# 汉字三码演进脉络与现存问题探析

王　迈

【提　要】内码、外码与形码是计算机表达汉字的三个接口。汉字是词符词素文字，具有字形复杂，区别度高、语义相关等特征。汉字的信息熵达到 9.65 比特，较印欧系音素文字的 4 比特高出约 32 倍，两者可以类比为不同数制的信息熵差异。依据香农信道编码定理，高熵值大集合的汉字需要双字节以上的编码空间。UCS 将多语种文字归入单一字符集，但未能在编码地图中反映字形、字义及历史文化的跨语种关联，而国际汉字单位的构想有助于实现体系化的编码布局。

【关键词】汉字　信息熵　语素文字　字符编码

DOI:10.14014/j.cnki.cn11-2597/g2.2025.09.039

字符是文字和符号的总称，由一组字符组成的集合称为字符集，如汉字字符集、拉丁字符集、数学字符集等。诺依曼计算机是二进制的机器，字符须改写成二进制码的形式，才能为计算机记忆和处理，这是字符编码的实质。

为了表达一个字符，计算机应该提供三方面的信息：其一是能够唯一鉴别字符身份的内码，也称机器码，它是计算机和字符的底层接口；其二是人们向计算机查询或选取字符所需的外码，也称录入码，它是字符与使用者的接口；其三是计算机表达字符视觉信息的形码，也称字形码，它是人眼与计算机显示设备的接口。

汉字是迄今世界上唯一仍在使用且没有断层的语素文字，它所对应的语言单位达到了概念级别，具有字形复杂、高信息熵、语义直接相关等特点，这使得汉字实现计算机存储、录入和输出相较于音素文字走过了更加曲折的历程。我们将结合汉字与计算机原理回顾这一历程，分析汉字三码的演进规律，并探讨当下仍未解决的一些问题。本文涉及的是汉字内码部分，外码与形码将另文探讨。

## 一、字符内码编码原则及基本编码方案

字符编码需要遵循两个原则：1. 香农信道编码定理。即每个字符获得的二进制码长不能小于该字符的信息熵。2. 唯一性原则。即在一个确定的编码空间内，每一个二进制码字获得的字符解释必须唯一。这两个原则保证了一个字符能够获得计算机充足并且专一的支持。

此外，好的编码方案还应具备如下特点：1. 系统性。编码地图能够反映文字系统的构成，其布局有规可循，具有模块化特征。2. 前向兼容性。保证历史文档的可持续利用。3. 后向可扩展性。为字符集的扩充预留空间。4. 灵活性。在同一编码方案下，可以实现灵活多样的转换格式，以适应不同场景的需求（详见后文 Unicode 下的 UTF8、UTF16 等）。5. 稳健性。也称鲁棒性，当传输错误、数据损坏等极端情况发生时，字符集能够把影响限制在较小的范围，减少对系统整体的破坏。

由于人类语言文字的复杂多样，加之早期计算机存储能力相对薄弱，以及各国机构的各自为政，文字编码的发展并非一帆风顺。

1967 年，美国国家标准学会（ANSI）发布了信息互换标准代码（ASCII），后被国际标准化组织（ISO）定为国际标准 ISO.646。ASCII 是针对英语国家的单字节编码方案，也是后续各类文字编码方案的源头，它使用 7 位二进制数表示一个字符，包括字母、数字、标点、图形、控制符共计 128 个，字节最高位保留为奇偶校验位。例如，单词 Cat 的 ASCII 码为 0x43 0x61 0x74。

ASCII 可以满足英语国家的需要，但是其他使用拉丁字母的语言多包含一些变音符号，如德语的

ä、ö、ü，法语的 à、é、ï 等，另有西里尔、希腊、阿拉伯，希伯来等差异更大的音素文字，ASCII 都未包含在内。由此 ISO 与国际电工协会(IEC)联合制定了 ISO/IEC.8859 系列标准，利用 ASCII 高位预留的 128 个字符空间作为补充，组织了共计 15 个语种的音素文字地图，这是 ASCII 的一次重要扩容，却并未涉及汉字等大字符集的解决方案。

