ISO

INTERNATIONAL ORGANIZATION FOR STANDARDIZATION

ISO-IEC JTC1/SC2/WG2

MULTI-OCTET CODED CHARACTER SET

ISO-IEC JTC1/SC2/WG2

N 561

February 16, 1990

Title : On HCC — More Facts and Analyses

Source : China

Requested Action : For Consideration at the Ad-hoc Meeting in Seoul

三论 HCC

目录

前言

1　从文字的定义看 CJK Unification

1.1 定义

1.2 我们的观点

2　编码的本质

2.1 联系与指针

2.2 命名

2.3 字义何在

2.4 字义字音的多元性与相对确定性

3　分析一种误解

3.1 字义字音不应是字符编码的依据

3.2 字的信息不因混合编码而丢失

3.3 同形字的统计同义性

4　编码字符怎样反映使用者的文化、历史和习俗？

5　HCC 的实现基础

6　HCC 的应用特点

6.1 字量

6.2 输入

6.3 输出

6.4 代码转换

6.5 排序

6.6 机助翻译

6.7 数据库

7　结语

前言

中国的 N480 提案，正式提出了中日朝汉字统一编码的建议 (HCC)，N513 又对 HCC 作了进一步的说明，本文是第三次论述 CJK Unification。

CJK Unification 的讨论，牵涉到一系列基本的概念和原则。其中最重要的是：

第一　什么是文字 (script)？

第二　编码的实质是什么？什么是字符编码和编码字符？

第三　字符的形音义是什么关系？字符怎样反映使用者的文化、历史、习俗？

此外还涉及一些实现与应用的问题。本文将针对上述各款阐述中国的观点。

1 从文字的定义看 CJK Unification

文字的定义不下十个。但我们只能从信息处理和字符编码的角度，基于各方同意的定义来讨论问题。

在 WG2 的安曼会议上，根据各国代表的意见，对文字(script) 的定义进行了修正。第2个 DP 10646 的4.2.10，如实反映了这一变化。

1.1 定义 文字是用于一种或多种书面语的图形字符的有鉴别性的完备集。

中国专家认为，这一定义，从信息技术和语言文字的角度出发，客观地、准确地描述了文字的特性，为我们讨论问题奠定了良好的基础。

■ 首先，文字是一个集合的概念，而且是 "完备集"，不是子集。任何个体的、局部的性态，都不能表征总体的文字。我们在讨论中应注意集合与统计的概念，不可以点代面、以偏概全。

■ 其次，文字是构成书面语的要素。文字的元素是字符，字符本身是一种字形，它只有和具体语言相结合，才有字音和字义。在这个意义上，字形与字音相比，显然具有更重要的地位。

■ 第三，文字反映了一种国际文化现象，即一种文字有可能用于记录多种语言。汉字即一典型。

■ 第四，有鉴别性特征(distinctive)(或"示差特性" - with distinctive features) 一词，较之一般的 "different"，有更严格的含义。作为一个集合，一种文字必须在本质上和数量上都具备有别于另一种文字的特征，才称得上 "distinctive"。

1.2 我们的观点

根据这一定义和大量的事实(请参阅 WG2 N513 和 1989.10 北京特别会议文件)，我们可以再次引出我们的观点：

■ **中日朝汉字，源于同一文字、属于同一文字。** 任何一国(地区)所使用的汉字，都不过是汉 script 的子集，一个国家或地区的编码字符集，只是反映这个子集的子集。它们之中的任何一个，都称不上是独立的 script。因此，在第2个DP 中，Japanese Kanji script，Hanzi script 等写法，都是与 script 定义相悖的。这个道理，就如同有了 Latin script 就不能再写 English script，或有了 Cyrillic script，就不可单称 Russian script 一样。再如，在南北朝鲜、中国、北美，Hangul 在使用中也都有些差异，但这种差异不能将 Hangul 分化为几种不同的文字。

■ 中日朝汉字之所以属同一文字，不仅是由于其编码字符集在**数量上**的大面积覆盖（依规则的不同，选材的不同，覆盖率为 60～90%），而且还由于这些文字在**本质上**的同一性：

