，以督促执行。

通过学习讨论，科组干部普遍认识到科组长的基本职责首先是搞好本科组的行政管理工作和思想政治工作，而科组长本身的素质，特别是政治思想素质高低对科组长来讲是十分重要的问题。要搞好科组工作，既要靠抓岗位责任制的落实，又要靠抓人的思想觉悟的提高，加强工作人员的职业道德修养。作为基层单位的管理干部，科组长的精神面貌、政治思想素质、管理能力如何，直接影响着科组的工作效率和全馆社会效益的发挥。研讨会上，大家认识到科组长工作在全馆十分重要，增强了在科组岗位上任劳任怨、尽职尽责为人民服务的自觉性，决心在加强自身建设上下功夫。大家明确了“管理就是生产力”，决心踏踏实实学管理；要有进取精神，进行目标管理；要把社会需要和社会效益作为科学管理的出发点；要用新的管理观念去完善和落实岗位责任制，落实业务规范，进行数据化管理；还要胸怀广阔，有“用人之道”。

北京图书馆面临新馆将于1987年建成并投于使用的任务，大家既受鼓舞，又感到有压力，一致表示要把压力变成动力，以创业者的献身精神，完成历史赋予的使命。

科组干部非常欢迎这种学习和探讨。大家希望能将这种培训方法制度化，以期不断提高科组干部的素质和管理水平，适应新时期的要求。

计算机开展多途径应用提供了良好条件。我馆准备利用配备国标字符集的计算机系统，通过这一支撑软件，识读台湾省发行的机读目录（MARC），加快台湾图书的编目、采购和检索工作。利用这一软件对书目排序，进行中文机读目录中汉语拼音数据的自动生成，统计本馆汉字使用频度，进行正体字与异体字的连接规范控制，利用小键盘终端快速录入数据。文字研究部门也可以利用其各种元件开展统计、研究。（北图供稿）