二、汉字的信息熵与双字节编码

汉字在信息处理领域的独特性需要引入熵的概念才能准确界定。熵(Entropy)原是热力学术语，用以度量系统微观粒子的无序程度，自香农(Shannon，1951)将其引入信息论以来，熵逐渐成为信息量大小的量度。假设一个 32 人的班级选举班长，在每个人当选概率相同的前提下，可以认为选举结果消除了 32 种存在的可能性，那么选举结果的信息量便是 $\log_2 32 = 5$bit。同样，如果在 8 人小组中选举组长，选举结果只能消除 8 种可能性，这时的信息量就只有 $\log_2 8 = 3$bit 了。我们使用语言文字表达思想，每一个语言文字单位的出现，都会降低表述内容的不确定性，就如同选举结果一样具有信息量价值。借助信息熵工具，我们可以对语言文字单位的信息量进行精确测量。

表 1　汉字与音素文字的信息熵

| 语种 | 汉语 | 俄语 | 德语 | 英语 | 西班牙语 | 意大利语 | 法语 |
|---|---|---|---|---|---|---|---|
| 文字的熵(bit) | 9.65 | 4.35 | 4.10 | 4.03 | 4.01 | 4.00 | 3.98 |

汉字是高信息熵的文字，单个字符包含的语义信息量达到了概念层级。冯志伟(1994)使用容量极限法测得汉字熵值约为 9.65bit，而印欧系诸文字的熵值只有 4bit 左右，两者的信息差达到了惊人的 $2^5 \approx 32$ 倍。从文字类型学的角度分析，汉字属于语素文字，其表达的语言单位是音义结合的语素(中古及上古时期表达的则是词符)，其指称的对象达到了概念级别，如“山”“水”“躬”“耕”；而印欧系诸文字属于音素文字，其对应的语言单位是音素，音素只表示一个发音动作(元辅音等)，不与语义直接相关。因而，我们的直观感受是，一个汉字足以表达一个概念，而拉丁字符集通常需要多个字母的组合才能达到概念层级。

以概念“火”为例，汉语需要一个汉字表达，英语则需要“fire”四个字母组合起来加以表达。我们可以将其类比为十进制与二进制的差异：数量“9”用十进制表达只需要一位，而用二进制表达却需要“1001”四位。概括地说，十进制相对于二进制的特点是：每一位的区别度高、信息量大(有 0-9 十种变化)；总体占用的位数少。这同样是汉字之于拉丁母的特点，也即语素文字之于音素文字的特点。

根据香农信道编码定理(Shannon's Channel Coding Theorem)，一个汉字的码长不应小于它的信息熵 9.65bit，但这是未计算冗余信息的理想状态，实际应用需要更大的空间。计算机通常以 8 个二进制位(Octet)作为存储单位，采用 2 个 Octet(16bit)表示一个汉字，可以编码 $2^{16} = 65536$ 个字符，这就基本满足了汉字编码需求。1980 年，中国国家标准总局发布了《信息交换用汉字编码字符集——基本集》，简称 GB2312 或国标码。它采用双字节编码一个字符，为了区别 ASCII，每个字节的最高位一律置“1”，这很大程度上避免了中英混排文档的乱码现象。GB2312 共收录了 7445 个字符，其中汉字 6763 个(全部为简体)，包括最基本的一、二级字库，是迄今几乎所有简体中文系统都支持的字符集。

三、多语种通用字符集与 UCS

与 GB2312 同时，各语言地区(特别是大字符集地区)也都先后研制了适合自己的字符编码标准，如通行于台、港、澳地区的繁体汉字 Big5 码，韩国语的 Wansung 码，日本语的 JIS-X-0208 码等。随着编码标准的增加，当采用不同字符编码的文档相互传递时，计算机经常因判断错误而不能正确切换解码标准，引起乱码现象；同时，计算机为了识别编码种类，必须在存储器中额外预存多种代码表，从而增加了系统复杂度。如果存在一种涵盖世界上所有文字系统的通用字符集，就可以在一个系统甚至文档中处理多种语言，这无疑是一种理想的解决方案。