● 中日朝汉字的基本**构形成分**（字根/部首，root/ radical）**基本相同**。

● 中日朝汉字的**结构方式相同**。日本藤堂明保先生在《汉字及其文化圈》中将汉字造字法归结为四种：象形、指事、会意、形声，而转注和假借则是派生出来的。即使在汉字发源地之外造出的汉字也符合上述结构方式，如"烟"和"畓"。

● 构字间架的**方块性**

● **表意特征及表意的一致性。** 中日朝汉字尽管经历了、并还继续经历随着时间空间的"语义演变"（semantic drift），同形汉字表达的语义仍具有很强的一致性（详见 3.3）。其基本构形成分（字根/部首，root/radical）所表示的**类属**义尤其如此。

此处指出这些同一性，并不意味着抹杀 CJK 汉字的差别。而只是说，这些差别远未达到使 CJK 汉字分化为独立文字体系的程度。

■ 字符集的**完备性**是一个字符集是否如实反映了一个 script 的重要尺度。尽管对表意文字，实现庞大的、无定界的"完备集"编码，有技术上的困难，但为了反映整体的特征，诸子集的代表性和典型性是至关重要的。至少，作为一个国际标准，常用的中文汉字（包括简、繁）、日文汉字、朝鲜汉字，其中任何一个部分都是不可或缺的。

2 编码的本质

为图形字符编码，是一个从真实世界向计算机世界（编码空间）映射与抽象的过程。编码，只是建立字符的另一种存在**形式**，而并不改变字符的**内容**。

2.1 联系与指针

真实世界的一个字（script 的一元素），在编码空间里，仅剩下两件东西，一是它的**代码表示**，二是它所具有的**图形符号**（graphic symbol）。

■前者(代码表示)无非是一个二进制的数码(pattern of binary digits)，用软件的术语，它是**不可视的指针**(pointer)，一方面指向它的图形表示，一方面指向真实世界该元素的丰富内涵。也就是说，我们借助代码表示建立了真实世界的字与代码空间的码位之间的**联系**（link）；

■ 后者（图形符号）是字符的**可视表示**（visual representation）*，它与代码表示密切相关，实现语义学上所谓的**"能指"**功能（the signifying function），而**"所指"**的内容（signified content），则是约定俗成的，是由真实世界的具体环境决定的。

二者都是**形式**，与真实世界的内容有密切联系，但它们本身不是**内容**。

* 注：这是 SC2/WG2 的定义，按 SC18/WG8 的观点，相对于具体的字形，它是一种广义的字形，一旦码位确定，其结构轮廓、组成部分也就大致确定了，但它还不是具体的字形或 image，它是独立于字体、独立于字号、独立于设计的 glyph（参见附录：Character/Coded character/Glyph/Font image 关系）。此处，**可视与否的差别并不影响我们的讨论**。

2.2 命名

除此之外，为了更确切描述和说明一个字符以避免歧义，在编码字符集中还往往为字符命名（Naming）。

需要特别强调指出：

■ Naming 目前只限于非表意字符的范围。对表意文字，命名的必要性和方法都是大可商榷的问题。事实上，无论哪一个国家或地区，都还没有对 Han Character 命名。

■ 尽管 Naming 有时涉及了字符的意义和应用，但那仅仅是对某个编码的进一步说明与补充，既不是对**各种用途**的总结，也不是对"所指"的**全面定义**。

■ Naming 所涉及的意义（meaning），就绝大多数情况来说，都是denomination下的"意义"，而不是语义(semantic meaning)和具体的 contextual meaning。

总之，命名是编码的一种**辅助手段**；命名并非定义字符的**内容**。

2.3 字义何在？

编码本身既没有责任、也没有能力去**最终地**确定每个编码字符的意义。

对给定的字符"－"，我们不知它究竟是破折号还是减号；

对给定的字符"．"，我们不能确定它是句号还是小数点；

这不是编码引起的歧义，而是真实世界的真实写照。代码本身，只是指针，指向真实世界的真实情况。在真实世界里，一个字符是多义的，在计算机世界里也必然是多义的。编码字符，在许多情况下，只能代表一种"类义"，即一种或多种**可能的**意义。

