文章编号：1003-0077(2015)04-0144-07

中文的同形异码字问题

张小衡

（香港理工大学 中文及双语学系，香港）

摘 要：同一个字符拥有不同的计算机内部代码，这意味着有两个或两个以上字形在人的眼中是同一个字，而计算机却认为是不同的字。这种“人机看法不一致”会给语言信息处理带来混乱，导致信息检索不全，统计数字不准，字词分类排序不一致等情况。该文结合 Unicode 实例专题讨论当前计算机上存在的中文同形异码字问题，包括（a）私人造字公有化所形成的同形异码字，(b) 兼容编码所形成的同形异码字，(c) 建立专门的笔画部首表而形成的同形异码字，(d) 半宽和全宽字形分别编码而造成的同形异码字等，并探讨解决问题的方法。

关键词：中文字符；同形异码；Unicode

中图分类号：TP391 文献标识码：A

Duplicate Encoding of Chinese Characters

ZHANG Xiaoheng

(Dept. of Chinese & Bilingual Studies, Hong Kong Polytechnic University, Hong Kong, China)

Abstract: A duplicate-encoded character is a character which has been assigned two or more code points in a coding system such as Unicode. When output in distinct codes, the glyphs of a duplicate-encoded character appear the same to human users, while in the computer, they are different characters. Such a human-computer inconsistency would cause confusion in language information processing, resulting in incomplete information retrieval, inaccurate statistic calculation, and inferior quality of data sorting and categorizing. This paper discusses duplicate encoding of Chinese characters in Unicode, MS Office and the WWW, including (a) duplicate encoding arising from new code assignment in the Unihan public area to characters already encoded in the private use area, (b) duplicate encoding caused by compatibility encoding, (c) duplicate encoding brought forward by building dedicated lists for CJK strokes and radicals, and (d) duplicate encoding of characters in half-width and full-width forms. Some effective solutions to the problems are also suggested.

Key words: Chinese characters, duplicate encoding, unicode

1 问题的提出

在繁体字 PDF 版《中华人民共和国香港特别行政区政府二零一一至一二年施政报告》中[1]，多处出现“劏房”① 这个词。如果您也关心香港的“劏房”问题，想找出施政报告中提到“劏房”的每一处，很可能会利用 Adobe Reader 的 Find 或 Advanced Search 命令。但是，不管您使用内地的微软拼音中文输入法（MSPY 2010）还是台湾的微软新注音输入法（New Phonetic 2010）输入繁体检索字“劏”，结果都是一样：连一个匹配都没有找到。但是，原文件中的确有“劏”字。如图 1 所示，在计算机报告找不到“劏”字的小窗口下面的原文件中就有“劏”字出现。

显然，这是一个相当严重的中文信息处理问题。

① 劏 tāng，粤语方言字，指杀。在香港，“劏房”指将大房间分割成（的）小房间，一般用于租借给贫穷人士居住。“劏房”和“豪宅”的并存是社会贫富悬殊的突出表现，是大家关注的问题。

