中 文 信 息 学 报
第3卷 第3期 JOURNAL OF CHINESE INFORMATION PROCESSING Vol.3 No3

中文信息处理系统的性能准则

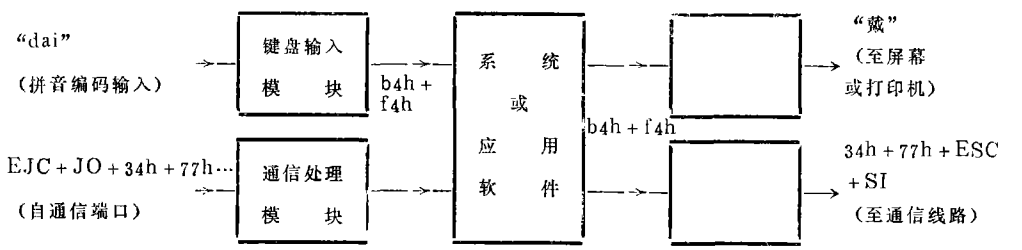
——内部码、汉化程度及其它

戴瀛洲 李成国
（中国电子设备系统工程工司）

一、引 言

计算机科学技术是当代信息社会进步的有力杠杆。这一高科技产品的发展之快，使得最古老的象形文字之一——汉字的计算机处理研究也具有了世界的规模。中国等国都在大量生产具有中文信息处理能力的计算机，仅在国内，就有超过50种以上的计算机系统和500种以上的汉字输入方案。这些系统各有特色并被广泛使用，然而并不统一。另一方面，ISO/IEC、JTCI／SC2倡导的关于NLS（National Language Support）标准化的研究已在多8位信息编码体系方面取得重要进展，具有统一字符集合、不同民族和文字的信息共享的计算机系统正在成为现实。基于上叙考虑，中文信息处理的标准化研究既是信息技术方面的重要基础工作，也是中文计算机开发应用实际迫切的需要。

中文信息处理系统（CIPS）的工作流程请见图一。一般地说，CIPS有两类信息输入源，其一是由标准键盘来的汉字输入码（例如“dai”），由键盘输入程序处理后产生汉字内部码；其二是由通信线路来的标准信息交换码（参见第2节），也变换为内部码。尔后内部码（例如高位置1的双字节ASCII码）由系统处理、存贮或作其它用途。最终，系统也有两类输出源，即用于显示打印的汉字输出码（常常是内部码）和用于通信的标准信息交换码。由图一可见，汉字“戴”的编码在处理过程中是变化的。



图一、CIPS工作框图，汉字编码是变化的。

由于中文信息处理的复杂性，在国家标准GB2312（“信息交换用汉字编码字符集（基本

本文1989年6月1日收到

集)”）颁布八年之后，对于输入／输出码、内部码等等仍未能有相应的标准，在设备选择，系统互连和用户应用等方面都引起一些困难。

本文重点讨论了中文信息处理系统的内部码设计与评价，系统和输入／输出设备性能的有关问题。在第二节提出了关于汉字内部码的12项准则、扩展CC和CNCC内部码的定义。在第三节介绍了关于系统汉化程度的概念。在第四节讨论了汉字终端的评价问题。在第五节简述了几个实际问题的中文信息处理系统及其性能。

在进一步讨论之前，列出本文所用的有关概念和术语如下：

汉字（Chinese-character）　也用作Hanzi、中文（Chinese），其图形模式称为点阵（Font）。

汉化（Hanzification）　此术语用来描述在西文计算机上的开发工作以使其能够处理中文。

汉化程度（Hanzification level）　请参见第三节。

汉字编码（Chinese coding）　汉字至少有四类编码：输入／输出码、交换码（Interchange code）和内部码（Internal code，也叫处理码（Process code））。