具体字义如何？这不是编码技术的范畴，它只能由具体的语言环境（context）决定，是一个**理解**的过程。然而，ISO DP10646 的引言指出："**本编码字符集允许用一切语言书写的文件进行交换，但它并不提供对各种语言本身的理解**"。这是切合实际的。

2.4 字义字音的多元性与相对确定性

表音文字的形与音结合得较好，但形与义没有直接的联系。

表意文字的形与音结合得较差（当然不等于没有结合，如形声字），但形与义的结合较好。 作为语素文字（morpheme script），汉字能够据形辨义。 孤立的汉字有可能(但不一定)表达一定的物体、行动、思想、概念。

但上述的 "结合"，无论是表音文字的"形音结合"，还是表意文字的"形义结合"，都是相对的，它们都还没有达到语音和语义的程度。

正象我们无法确定孤立的 a 读作 [ə],[ei] 还是[æ]一样； 我们也无法确定孤立的一个汉字 "火"，究竟意味着 fire，还是 anger，或是其他的意思（中医的内火- internal heat，firearms，top urgent 等等）。

我们要说明的是， 字义字音的多元性，是真实世界固有的现象，同时，它们又存在于同一字形的统一体里， 这也是事实。 字符编码，应该反映这样的现实，反映的方法， 即是 "Let them have as many meanings/pronounciations as they have in the real world"。

下面将对表意文字作进一步的说明。

■ 汉字的多义性

汉字字义的多元性，表现在同义、多义和反义现象上。 一字多义，是个突出的问题。

汉字原本是一字一义的。但几千年来，发生了从一字一义到一字多义的演变。即使在一个国度里， 一字多义已变成了司空见惯的普遍现象。 例如，《说文解字》 中，有928个当今的常用字(中华人民共和国教育部 1952年颁布)，在1800年前， 其中有二个意义以上的只有22个字，其它都是一字一义。发展到今天，其中一个义项的只有72个字， 二至三个义项的却有 387 个字， 四至五个义项的字有 295 个字，六个义项以上的有 174 个字，有少数字的义项甚至多达 15 个左右。(据中国人民大学语言文字研究所提供的资料)。 一个简单的"行"字，(xing)，单在中国境内，就有从 walk 到 OK 的十几种意义。

显然，在汉字编码设计时，只能参照，而不能依据字义去进行编码。所谓在编码时 "参照字义"，体现在

● 控制同义的变体字（异体字）的数量（如 GB 2312 所为）

● 将同义的变体字编组在一起，或用交叉索引将它们联系在一起（如 JIS x 0208 所为）

● 将极个别的同形字符分别编码， 在 Naming 时加以注解， 如 μ/μ 和 Ω/Ω 分别注以 "MICRO SIGN"/"GREEK SMALL LETTER MU" 和 "OHM"/"GREEK CAPITAL LETTER OMEGA"。应强调指出，这是极其个别的情况，若把它夸大为普遍的原则并用于汉字编码，将会造成万千的码位的浪费。

■ 汉字的同音与多音现象

同音:汉字里同音字很多。在中国,YI 音有177个字, JI 音有163个字、YU 音有139 个字、 LI 音有133个字、XI 音有130个字、 ZHI 音有128个字、 JIAN 音有119个字、 FU 音有118个字、 LU 音有115个字、QI 音和 WEI 音各有111个字、SHI音有109个字, 还有218个音由20字起到94字不等。 以上字数加起来,一共有7901个字,是现代汉语词典一万零一些字中的百分之七十九。(安子介:《汉字的再认识》)

一字多音也是多见的。"行" 在 "自行车" 中读作 xing, 在 "银行" 中读作 hang。 但在广东话中, "自行车" 中的 "行" 却读作 hang。 方言如此, 异国语言便可想而知了。

因此, 读音也不能作为编码的依据。 **同音字可以用字形来区别; 而多音字则可靠词语来最后定音。** 在这里, 我们再次看到了**字形的威力和语境的重要**。

在这种情况下, 我们进行字符编码时, 字形客观上成为了最重要的因素。 在进行 CJK 汉字混合编码时, 从字形的异同, 决定代码的赋值, 成了方便的途径。因为, 尽管一个编码字符的内容(音、义) 和图形符号(广义字形)都是相对的, 但广义字形毕竟比代码表示的内涵(字义、字音) **更具确定性**, 更易表示。

