超大字库中异体字形对应关系整理的设想

张再兴

（华东师范大学 中国文字研究与应用中心，上海 200062）

摘要：随着汉字字库的不断扩大，字库中所收历史异体字不断增加。这在为全面解决计算机用字问题提供了条件的同时，也给汉字输入、繁简转换、古文献全文检索等带来了新的问题。为此，必须对字库中的字形进行整理。这种整理不仅仅是字形的精简选择，更主要的是对异体字形之间的关系进行全面的清理，从而建立起异体字形之间多层面的内在联系。这种整理也包括给字形标注多种来源属性。通过这样的整理，才能真正让超大字库发挥效用。

关键词：超大字库；异体字形；字形关系整理

中图分类号: H087　　文献标示码:A

超大容量的汉字字库的建立，本应为全面解决计算机用字问题创造条件，但从目前通用的大字库的实际状况来看，距离这一目标还有很大的距离。个中原因很多，其中十分重要的一点，就是字库中的大量历史异体字未经整理系联。本文所谓的“异体字”是从广义的角度来讲的，包括意义完全相同的异体和部分意义相同的异体。由于汉字使用的时间非常长，累积到了今天，汉字的数量已十分庞大。而在这层累起来的汉字字形集合中，大量的字形只是异体字。以今日收字最全的《汉语大字典》为例，共收字 56000 左右，其中所收异体字主要是历代字书、文献中的楷书异体，约 11900 组，以平均每组包含 2 个异体字计算，包括繁简异体，共有 23800 个异体字，占到了总字数的 42％。由书体的不同而导致的异体数量更是十分惊人。据统计，甲骨文、金文等古代书体中的手写异体字形便有十几万。这样一来，“围绕着一个规范的字形，会出现一大串异写字与异构字”[1]。而古文献的计算机信息处理又要求这些不同时代的多种异体并存于一个大字库中。因此，字库越大，其中所包含的异体越多。

这样一来，就不可避免地产生一些问题。首先是众多字形的检索输入问题。GBK 全拼输入法中的大量重码字已经令人头痛，可它还只有两万多个字。但是面对几十万、上百万的异体字形和历史字形，光靠在输入法编码上下功夫是远远不够的。

其次，计算机信息处理中还会出现另外一些问题。（1）繁简转换、通用体与异体的转换无法准确地自动实现。传统的汉字繁简自动转换都是通过建立一个与简体字一一对应的繁体字库，部分不能对应的繁体则置于繁体字库的另外地方来实现的。因此，这种方法很难实现具有一对多关系的简繁字及大量其他异体字之间的转换。如 GBK 中的宝、寶、寳[2]、寚[3]、寶[4]、珤[5]六个形体，“宝”是简化后的通用形体，“寶”是繁体字形，后四个形体都是不同时代产生的异体。利用 Microsoft office2000 提供的繁简转换工具，在繁体转换为简体时，只有第二个字形能转换为简体，其余形体无法转换为通用形体。而在简体转换为繁体时，转换结果也只能是第二个形体。（2）在古文献的全文检索中，同一个字的多种异体由于分处字库的不同码位，现有系统或者无法认同为一个字，降低了查全率。或者将不应该认同的字认同，降低了查准率。如《四库全书》全文检索中的异体匹配模式，由于检索数据量的巨大，无关的数据检索结果比重相当高。

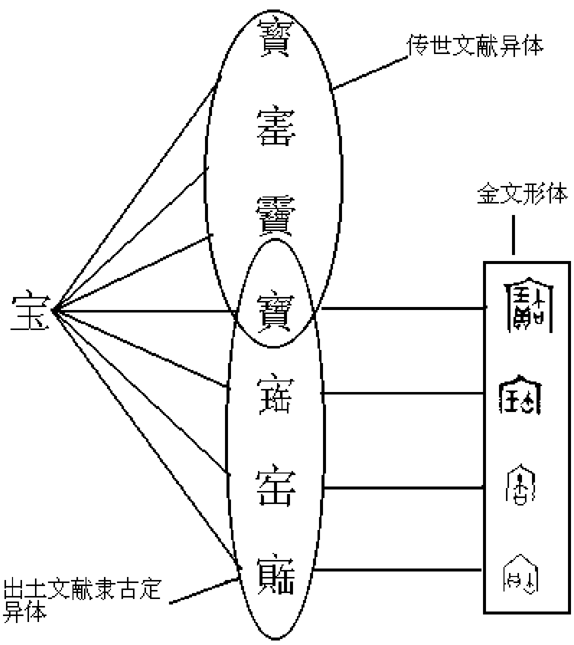
因此，要解决这些问题，让超大字库真正发挥作用，就有必要在同一个超大字库中，根据字形特点，对字形关系进行全面整理，即建立不同异体字形之间的内在联系和标注各个字形的多种来源属性。

以往谈到的字形整理，往往是指对字形的选择，例如，根据使用频度、规范性等原则，对进入字库的字形进行选择。但是，根据计算机技术对字库的要求，每一个汉字字形必须做成字库中的一个字，在字库中占有一个码位，才能在计算机中使用，否则计算机就无法调用。因此，理论上要求历代使用过的所有汉字都要在字库中有一个字形。我们不会因为某个字形不规范，在学术研究、古籍整理出版

中就不用这个字形。这样，超大字库研制中的字形整理就不应以选择典型形体或规范形体，精简字形为主要目标。因此，我们这里所指的字形整理主要是指对字库中各个字形之间特别是异体之间的关系的清理和相互对应以及标注各个字形的多种来源属性。