中文信息处理系统（CIPS）　具有中文操作系统和中文文本处理实用程序的计算机系统。

二、汉字内部码的12项准则

中文信息编码有着其内在的特殊性。首先，汉字编码必须与现行的单字节西文字符集合兼容，同时由于汉字集合的大基数必须用多字节编码。其次，尽管西文字符的信息交换码和处理码具有一致性，汉字的内部码（处理码）一般不同于交换码。第三，编码不仅基于中文、信息和计算机科学理论，也具有工程性。目前流行的CIPS几乎全是国内开发的系统，将来的编码标准也颇可能会是某个“事实的（de facto）”标准，它具有相对于实现可能性而言是最优的性能。

表一、字符编码集合的基本标准（单字节）

标　　准	IS 646	IS 2022	IBM
相应国标	GB 1988	GB 2311	
等价集合	ASCII	扩展 ASCII	EBCDIC
控制码子集	C0 = {[0,1F],7F}	C0 = {[0,1F],7F C1 = {[80,9F],FF}	CE = {[0,3F],FF}
图形码子集	G0 = {[20,7E]}	G0 = {[20,7E]} G1 = {[A0,FE]}	GE = {[40,FE]}
应用举例	大多数计算机	IBM PCs, DEC VAX （多国字符集）	多种中大型机

有关西文字符编码的基本标准、常用的汉字内部码方案分别在表一、表二列出。ASCII码无须赘述。扩展ASCII码由DEC VAX和IBM PC等计算机采用，然而关于GI集合的定义是不同的。EBCDIC码的GE集合中约有90个尚未定义的空码位。这些编码集合满足交换码／处理

码的一致性，且至少其中的ASCII码等价子集已由商品化的通信、网络和数据库软件自动转换，以满足系统互连时的要求。

上述优点对于汉字编码都不再存在。标准信息交换码（GB2312）仍未见有直接用作内部码的实例，除非附加了有关标识。CC（Chinese-character）内部码虽然为许多计算机系统所采用，但是一般系统的汉化水平都不高，不能保证汉字处理的完整性。DBCS-HOST码采制SO/SI标识字节，使得有关软件的汉化非常困难，而且没有第三方的产品支持。FP(Free-Position code)码系由国内首先用于IBM中大型机，然而受到EBCDIC集合中空码位重新定义的限制。

内部码的设计和标准化是当前中文信息处理学科的一个主要问题，也是一大难点。有关设计和标准化工作应当以性能评价为基础。根据在微机和小机CIPS开发中的经验教训和对各类内部码的分析测试，并且借鉴国内有关的研究成果，我们总结出汉字内部码设计和评价的12项准则。

1. 编码效率准则。该准则要求实际使用的码长应短，然而这受到诸如系统汉化水平等

表二、常用的汉字内部码方案（多字节）

名　称	GB2312	CC	DBCS-PC	DBCS-HOST	FP空位码
描　述		国标移位	分段映射	分段映射	码位映射
定　义	B1=[21,7E] B2=[21,7E]	B1=[A1,FE] B2=[A1,FE]	B1=[81,FE] B2={[40,7E], [80,FC]}	B1=[40,FE] B2=[40,FE]	B1—空位码 B2—任意码
附加标识	需　要	不需要	不需要	需　要	不需要
标识方式	ESC+SO, ESC+SI	MSB=1	MSB=1 与 整字处理	SO, ST 方　式	MSB 方　式
应　用		PC-CCDOS, CUNIX, VAX/CCVMS	IBM PC3350	IBM 中、 大型机	IBM 4300, 3080
集　合	ASCII			EBCDIC	

表三、扩展CC和CNCC内部码的集合定义

名　称	ECC Level 1	ECC Level 2	ECC Level 3	ECC Level 4	CNCC
定　义	B1=[A1,FE] B2=[A1,FE] (与CC码相同)	B1=[80,FE] B2=[80,FE]	B1=[A1,FE] B2={[A1,FE], [20,7E]}	B1=[80,FE] B2={[80,FE], [20,7E]}	B1,B3,B4= ["A","Z"] B2=["0","9"]
标　识	不需要附加标识				
应　用	参见第5节				

等因素的制约。例如，系统文件名或数据库汉段名的定义集合基数通常小于40('A'－'Z','0'－'9'等)，为直接满足这些名称汉化要求的CNCC（character-Numbercharacter-character）内部码就需要四个字节。