3 分析一种误解

有一种担心, 或一种指责, 认为以字形为主将 CJK 汉字统一编码会抹杀字义、字音的属性。

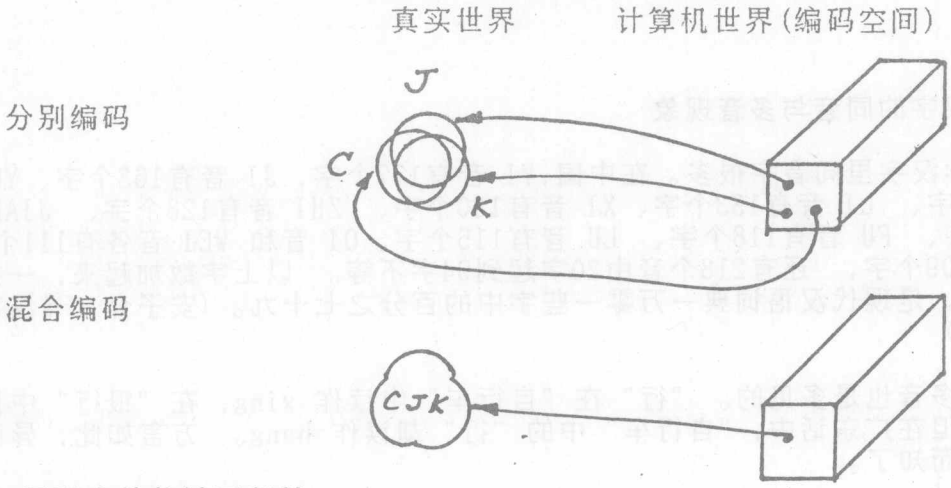
稍加分析便会发现, 这种担心是没有必要的, 作为指责也是没有根据的:

3.1 字义字音不应是字符编码的依据

同形同码并不意味着同音同义, 不同语言里的同形异义现象, 无论在表音文字里, 还是在表意文字里, 都是屡见不鲜的。 如 don, 英语为"装饰", 西班牙语为"先生", 法语则为"礼物"; 再如中文的"汽车", 在日文和朝文中则为"火车"之意。 甚至在一种语言里, 由于国度的不同, 也有不同的含义, 比如"football" 在英国为足球, 在美国则为橄榄球。 这样的例子还可以列举很多。 然而它们所引起的歧义, 只要从单字进入词、语和 cotext 以及具体的情景, 问题便迎刃而解了。 因此, 没有人会根据 football 在英美的意义不同而建议将英文字符和美式英文字符分别编码。

3.2 字的信息并不因混合编码而丢失

如前所述, 编码的本质是设置指向真实世界的指针而未对编码字符的属性、内涵加以任何限制。 CJK 汉字混合编码并非用任何一国的字义去取代其他国家同形字的字义。 CJK 混合编码无非是**扩大了真实世界所指的范围**。原来的三根指针, 指向三组可能的含义, 现在变成了一根指针, 指向了**三组含义的并集**。真实世界的含义发生了变化, 编码字符的含义也随之变化, 这便是代码指针的灵活性, 恰好反映了汉字适用于多种语言、多种用途的能力。



3.3 同形字的统计同义性

CJK 混合编码是否会把意义不同的大批汉字"混为一谈"， 赋以同码呢？

中国专家进行了认真的分析研究， 以中日形同形近的汉字和词为例，得出了如下结论：

■ 在字一级(含单字词)，尽管我们可以找到同形异义的典型,如日文的"芸"(蕓之简体)与中文的"芸"，日文的"浜"(濱之简化) 与中文的"浜"(小河之意)，但这只是屈指可数(万分之一二)的个例。

经过对1936个日中常用汉字的字义对比，以及两次分别抽样100个非常用汉字的分析，得出如下数据：

	常用字 (1936字)	非常用字 (200字)
① 基本义 、常用义相同	89%	95%
② 基本相同、稍有不同	5.6%	3.5%
③ 基本不同、但意义相通	2.7%	1.5%
④ 完全不同	2.7%	0