收稿日期：2013-09-27 定稿日期：2014-04-17

基金项目：PolyU RGC Direct Allocation Fund. Project Account Code: A-PK14

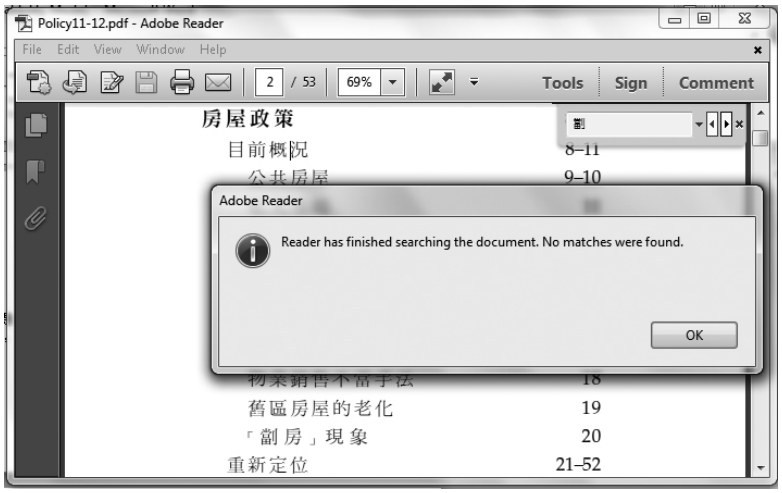

图 1　Adobe Reader **报告找不到“劏”字，但该字就在文中。**

2　原因分析

为了寻找问题的原因，我们从施政报告原 PDF **文件上直接将“劏房”拷贝到** Adobe Reader **的搜寻文字框里，发现该词在文字框里变成“・房”，不能正确显示。但点击右边的"** Find next**"按钮却能顺利查找，如图** 2 **所示。**

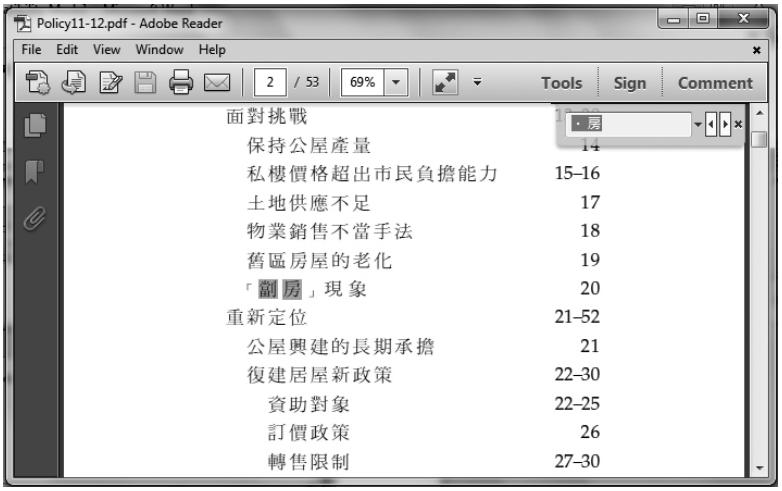

图 2　**拷贝原文中的“劏房”，在** Adobe Reader **上显示成“・房”，但能顺利查找。**

显然，在计算机的“眼”中，通过微软中文输入法输入的“劏”和施政报告中的“劏”是不同的字符。但是在人的眼中，从语言文字应用的观点来看，这两个“劏”完全是同一个字。为了进一步验证这一看法，我们又将施政报告中的“劏房”复制到 MS Word 2010 **上。用宋体（**Simsun**）显示，得到的也是“・房”。改为大字符集的** Simsun-ExtB **或** MingLIU-ExtB**，依然未能正确显示。但是使用支持《香港增补字符集》[2]的** MingLiu-HKSCS **或** MingLiu-HKSCS-ExtB **字体，则能得到正确的“劏房”字样。详细情况请见表** 1**。**

表 1　从香港政府施政报告中复制的“劏房”在 MS Word 2010 上的不同字体输出效果

字体	显示效果
Simsun（宋体）	・房
PMingLiu（新細明體）	・房
Simsun-ExtB（宋体-ExtB）	・房
MingLiu-ExtB（細明體-ExtB）	・房
MingLiu-HKSCS（細明體-HKSCS）	劏房
MingLiu-HKSCS-ExtB（細明體-HKSCS-ExtB）	劏房

可见，香港政府施政报告中的“劏”字使用了《香

港增补字符集》的编码。在 MS Word 上将光标移至“劃”字后按 Alt+X 键可查得该字的 Unicode 编码是 EB7B，用同样的方法可以看到通过微软中文输入法输入的“劃”的 Unicode 是 528F。显然这两个字符同形异码。同形异码字由于机器代码不同所以在计算机的眼中是不同的字，但是由于字形相同所以在人的眼中是同一个字。这就解答了上文关于“劃”字为什么不能正常检索和显示的疑问。