1993 年发布的 ISO.10646 定义了这种通用字符集(Universal Multiple-Octet Coded Character Set，

UCS 或 Unicode)，它是世界上所有已知字符集的超集，并为将来可能出现的字符预留了足够的空间。UCS 规定了一个 4 字节(31 位)的编码空间，并按字节由高至低分别定义为 Group、Plane、Row 和 Cell 四个级别，理论上可以存储 231 个字符，如此庞大的空间几乎是不可能被填满的，迄今被使用的也只有 0x0000～0xFFFD 共 65534 个码位，相当于第一个 Plane，称为基本多文种平面(Basic Multilingual Plane，BMP)。BMP 是 UCS 的一个子集，Unicode 通常就是指对这个子集的实现，但仅仅是这一个子集，就几乎涵盖了世界上所有文字的主要字符。其中，大部分拼音文字(包括 ISO.8859 中的大部分字符和日语假名等)位于 A-Zone 区；大部分汉字(包括日、韩、越等国使用的)位于 CJK 区；增补的扩展汉字集位于 CJK ExtensionA 和 0xFFFF 后的部分空间；韩语文字位于 Hangul 区；YI 区包含彝、八思巴、爪哇等一些使用人口不多的文字；S-Zone 为 UTF16 转换方案专用区；EUDC 为保留私用区，用户可根据需要增加自创文字；R-Zone 为限制使用区，包含一些特殊字符和兼容字符等。

UCS 给每个字符分配了唯一的代码，使得现存各类字符集中的任何字符都可以在 UCS 中找到映射。例如，汉语“上外”的 Unicode 是 U+4E0A U+5916；日语 “はな”是 U+306F U+306A；韩语“학교”是 U+D559 U+AD50；其他文字类同。这样，UCS 就实现了世界文字符号的大一统，可以在同一个文档的同一个字符集中处理多国语言。

UCS 为我们提供了字符编码方案，但并未规定具体的实施细则。在英语国家，传输纯粹的英语字符只需要 8 位，如果仍旧用 2 个字节编码，高字节就得全部置 0，造成空间浪费，由此在 UCS 基础上制订了 UTF-8(UCS Transformation Format 8)，这是一种传输标准，是对 UCS 编码的具体实现。同样，我们可以根据实际需要制定多种实施标准，除了 UTF-8，常见的还有 UTF-7/UTF-16/UTF-32 等。以 UTF-8 为例，它采用变长技术，用 1 个字节传输 UCS 中的 ASCII 字符；用 3 个字节传输 UCS 的 2 字节字符(例如汉字)。仍以“上外”为例，U+4E0A U+5916 的二进制串分别嵌入 UTF-8 掩码，得到的二进制串用表示为十六进制，即为 UTF-8 的编码%E4%B8%8A 和%E5%A4%96。同理，如果是拉丁字母，只需嵌入模板第一行，采用一个字节就可以完成传输。

Windows 系统的内核已经使用 Unicode 编码，这就做到了从底层支持世界上所有语言共用一个平台。但是，现存的大量文档和软件仍是传统字符集下编写的，为了向前保持兼容，Windows 提供了代码页(Code Page)机制，这是一种折中方案，用户根据所处地区设定系统的缺省代码页，系统则根据代码页将软件或文档中的 ANSI 字符自动映射为 Unicode 字符。仍以“上外”为例，其在地方软件中的内码为 0xC9CF 0xCDE2，Unicode 编码为 U+4E0A U+5916，代码页其实就是这两种编码的映射关系表。常见的代码页有：CP936 对应 GBK；CP950 对应 Big5；CP932 对应日语 Shift-JIS；等等。随着 Unicode 的普及，代码页的使用场景将越来越少。