1．清理字库中不同形体之间的内在关系，在同一个字的多种异体之间建立起内在的关联，从而建立起各个层面的异体认同机制，使得计算机能自动认同一个字的不同异体。

现有大字库中一个字的多种异体实际上处在一种无序无关的状态。如 GBK 中的宝、寶、寳、寚、寶、珤六个形体分别位于 B1A6、8C9A、8C97、8C87、EC64、AB92 六个码位。这些字形相互之间没有建立起联系。通过在异体之间建立内在的关联，我们可以将“宝”的各种传世文献异体和出土文献异体之间以及出土文献异体与各类古文字形体等几个层面之间都建立起双向对应关系。如附图所示。



附图　“宝”的异体关系

这样一来，在繁简转换及通用形体与异体之间的转换过程中，就可以改变异体之间（包括繁简体）之间的“认同”通过不同代码页中相同码位的对应来实现的现有模式，而是在同一字库的不同码位之间通过内部联系来建立认同机制，这样，既可以突破异体之间只能一一对应的限制，还可以在上述的不同层面的异体之间建立关系。在转换时，如果一个通用形体对应于多个繁体或异体，可以列出所有的繁体或异体选项，通过程序与用户对话的方式由用户实现自主选择，也可以使系统通过上下文的内容环境进行更智能化的自动选择。如根据上下文将头发的“发”转换为“髮”，将发财的“发”转换为“發”。这样就可避免一般转换程序存在的需另外安装繁体字库和部分繁简不一一对应的字在转换过程中出现转换错误等问题。而在古籍的整理、普及出版中，还可以让系统自动为文本中的各种异体字加注通用形体。在古代文献的全文检索中，则可以通过控制是否与通用字匹配来兼顾查全率和查准率。

在字形检索输入方面，现有输入法由于基于较小型的常用字库，没有大量异体的障碍，因此通过输入法的改进可以大大减少重码字。但是对于超大字库而言，大量的异体字重码是在所难免的。因此，除了用增加码长等方法外，还可以通过建立字形之间的内在联系，根据其联系对大量异体字形进行分层面检索。这样，将传统的码表映射方式，改为底层数据库映射方式，检索出来的异体重码字可以是一张字形之间关系条理清晰的异体重码字关系表，而不是现在通用的一串线性排列的字形队列。例如，

用全拼输入法输入“bao”时，与“宝”字相关的异体字形在候选窗口中不与其他无关的字形一起呈线性排列，而是如附图所示，在一个二维平面上由通用字形“宝”统领的多层面异体重码字关系表。

另外，就广义的角度而言，许多异体字只有在部分义项上重叠。因此，对于异体字关系的梳理还应该充分照顾到不同字形之间的意义关系。只有梳理出完整清晰的意义关系才能更明确地区分异体字形之间的内在关系。

3．给字形标注多种来源属性，如：使用时代、所属书体、字形出处等。以往的汉字信息处理中，对汉字属性的关注主要集中在字形本身和计算机技术方面，如笔画数量、笔画顺序、偏旁部首、读音以及码位等。而对汉字流传的历史环境缺少关注。而汉字是历史层累的产物，超大字库的使用场合也多与历史文化相关。因此，我们应该对汉字的来源属性给予足够的关注。有了这些属性之后，在使用中可以根据需要，自动动态检索生成相应字形分库。例如，在编辑、研究某一时代的典籍时，可以根据需要调用某一时代的字形分库。也可以根据书体动态调用甲骨文、金文、小篆等字形分库。这种字形分库在使用中应该比整个超大字库要方便得多。

不过，上述整理对字库建设提出了更高的要求，不仅仅是将收集到的字形按照笔画、偏旁等外在属性排列，而是要整理字形与字形之间的复杂内部关系。这也就要求未来的字库建设更多地需要文字学家的参与。

Ideas on Organizing the Relationships among Variant Forms of Characters in Super Chinese Character Database

ZHANG Zai-xing

(Center for the Study of Chinese Characters and Their Applications, East China Normal University, Shanghai 200062,China)

Abstract: With the expanding of Chinese character database, historical variant forms of Chinese characters collected in the database are increasing. While making the complete computerization of Chinese character possible, the expansion of the database brings new problems to character input, form transition between simplified and complex, and indexing of historical documents. To solve these problems, we should organize the character database. In addition to selecting characters, this organization work will mainly focused on clarifying the relationships between the variant forms of characters, building up the multi-levels of inner joins, and marking the source property of each character. With this kind of work done, the Chinese Character database can be fully utilized.

Key words: super Chinese Character Database; Variant forms of Chinese characters; clarifying the relationships between the variant forms of characters

收稿日期: 2003-5-30

作者简介: 张再兴（1968-）男（汉族），浙江新昌人，史学博士。华东师大中国文字研究与应用中心副教授。

1 王宁.《计算机古籍字库的建立与汉字的理论研究》，载《语言文字应用》1994 年 1 期。

2《改併四声篇海》引《玉篇》字形。

3《说文・宀部》:“窻，古文寶，省贝。”

4《字汇补・雨部》:“霦，篇韵与寶同。”

5《玉篇・玉部》:“琖，《声类》云：古文寶字。”《后汉书・光武帝纪上》:“今若破敌，珍琖万倍，大功可成。”