2. 集合基数准则。集合基数定义为可以编码的汉字和符号的个数及其扩充性。汉字集合的基数对于一般的使用需求为3K，对于诸如国家图书馆级的要求为30K以上。

3. 与国际性的标准的兼容性。它定义为汉字编码与有关国际的标准的兼容和符合程度，涉及国际性标准化组织的标准如IS，第三国的标准如JIS，工业标准如EBCDIC。而且，汉字编码应当与控制码、保留码和其它字符集合相区分。

4. 与国家标准的关系。它定义为与有关的国家标准例如GB2312的一致程度、映射算法的复杂性等。

5. 等长性。汉字内部码也即处理码应当具有相同的码长，并且应与计算机的指令字长相适应，否则处理和存贮均有不便。

6. 保序性。意指编码与二进数字排序规则的符合程度。在中文信息处理中经常要求汉字按拼音或部首排序，如果汉字编码自然有序将是方便的。顺便指出，虽然GB2312的一级和二级汉字子集各为拼音排序和部首排序，整个集合却是无序的。

7. 连续性。过多的非连续码位可能会影响处理效率。例如，FP码必须使用查表来判定数据的有效性。

8. 独立性。通常希望编码与外部因素如环境和实现相独立。编码不应受诸如系统结构、软件应用、通信规程等等因素的影响。

9. 定义的完备性。它涉及汉字定义的唯一性和确定性、与ASCII和EBCDIC码的区别、内部码与交换码的一致性等等。唯一性意味着所论集合中的内部码位应当与汉字一一对应（CNCC码可能不满足这点），确定性（亦称防移性）意味着汉字的确定不受多字节内部码字节偏移的影响。

10. 编码的一致性。它涉及汉字显示、存贮和处理的字节映射关系。内部码为两字节长一般是合适的，因为一个汉字的显示宽度常为西文字符（ASCII码）的两倍，I/O驱动软件也因此无须特殊处理。使用SO/SI标识字节的内部码一般不满足处理的一致性，例如，包含两个汉字的字符串的长度并非包含单个汉字的字符串长的两倍，因此需要额外的处理。

11. 操作的完整性。作为最小处理单位的汉字处理操作应当具有类似transaction（译作事务或者交易）的属性，即它必须是完全成功或者彻底回退的。完整性有三个基本含义：处理不受西文字符操作例如大小写变换的影响；必须每次处理一个完整的内部码（亦称汉字整字处理），或者至少能检测内部码的残缺状态，而不致引起字节偏移的位置错误（参见第4节）；边界处理不应引起内部码的残缺错误，这类错误可能发生在行末、块或窗口边缘，并且可能是动态的。完整性与系统实现有关。

12. 非均质环境下的适应性。内部码的设计必须考虑包括不同计算机，不同操作系统和不同编码体制的分布式信息处理系统的综合要求。考虑到ASCII和EBCDIC编码体制的共存，汉字内部码至少应在这两个体制下分别统一、否则在通信、网络、数据库和其它软件环境下系统互连所要求的内部码转换将会非常复杂，系统汉化也会非常困难。这一问题涉及许多方面，本文因篇幅所限只能点到即止。

不难看出，汉字内部码的性能与系统实现有关，离开实现的可能性和水平去谈内部码的设计是没有意义的。另一方面，ISO提出的多字节编码体制指出了解决上述问题的新方向，对将来的新的IS标准理应予以足够的重视。

三、系统汉化程度

系统汉化是一门非常复杂的技术。计算机系统对于中文信息处理的能力由于系统结构、所用字符集合基数、系统汉化水平的不同差异很大。粗略地说，系统汉化分为三种模式：初级模式是接插兼容模式，通过使用兼容的汉字I/O设备的适宜的内部码类型实现汉字处理，计算机主机软件不作改动。这种模式不能保证操作的完整性（整字处理）。如果使用两字节内部码，则计算机必须具有8Bit字符集，而且汉字只能在数据一级使用。其次是预处理模式，它在第一级的基础上，通过开发上层人机接口软件部分地实现整字处理。水平最高的是核心改造模式，它可以取得最好的性能，然而开发难度也最大。这种模式基于系统分析和跟踪，修改系统和其它软件的核心部分，从而保证在任何数据结构中的汉字适应性和操作的完整性。诸如内部码截断和字节偏移等问题在开发中逐一予以解决。