因此，我们可以得出统计性结论：**中日常用汉字 90% 以上字义相同或基本相同。 对非常用字，比例更高。**

■ 在词一级 (不含单字词)

① 日本早稻田大学语言教育研究所日本语科曾对日本的初、中级教科书中的汉字词汇进行过整理， 结果发现，这些教科书中使用的汉字词竟有三分之二与中国使用的词同形同义，或形同而意思接近(日本文化厅《中国语と对应する汉语》)。

② 中国人民大学对外汉语教学中心的中国学者对2023个形同形近的中日汉字词汇进行对比，发现词义相同＋词义基本相同的比例达 89% (58＋31%)。

这种宏观上的同形即同义的**统计规律**，也为 CJK Unification 提供了客观依据。 换言之，我们所建议的 HCC，并非 "purely shape-based"，而是**由形及义**的，这是由 Han Character 的特性所决定的。因为汉字的形与义有**内在的有机联系**。

4 编码字符怎样反映使用者的文化、历史和习俗？

日本的国家立场 (WG2 N518) 强调，图形符号是**纯粹**的形，而编码字符则反映使用者的文化、历史和习俗。

这种提法是不全面的。实际上这仍然是字形与字义的关系问题。 字形与字义的关系， 即是形式与内容的关系，世间上没有无内容的形式，也没有无形式的内容。把形与义截然分离是不正确的。 对表意文字，更是大谬。

我们的看法是：

■ 第一，编码字符集中的某些字符确实反映了使用者的文化背景，如日本的"和字"/ "国字"，中日不同风格的简体字；但这毕竟是很有限的，**并不是所有的**，甚至**不是大多数**字符都反映这种背景。

■ 第二，这种文化背景的体现，只能通过两个途径：一是**遣词造句** (字符的组合搭配)，日文中有大量的汉字词语，在中文中没有对偶；中文中也有大量的汉字词语在日文中也没有对应物，这些都是语言民族性的反映。 但与字的编码毫不相干。 二是通过其外现的**图形符号，具体的字形**表现。 **显然，这都超出了字符编码的范围**。

尽管如此，我们仍认为，字符使用者的文化、历史和习俗是重要的。 因此，

■ 中国的 WG2 N508 提案 (认同规则) 已经建议应要在 HCC 中**保持现有的差别**， 如， 由于不同简化方式形成的变体字， 如 团/団，对/対，图/図 等等，并赋予它们以不同的代码。

■ 同时，我们还在 WG2 N480 提案中，强调了应允许不同的国家和地区使用不同的字形 (FONT image)。 比如，日本奉明朝体 (Mincho) 为正体，而中国则以宋体为 "报版字" (最流行)，二者的细微差别实际上局限在笔画的风格上，与其称为民族文化之差,不如说是书法之别， 如文/文，前者是中国的规范写法 (斜点)， 后者是日本的正体(竖点)， 但日本的楷书也采用前一种形式(文)，中国台湾又较多地采用竖点。 无论如何，这是用户在 Glyph rendering 层次上的选择。CJK Unification 对此毫无约束。

字/字的组合/文句，和字符/ Glyph/ Font image 一样， 是不同层次的两串概念。**字和字符**， 均是处于最抽象的层次，也是民族性最弱的层次， 集中体现了共性， 而不是差别。 不难想见， 民族性和文化差别或许更容易在 "Text" 和 "Font" 的层次上反映出来。 这个概念可以用下图表示。

字符 / 图符 / Font image

字 / 字的组合 / 文句

抽象－－－－－－－→ 具体

民族性

弱 －－－－－－－→ 强

国际性

强 －－－－－－－→ 弱

正如我们多次指出的， 中日朝汉字的现状是 "大同小异"， 汉字编码和汉 Glyph ID 赋值的策略则应是 "求大同存小异"。 在一定的意义上，在字符一级体现较多的"大同"， 在图符一级和具体的 image 上，则体现更多的 "小异"，是合逻辑的。