当然，同形异码不会局限于一个“劃”字。其他《香港增补字符集》的字也可能同形异码。而且增补字符集之外，还有同形异码现象。下面将结合从香港政府施政报告、澳门政府施政报告等重要文献中发现的实例，分类讨论当前计算机上的中文同形异码字。我们将主要讨论 Unicode 的同形异码字问题，因为 Unicode（等效于国际标准 ISO/IEC 10646）正在逐渐成为全世界范围内信息技术产品开发的主流，而且将逐步取代现有的国标（GB）码和大五（Big5）码等汉字编码标准[3]。其实，Unicode 已经是国际上影响最大的语言文字代码标准，广泛应用于 MS Windows，MS Office 和 WWW 等人们几乎天天都要用到的计算机工具。据谷歌官方博客上的最新报道[4]，WWW 上有超过 60％的文字内容是用 Unicode 编码的。

3 私人使用区造字引起的同形异码字

香港政府施政报告 2011-12 繁体字 PDF 版中的同形异码字不局限于《香港增补字符集》中的“劃”等方言字。例如，施政报告中“罰款”的“罰”字（须用 MingLiU_HKSCS 字体显示）与 Unihan 的“罰”同形不同码，前者的 Unicode 是 EE05，后者的 Unicode 是 7F78。因此，用其中任何一个“罰”都匹配不到另一编码的“罰”，严重影响检索结果，进而影响到其他语言信息处理。例如，在 Word 2010 上将同形异码字“罰（EE05）、罰（7F78）”转换为简体字时，得到的是“罰（EE05）、罚（7F78）”。后一个字能正确转换，前一个字却不行。因为转换字典中只收与一个字码对应的简繁体字对。

我们在施政报告上发现的同形异码字总结如表 2 所示。

表中第一竖列是同形异码字，共八个，第二列是该字在《香港增补字符集》中分配的 Unicode 编码，第三列是该字在 Unihan 中分配的专用 Unicode 编码，最后一列是取自施政报告原文的应用例子。

表 2 香港政府施政报告上的同形异码字

字形	Unicode (HK SCS)	Unicode (Unihan)	原文例子
羣	E373	7FA3	弱势社羣，目标组羣
着	E2E5	7740	多方面着手
劃	EB7B	528F	“劃 房”现象
鈎	E480	920E	挂鈎
邨	E473	90A8	屋邨，工业邨
牀	E27C	7240	病牀
衞	E40C	885E	医疗衞生
罰	EE05	7F78	罰款

《香港增补字符集》（HKSCS，Hong Kong Supplementary Character Set）是香港特区政府资讯科技处研制的收集了（当时）计算机上没有而市民需要的字符集合。最新版本是 2008 版，共收 5 009 个字，在 Unicode 编码空间中的私人使用区（Private Use Area，Range：E000-F8FF）编码。后来，新版的 Unicode（包括 Unicode 6.0 以后）的公共中日韩（CJK）汉字集 Unihan 收录了《香港增补字符集》，给每个字符又安排了专用码点，于是，这些字变成了同形异码。例如“劃”字在 Unicode 中既有私人使用区的代码 EB7B，又有 Unihan 的代码 528F。

在私人使用区造字而引起的同形异码字有两种情形。一种是在私人使用区造的字后来在公共空间又分配了码位，例如，《香港增补字符集》的情况；另一种是原本 Unihan 中有某字（目前 Unihan 收字达七万多个），但是用户又在私人使用区重造这个字。这可能是因为用户以为电脑上没有该字，也可能明知有该字，但由于特殊原因而再造一次。例如，学者张小衡和李笑通[5]在给汉字作笔顺排检时，就将笔形字符在 Unicode 的私人使用区按照比较合理的代码顺序重新编码，以便自动排检。

私人使用区的同形异码字在公开使用时，最需要统一到公共代码空间上来，因为不同的用户可能用同一个私人使用区的代码造出不同的字形来，造成更大的混乱。例如，Unicode EB7B 在 MingLiU-HKSCS 字体中对应的字符是“劃”字，但用 MS Outlook 字体显示的字形是“□”，用 Wingdings 字体显示字形是“※”，用 Adobe Arabic 字体显示字形是“⌊”，简直是五花八门。