汉化系统的性能指标包括与原系统的兼容性、汉化水平、汉字码标准、系统可靠性，以及系统开放性。为了评价系统汉化的水平，基于经验的总结上，我们引入了汉化程度的概念。汉化程度系指系统汉化或汉字处理的细微程度，它反映了计算机系统对于中文信息的处理能力。汉化程度定义如下：

一级（大粒度）：汉字能够在系统中正常输入／输出，作为数据或字符串常数。这也意味着汉字不受诸如大小写字母变换，位屏蔽等操作的影响。隶属度可以定义为：

完整汉字集合的I/O功能	0.4
汉字作为数据	0.25
汉字作为字符串常数	0.25
有关信息汉化	0.1

二级（中粒度）：在一级的基础上，操作系统、实用程序和其它软件的核心部分均能正确接纳与处理汉字。例如，文件名和数据库字段名能用汉字定义，程序接口的中文信息处理不受限制。其隶属度通常取为：

核心汉化	0.35
系统数据结构（文件名、命令名…等）	0.15
软件的字段名、变量名…等	0.25
程序员接口	0.15
有关信息汉化	0.1

三级（小粒度）：在二级的基础上，实现有关软件的汉字整字处理。汉字处理不再出现字符相似匹配、边界内部码截断或控制夹插等错误，因而操作的完整性得以保证。其隶属度不妨取为：

整字光标（移动、删字或插字）	0.1
在边界或行末的内部码截断	0.15
字符串相似匹配	0.15
控制码夹插	0.35

按拼音与部首排序	0.15
完全的中文界面	0.2

四、汉字I/O设备的评测

汉字I/O设备的质量对于CIPS的整体性能具有很重要的影响。某次上十种国内外生产的汉字终端的测试结果表明，所测设备几乎无例外地在汉字处理方面的性能远不够理想。汉字终端最主要的问题之一控制码夹插仍未得到足够的重视。

代码夹插是指一个汉字的多字节内部码中由于操作而夹入某些其它字符，从而操作的完整性受到破坏。一个类似的情况是内部码字节偏移。这些问题在多用户计算机系统中普遍存在而且变化多端。它们本质上是汉字整字处理的问题，然而解决这些问题所要求的汉化水平常常高得难以实现，目前看来，从系统和终端两个方面同时对其加以限制是有利的。

因此，汉字终端的测试和评价具有重要意义。评价包括四个方面：基本功能练验、仿真性能测试、汉字适应性测试和实用软件运行考核。

基本功能检验包括通用指标的检验、屏幕闪烁及余辉、显示速度、接口及汉字点阵等有关标准，汉字输入／输出设计的适用性等。

仿真性能测试系指所测终端与仿真的西文终端及有关标准（例如VT100）的性能一致程度，包括键盘定义、ESCAPE序列全集的处理等，以保证仿真设备具有同等功能。

汉字适应性测试是指汉字设备所特有的扩充功能的测试，例如编码标准，汉字／西文字符集合的兼容性、编码的一致性和操作的完整性。我们强调指出，此处适应性定义为基于相应的终端标准和中文信息处理标准的扩充性能，并不局限于原西文终端标准。从这个意义上说，所谓100%的仿真尚不足以保证汉字处理的正确性。