5 HCC 的实现基础

CJK Unification 的难点之一，是如何建立一个公认的字符认同规则。 实际上，这一规则正在逐渐形成：

■ 日本 JIS x 0208 的3.4节， 关于 Kanji Variants 的处理(Handling)已经提出了许多重要而正确的准则。 特别是解说表6，对于分清汉字与汉 Glyph，减少重复编码，是大有助益的。

■ 中国的提案 (WG2 N508, WG2 N508 Rev.)， 正式提出了处理字形相近的中日朝汉字的规则。 将这个规则与 JIS x 0208 的表六相对照，我们会发现二者的思路是十分相近的。

■ 美国 Unicode 中的汉字字符的区分与归并，与上述规则是并行不悖的。最近， 中国专家在对 Unicode 的 10496 个 Cell 中的汉字进行了 Review之后认为其认同规则的运用是合理的， 除了个别的错误 (0.25%)外，整体上是完全可以肯定的。 换言之， Unicode 中的汉字 Collection 可以作为C/J/K 各国家委员会讨论的基础。

■ 日本 SC18/WG8 提出的 "Proposal for Variant Discriminating Scheme" (AFII/89-50)，尽管是区别 glyph/variant 的，但对字符的认同也有很大的参考价值。

借助上述的规则或经验判据， 我们可以大大地削减字符编码中的冗余，更加合理地安排 ISO 10646 中的基本多文种平面的空间。

最近的工作进一步证实了这一点：

■ 根据 Unification 的原则，GB 2312-80 所对应的2200多个繁体汉字中，至少有1800个无需新的编码即可进入基本多文种平面。 (因与 JIS, KS 汉字相同)

■ 日本的 JIS Ⅲ (draft) 的5800个汉字中，至少有 4000个无需新的编码(因与 CNS 非常用字相同)。

因此，可以肯定地说，按照 HCC 方案，在基本多文种平面，不但可以安排原有的 GB 2312, JIS x 0208 和 KS C 5601 中的全部汉字，还可以容纳

■ GB1 （2200个繁体汉字） 和 GB3 的绝大部分汉字

■ CNS （13000个汉字）

■ JIS Ⅲ （5800个汉字）

这是符合各国人民的共同利益的。

6 HCC 的应用特点

基于 HCC 的应用软件，较之 C/J/K 分别编码的软件，有一系列的特点。通过下面分析，我们可以看出其中的利弊。总的情况是，软件简化了，使用方便了；但有的情况则相反。需要强调的是，ISO 10646 是作为多文种的国际标准提供的，其优点，只是在相对复杂的应用环境才显现其优越性；对于最简单的情况，如单文种或 Bilingual，则未必方便。这一点在 10646 的引言中已有说明：

"The primary value in using this form of coding is that software may be much sumplified compared with the case where other types of coded character sets are used **(except the simplest cases)**"

以下，■ 表示HCC 之优点 □表示缺点 ◪ 表示情况复杂，难以定论。

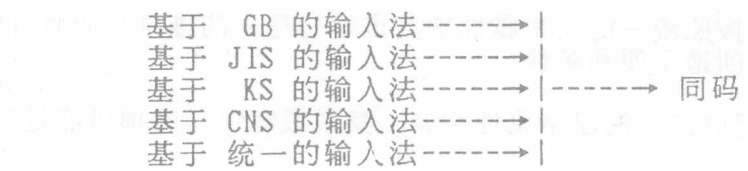
6.1 字量 (Characters' cardinality)

■ 以两字节、定长方式可提供的汉字大幅度增加。因此，在应用中，需要用户造字的概率将大大下降。对绝大多数的应用，3 字节和 Single shift 都可以避免，而且可以减少 Locking shift 的使用。

6.2 输入

■ 除了 Language dependent 的输入方法外，还可以开发统一的 Language independent 输入法 (如以 radical/stroke 为基础的输入法)。

■ 录入员或 OCR 读入识别系统无需考虑待输入字的语言背景，便可用其熟悉的方法将这个字输入进计算机系统，而不致发生错码。



□ 对于统一的输入方法

- 不可采用按音输入的方式
- 使用较大的输入码/内码对照表
- 待选字可能较多

6.3 输出

■ 如进而实现 Unifont，则有可能大大减少 Font 开发的投资，促进字形的规范化，实现 Font 共享。

□ HCC 分级必须优化，因为随着可供字量的增加，字库中的冗余部分也相应增加了。

□ 在**多文种混用**和**要求严格**的场合，必须使用 Multifont（即使是同一字体），以反映各国的差别(如与/与)。在这种情况下，必须要有指明 font、调用 Image 的标准机制（而非 Language ID)。实际上，目前已有多种方式支持这种机制：