4　兼容编码引起的同形异码字

跟香港一样，澳门特区政府的重要文献上也有同形异码字。表 3 是我们在《中华人民共和国澳门特别行政区政府二〇一二年财政年度施政报告》[6]繁体字版中发现的同形异码字。

表 3　《澳门特区政府二〇一二年财政年度施政报告》上的同形异码字

字形	Unicode (CI)	Unicode (Unihan)	原文例子
勵	F97F	52F5	鼓勵
離	F9EA	96E2	離船，離任，離不开
輪	F9D7	8F2A	輪候
領	F9B4	9818	領域，領导
數	F969	6578	數据，數量
臨	F9F6	81E8	臨时
類	F9D0	985E	各類，蛋類，人類
參	F96B	53C3	參与，參考
讀	F95A	8B80	就讀
療	F9C1	7642	医療
樓	F94C	6A13	酒樓，大樓
樂	F914	6A02	娱樂，樂观，安居樂业
藍	F923	85CD	藍图
靈	F9B3	9748	靈活
狀	F9FA	72C0	狀况，现狀
歷	F98C	6B77	歷史，歷程
識	F9FC	8B58	认識，共識，識法，意識
裡	F9E8	88E1	日子裡
聯	F997	806F	关聯，聯系
勞	F92F	52DE	勞动，勤勞
兩	F978	5169	一国兩制，兩个
論	F941	8AD6	論证
車	F902	8ECA	列車，車辆，环保車

表中第一竖列是 23 个同形异码字。施政报告全文版用的是第二列的代码，Word 2010 不能正确转换成简体字；而施政报告文本版用的是第三列的代码，Word 2010 能正确转换为简体字。两份文件取码不一，影响计算机信息处理。

在 Unicode 的编码空间中，上表第三列的代码都属于中日韩 Unihan 字符，第二列的代码都属于中日韩兼容表意文字（或称中日韩兼容汉字）。兼容表意文字一部分位于 CJK Compatibility Ideographs 编码区域（U＋F900 至 U＋FA2D），其它的属于 CJK Compatibility Ideographs Supplement 编码区域（U＋2F800 至 U＋2FA1D）。详细说明请见 Unicode 的相关码表，网址 *http://www.unicode.org/charts/PDF/UF900.pdf* 和 *http://www.unicode.org/charts/PDF/U2F800.pdf*。

兼容表意文字共有 400 来个，与原有的汉字同形异码。其中不少是常用字，例如，“理”字在 Unicode 中有 Unihan 代码 7406 和兼容码 F9E4。而且这两个字形都可以用 SimSun 和 MingLiu 字体正常显示。又如常用字“女”也有兼容码 F981 和 Unihan 码 5973。还有“年”（5E74，F98E），“度”（5EA6，FA01），行（884C，FA08）等。

根据 Unicode 官方文献说明[7]，中日韩汉字兼容区（U＋F900 through U＋FA2D）与主要的 CJK 统一表意文字块（primary CJK Unified Ideographs block，即 Unihan）分开编码，是为了方便同韩国标准 KS C 5601-1987 的往返转换（round-trip conversion with KS C 5601-1987），这些兼容码不可应用于其他目的。希望今后的澳门政府工作报告能注意到这一点。

5　建立笔画部首表而形成的同形异码字符

同形异码问题不仅存在于汉字，还存在于非成字的中文字符。非汉字的同形异码中文字符一般是汉字笔画和部件（包括偏旁部首）。这些字符原先是零星出现在 CJK 统一汉字集中，后来又建立专门的部首表和笔画表[8-10]，于是造成同形异码。例如，笔形“𠃌”在汉字集（Unihan）中的代码是 U4E5B，在中日韩笔画表（CJK Strokes）中的代码是 U2E82（可用 Adobe Song Std 字体显示）。由于 Unicode 笔画和部件表中有些代码的字形还未能在 SimSun 和 MingLiU 等主流字体上正确显示，目前使用 Unihan 字符尤其是其中的 BMP（基本多文种平面）字符比较安全。