四、UCS 的优化空间与未来展望

UCS 将各语种文字归入单一字符集，这是了不起的成就，但它并非完美无缺。我们从两个方面探讨 UCS 尚待优化的问题。

首先，UCS 不兼容简体汉字基本集 GB2312，既有软件和文档必须依赖代码页映射，对执行效率和稳定性存在潜在影响；同时，随着考古事业的发展以及文献古籍的解读，陆续有一些生僻字形纳入到汉字库中，这些字形没有得到 UCS 的很好支持。

为此，中国信息技术标准化委员会发布了 GBK(国际标准扩展码)，它向前兼容 GB2312，收录字符达到 21886 个。GBK 基本涵盖了 Unicode 中的汉字集，除与 GB2312 保持兼容的 6763 个汉字，其余排列基本是 Unicode 编码的平移，从而保证了两个字符集间的映射关系。GBK 在古籍数字化处理中发挥了重要作用，但同时必须承认，受 Unicode 的影响，GBK 还是在收字和编码方案上做出了很大妥协。

2000 年和 2005 年，又有两个大字符集标准 GB18030-2000 和 GB18030-2005 发布。与 GBK 不同，GB18030 采用多字节编码，可以用 1、2 或 4 个字节表示一个字符：其单字节部分与 Unicode 兼容；双

字节部分与 GBK 兼容，并适当调整了一些与 Unicode 的映射；四字节部分映射了单双字节没有映射到的 39420 个 Unicode 的 BMP 码位，以及 16 个辅助平面（Plane）共计 $2^{16}*$ 16 个码位。如此庞大的字符空间几乎就是另一个版本的 UCS，或者可以看成与 GBK 兼容的 Unicode 本地化版本。

另一方面，我们应该看到，随着计算机硬件和通信技术的发展，存储空间早已不再是字符编码的掣肘因素，然而，如何更好地在编码地图中科学反映语言文字特点和民族文化内涵，仍是需要进一步探究的课题，尤其是站在汉字文化圈整体视角加以考量。

①发 U-53D1　②發 U-767C　③髮 U-9AEE　④発 U-767A　⑤髪 U-9AEA

上述五例汉字在字源、字义、字形、字音、使用时代和使用区域等方面有着错综复杂的联系。其中②③为繁体中文汉字，②为“發展”之“發”，③为“頭髮”之“髮”；①为②③的简化汉字，简化后不再区分；④⑤为②③对应的日本汉字（须注意③和⑤不同，⑤比③少了一点）；字后附上了相应的 Unicode 编码。如果按照字源和字义，应分为①②③和①③⑤两组；如果按照字形，应该分为①、②④和③⑤三组；如果按照语言或使用地区，应该分为①、②③和④⑤三组；如果按照使用频度，则可以安排在一组。观察它们的编码，可以知道 Unicode 是按照字形归类的，即①、②④和③⑤分组法，这是一种简单易行的方案，但是未能体现汉字间的联系。

朱一星（2012）、王迈（2014）提出了国际汉字单位（ICCG）的构想，它建立在自顶向下的汉字系统观之上，将跨语种同位汉字的多维联系体现在体系化的汉字编码布局中，可以逐步解决汉字同位异码问题，促进汉字跨语种的整理和规范化。概括地说，有历史渊源关系且字形相近、字义相关的汉字，可以归并为同位汉字，这就如同氢的同位素包含氕、氘、氚一般，可以提供分类相同又互有区别的体系化信息。

## 参考文献

冯志伟　1994　《汉字的信息量大不利于中文信息处理》，《语文建设》第 3 期。

冯志伟　2024　《汉字熵值计算及其科学意义》，《北华大学学报》第 1 期。

王　迈　2014　《同位汉字的编码问题》，《数字化汉语教学》，清华大学出版社。

朱一星　2012　《如何界定汉字的理论单位》，《研究论丛（日）》第 1 期。

Shannon，C. E.　1951　*Prediction and Entropy of Printed English*. The BELL System Technical Journal.

（通信地址：200083　上海外国语大学国际文化交流学院）