Font designation is supported in a variety of ways: At the system interface or toolbox level (e.g. Macintosh, Presentation Manager, NeXT), document interchange standards (MacWrite on the Macintosh, Microsoft Rich Text Format for Macintosh-PC document conversion, Office Document Architecture [ODA], Standard Graphics Markup Language [SGML], CALs (document standards used by the US government), and page description languages (Postscript, Interpress).

6.4 代码转换

□ 必须使用大对照表，而不能使用简单的 Logic 或公式。不适于在容量小、速度低的计算机上进行转换。

□ 从 10646 的 HCC Code 到 2022 的 C/J/K Code 的转换，必须指明 HCC Code 的 Language ID，否则会错位。(6.2 的优点，在这里是缺点)。

```
                +   C info  →    GB Code
HCC Code        +   J info  →   JIS Code   异码
                +   K info  →    KS Code
```

6.5 排序

– 综合排序 Sophisticated Sorting

■ 可以按照统一的汉字属性字典进行有意义的排序；而且 HCC 为建立这种统一的属性表创造了便利条件。

□ 必须建立统一的汉字属性字典（其中最有意义的项可能是字根，笔画数，笔形等等）

– 简单排序 Binary Sorting

◪ 可以有意义，取决于 HCC Level Ⅰ/Ⅱ/Ⅲ 的安排。而在 C/J/K 分别编码的场合，对 CJK 混合的 item 进行 Binary 排序，则是完全无意义的。

6.6 机助翻译(MAT)

◪ 需要在篇章或段落一级提供某种 Language Information(无需在字符一级

◪ 可对篇章或段落进行抽样预扫描以判断文种，其中有 Kana 者为日文，有 Hangul 者为韩文。

6.7 数据库 (RDB)

◪ 为避免极个别的同形词（如“汽车”）引起的歧义和误操作，在检索、运算等操作前，应进行相应的 "质量控制" (authority control) 或“组合条件”查询。

实际上，在数据库中，在某些关键项中内容相同的情况(如姓名相同) 是屡见不鲜的。实践中有多种办法排除歧义而无需专为这一项的内容增加标识信息。

■ 检索：由于C/J/K 分别编码，在检索时，可能会发生**一字多码**的困扰。较之 HCC 中的同形异义词的麻烦，一字多码的困扰可能更为严重，概率更高。相形之下，HCC 的**一字一码**，“内外一致”,则显现了优越性。

7 结语

ISO 10646 的发展，目前面临着重要的抉择。

汉字在基本多文种平面，占据了 55% 以上的 Cells，这个空间使用是否得当直接会对整个标准的实用性产生重大影响。

在 2nd DP 中 CJK 汉字的分离,不仅浪费了宝贵的多文种平面、排斥了有重要实用价值的中国繁体汉字，而且直接违背了 ISO 10646 按 script、而不是按 Language /Country 编码的原则；而由此引起的一字多码，对未来多文种应用环境，将带来严重的困扰。因此，**中国 National Body 不同意目前第 2个 DP 10646 的对汉字字符的安排。**

中国专家经过认真的、反复的分析研究，认为：尽管各种解决方案都有利弊，但 HCC 方案较为合理。中国的出发点，不仅是中国的利益，而且是整个汉字文化区的共同利益。我们愿意与各国专家共同协商，寻求公正的解决方案。

我们认识到，CJK Unification 的实现，还是一件很紧迫的任务。由于数据的积累。软件的开发日新月异，因此，CJK Unification 的实现，明天比今天要付出更大的代价。**与其晚实现，不如早实现。**中国 National Body 愿为此作出牺牲和贡献。

ANNEX

Character/ Coded Char/ Glyph/ Font image

字 / 编码字符 / 图符 / 具体字形 的关系