表 4 是从文献[11]整理出来的 61 个非汉字同形异码中文字符及其 Unicode 代码。

表4 非汉字同形异码中文字符及其Unicode代码

中文字符	Unicode	中文字符	Unicode	中文字符	Unicode
𠂉	U4E5B, U2E82	龜	U9EFE, U2EEA	𠂤	U2E8F, U5C23
𠃌	U4E5A, U2E83,U31DF	𠚤	U7F52, U2EB2, U2EAB	𠂇	U723B, U2F58
〇	U3007, U25CB, U25EF	𠚥	U7F53, U2EB1	飠	U98E0, U2EDF
镸	U9578, U2ED2	刂	U5202, U2E89	灬	U722B, UFA49,U2EA4
耂	U8002, U2EB9	爿	U723F, U2F59	屮	U5c6e,UFA3C
襾	U8980, U2EC3	艸	U8278, U2F8B	纟	U7E9F, U2EB0
覀	U897E, U2F91	𠂢	U5F51, U2E94	幺	U5E7A, U2F33, U2E93
艹	U8279, UFA5E,U2EBE	氺	6C3A, U2EA2	糸	U7CF9, U2EAF
无	U65E1, U2E9B	钅	U9485, U2ED0	厶	53B62F1B
扌	U624C, U2E98	攵	U6535, U2E99	齐	U6589, U2EEB
尢	U5C22, U2E90	殳	U6BB3, U2F4E	竜	U7ADC, U2EEF
彐	U5F50, U2F39	龜	U2EF2, U4E80	辶	U8FB6, UFA66,U2ECC
忄	U8080, U2EBA	肴	U2EC7, U278B2	讠	U8BA0, U2EC8
已巳	U353E, U2E8B	饣	U9963, U2EE0	礻	U793B, U2EAD
疋	U758B, U2F66	夂	U590A, U2F22	衤	U8864, U2EC2
癶	U7676,U2F68	亻	U4EBB, U2E85	丬	U4E2C, U2EA6
阝	U961D, U2ECF, U2ED6	冒	U2ED5, U28E0F	氵	U6C35, U2EA1
攴	U6534, U2f41	辵	U8FB5, U2FA1	忄	U5FC4, U2E96
齒	U2EED, U6B6F	犭	U72AD, U2EA8	灬	U706C, U2EA3
歹	U6B7A, U2E9E	釆	U91C6, U2FA4		

上表中的十六进制Unicode数码前用U标示，例如，U4E5B表示Unicode 4E5B。某些代码的字形需用Adobe Song（宋）字体显示，例如，U2E82的“𠂉”。

Unicode笔画表和部首表中还包含一些成字字符，而且其中也有同形异码字，例如，CJK部首表中有“水”(U2F54)，Unihan中有“水”(U6C34)。又如，汉字“一”(U4E00)、部首“一”(U2F00)和笔画“一”(U31D0，在Word 2010上还不能正常显示)。

6 半宽和全宽字形分别编码而造成的同形异码字

有些占据书面空间大小不一的同形字符也使用不同编码，例如，ASCII逗号“,”(U002C)与中文逗号“，”(UFF0C)，ASCII“?”(U003F)与中文“？”(UFF1F)在中文书面语中混用时常被人们看作等价字符。当用电脑检索统计时，却是有区别的。有趣的是，微软新注音中文输入法的双/单字节(Double/Single Byte)逗号分别是“，”(FF0C)和“,”(U002C)，而微软拼音中文输入法的全形/半形(Full/Half Shape)输入的逗号都是“，”(U002C)。

“单/双字节字形”和“半/全形字形”在Unicode标准中称为“半宽和全宽字形”(Halfwidth and Fullwidth Forms)。把同一个字母的半宽和全宽字形分开编码，这也是一种同形异码现象。Unicode的全宽拉丁字母(Fullwidth Latin Letters)表的网址是*http://www.unicode.org/charts/PDF/UFF00.pdf*。收录有一百多个字符，包括字母、数字和标点符号等。