运行考核可以弥补测试的不完备性。系统的标准验证过程、屏幕或终端管理软件、实际的应用软件以及专门开发的程序均可用于运行考核。

五、中文系统的实例

本节给出几个实际的系统例子用以说明前几节所述的观点。这些系统在国内都有一定影响而且具有优良的性能。

CCVMS中文信息处理系统。它在VAX系列小型机上实现，通过操作系统、I/O驱动程序和信息软件的核心改造，在各级和各种数据结构中的汉字接纳与处理都是比较理想的，在整字光标、边界截断、系统服务方面实现了整字处理，从系统和终端两侧同时限制了控制码夹插问题。汉化程度接近三级。CCVMS与原VMS系统完全兼容，无附加开销，无附加操作。它所采用的二字节ECC内部码（参见表3）的最大集合基数为28K，通常使用的CC内部码具有普适的意义。

ORACLE数据库管理系统。它是我方与美国ORACLE(中国）公司合作开发的中文版本。通过修改源码实现了三级汉化程度，这是国内所有中文软件的最高汉化级别。汉字整字处理在整字光标、相似匹配（SQL语言LIKE子句）、边界的内部码截断、控制码夹插、汉字定位、子串截取等方面均得以实现，并且提供了拼音和部首的排序函数（SQL语言ORDER BY

子句）。在ORACLE中可望使用ECC内部码的全集。

CAW-DOS中文操作系统。它在IBM PC-DOS系统上实现，除具有CCDOS（国内最普遍的IBM PC微机中文操作系统，汉化程度约为二级）类似的性能之外，CAW-DOS的特点是同时采用了CNCC和CC两种内部码类型。CNCC内部码可以适用于各种计算机系统并且实现完全的插接兼容。因此这一系统可以用作尚未汉化的计算机如IBM中大型机的汉字终端。为了避免汉字与可能的具有C-N-C-C类型的字符串发生冲突，可以定义冗余编码集合，使得用户禁止的字符串组合不致影响汉字集合的基数。

六、结　论

容易看出，上面所有的讨论都是针对通用计算机系统中文信息处理某些特定的方面进行的。虽然关于内部码是这一领域中的热门论题，然而将编码与系统实现中的一些难题诸如编码的基本规则、汉化程度、系统I/O设备的评价等联系起来进行讨论的文章尚不多见。作者希望本文的观点能有助于这方面的研究，并愿意展开进一步的讨论。

七、鸣　谢

作者感谢华北计算技术研究所王之灌先生、国家科委信息系统研究所张轴材先生给予的有益启发。作者也感谢所有为本文提供帮助和给予建议的朋友。

参 考 文 献

[1] 张轴材，汉字内部码和汉字数据类型，ICCIP'87，北京（1987）

[2] 江　涛，VAX汉字系统的若干问题，《(中国)计算机世界》，1987.12

[3] 钱培德，中国古文字信息处理系统CAW-DOS的设计，《中文信息学报》，Vol.2，No.2，1988.

[4] 戴瀛洲，VAX汉字外设的若干问题，中国计算机学会中文信息技术专委会年会论文，常州.（1988）

[5] 戴瀛洲、韩柯，ORACLE中文分布式关系数据库管理系统，中国计算机学会分布式系统专委会年会论文，旌德（1988）

SPECIFICATIONS OF CHINESE INFORMATION PROCESSING SYSTEMS

——Internal code, Hanzification level and so on

Yingzhou, Dai　　Chengguo, Li

(China Electronic System Engineering Company

6 Wanshou Road, Beijing, P.R.C.)

ABSTRACT

This paper begins with basic concepts of Chinese information processing, introduces popular Chinese internal code setups, discusses deeply the standardization of

the codings and the evaluations of Chinese information processing systems.

The paper points out that the Chinese coding has its inherent specialarity and proposes at the first time the twelve basic rules for Chinese internal code. These rules are coding efficiency, set cardinal, foreign standard compatibility, national standard relation, length equality, order-preservary, continuity, indepedence, definiton perfection, coding consistency, operation integrity and suitability under non-homogeneous circumstances.

It presents that one of the main target of Chinese systems can be evaluated by the Hanzification level which is defined having three level, discusses the evaluations of Chinese I/O devices and emphasizes that the so called 100% compatibility with west standard devices may not be adequate for Chinese information processing.

Sevelal actual Chinese information processsing systems that are of influence in China are also introduced.

Keywords: standardization, internal code, Chinese information processing, Chinese computer.