如果全宽和半宽字形的应用是为了文字排版外观的需要的话，则大可不必动用两套代码，用字体或字型(font)技术实现就行了，类似于楷体字和宋体字的处理方法。

还有一些其他类型的同形异码字。例如，拼音标调字母“ǹ”有两个Unicode代码：MSPY输入的

字符代码是 UE7C8，而 Unicode Latin-Extended B 表上的字符代码是 U01F9（The Unicode Consortium，2012e）。“ḿ”也有类似的情况。当我们在 Word 2010 上将字符串“ḿ（UE7C7），ḿ（U1E3F），ǹ（UE7C8）和 ǹ（U01F9）”转化为大写字母形式时，得到的结果是：“ḿ（UE7C7），ḿ（U1E3F），ǹ（UE7C8）AND ǹ（U01F9）”，前面的 ḿ 和 ǹ 都没有变化。其他的例子有“。（UFF61）”与“。（U3002）”，“、（UFF64）”与“、（U3001）”等。

7　同形异码字问题的解决方法

一般而言，同形异码字是信息处理的累赘。最彻底的解决方法是通过代码整理，尽量使得每一个中文字符对应一个 Unicode 代码，类似于汉字规范工作中的异体字整理[12]。应该在同形异码字的几个编码中选用一个作为该字的正码而淘汰其它的异码。关于如何选择，上文针对不同种类的同形异码字已经有所交待。

我们提倡使用正码字符，但是异码字符不可能马上销声匿迹。因为有些现有的文件已经含有异码字符，而且以后仍有用场，今后产生的文本也有可能出现异码字。换句话说，中文信息处理在相当一段时间内仍会遇到同形异码字问题。对于含有同形异码字的文本，可考虑以下处理方法。

1）在数据层面上，可研制专门的代码转换程序，将含有同形异码字的文本预先转换成纯正码后再作其他处理。例如，将所有的 Unicode 为 F9E4 的兼容字“理”都改为 U7406 的 Unihan 正码字“理”。

2）在程序层面上，可以在信息处理软件中加入同形异码字处理功能，使得无论用户用哪一个异码查检，计算机都会找到所有的同形字符。例如，当用户用 Unicode 7406 或 F9E4 的“理”检索时，计算机都会把两种代码的“理”都找出来。目前许多重要应用软件都未具备这种功能。我们在 Google 上搜索含“理”（U7406）的“理工”和含“理”（UF9E4）的“理工”，得到的结果数目分别是约 165 000 000 条和 224 000 000 条。百度的搜索结果也不同。而 Word 2010 在寻找（find）"理"（U7406）时，则明显地忽视了“理”（UF9E4）。图 3 是 Word 在本段文字上的搜索结果屏幕剪影。

在 jfj 汉字简繁体转换软件中，为了解决同形异码字问题，转换字典收录了与每个字码对应的繁简体字对。例如，既收字对“罸（U7F78）-罚”，也收“罸（UEE05）-罚”[13]。

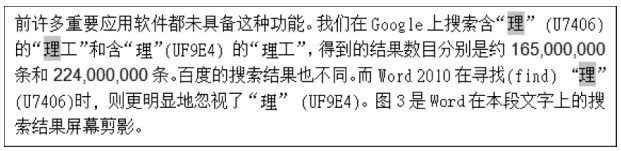

图 3　Word 在本文上段文字中寻找"理"（U7406）时忽视了“理”（UF9E4）

3）在用户层面上，首先，应该考虑让用户根据需要来设置是否区分同形异码字。在这方面，Excel 2010 的处理方法值得我们参考：当用户选择区分大小写字母（Match case）时，只匹配同码字；不区分大小写字母时，则匹配所有同形异码字。此外，还需要更好地利用 Unicode 的规范表达式（Normalization Forms）技术[7]。可以建立一个较为全面的同形异码字表，以便用户在使用信息处理软件时考虑到同形异码字的各个代码。例如，在 Word 2010 文档上寻找“理”字时，既寻找 U7406 的“理”，又寻找 UF9E4 的“理”，然后将两部分结果综合起来。

最根本的解决方法是 Unicode 本身的改善，尽量消除同形异码，做到字-码一一对应。中日韩汉字兼容区的异码字如果实在有需要保留，可考虑在字形上加标记（例如用较小的 k 表示韩语专用），以免误用于其它场合。此外，应该尽量少造字，不造计算机上已有的字。私人用字一旦纳入公共编码供大家使用，就立即停止使用私人造字时所用的代码。另外，还要改善中文输入法等文本生成软件，使计算机上新产生的中文字符都使用优选的汉字代码。

8　结论与讨论

同形异码字同一个字符拥有两个或两个以上的计算机代码，不仅浪费代码资源，还会直接影响语言文字应用与研究。例如，我们在前文已经看到，同形异码字的存在使得计算机检索不到完整的信息，使得某些汉字的繁简体转换被遗漏。关于同形异码字的讨论还未见有文章发表，上文主要讨论了同形异码字的不同类型及其相关问题的解决方法，希望能起抛砖引玉的作用。下面补充几点说明和讨论。

本文使用的语料实例主要来自香港、澳门和 Unicode 官方网站，但其影响并不局限于这三个方面。在中国两岸四地文化交流不断发展，中文信息处理日益国际化的今天，地区性的问题很容易影响到外地。例如，《香港增补字符集》的字可能会传到

外地，港澳政府的工作报告也可能会作为重要文件被收入台湾、内地或国际的现代汉语语料库。Unicode CJK 兼容表意文字中有许多是不同地区的共有常用字，例如，“理”、“領”、“車”等。而非汉字的中文字符在国际中文教学与应用中也起到相当重要的作用。

中文之外的字符也存在同形异码现象。例如，瑞典语、丹麦语和挪威语的字母 Å 的 Unicode 是 00C5，Å 的 Unicode 是 212B。又如 x_2 和 x_2 的下标“$_2$”的代码分别是 U2082 和 U0032，前者使用了专门的下标字符[14]，后者是传统 ASCII 字符的下标格式。此外，上文已经提到过，英文字母的全宽和半宽字符（fullwith and halfwith）也使用不同代码，例如，全形的 A（UFF21）和 a（UFF41），半形的 A（U0041）和 a（U0061）。因此，本文的讨论对于了解和处理中文之外其它语种的同形异码问题也有参考价值。

与同形异码相对应的是同码异形问题。有一些汉字新旧字形用不同代码。例如，户（U6237）-戶（U6236），术（U672F）-朮（U672E），强（U5F3A）-強（U5F37）；但是也有一些新旧字形用相同的内码，例如：骨（U9AA8）-骨（U9AA8），辶（U8FB6）-辶（U8FB6），反（U53CD）-反（U53CD），襯（U8941）-襯（U8941）（这里的异形是通过使用 SimSun 和 MingLiu 字体来实现的）。其中最令人费解的是“强（U5F3A）-強（U5F37）”异码，而“襯（U8941）-襯（U8941）”却同码。这种不一致现象也是 Unicode 不尽如人意的地方。

看来 Unicode 在统一取代 ASCII，GB 和 Big5 等其他代码标准的同时，其内部也需要进一步统一完善，尽量减少没有必要的或得不偿失的同形异码现象[15]，以便在计算机语言信息处理中更好地发挥作用。

参考文献

[1] 曾荫权. 中华人民共和国香港特别行政区政府二零一一至一二年施政报告：继往开来[R]. http://www.policyaddress.gov.hk/11-12/chi/pdf/Policy11-12.pdf. 2011

[2] 香港政府资讯科技总监办公室(2008). 香港增补字符集[S]. 香港：政府资讯科技总监办公室 http://www.ogcio.gov.hk/tc/business/tech_promotion/ccli/hkscs/.

[3] 陈壮. 中国在 ISO/IEC JTC1/SC2 的活动与中文编码的国际标准化[J]. 中文信息学报，2007，21(4).

[4] Google. Unicode Over 60 Percent of the Web [EB]. Posted on Google Official Blog by Mark Davis, International Software Architect, http://googleblog.blogspot.hk/2012/02/unicode-over-60-percent-of-web.html 2012.

[5] 张小衡，李笑通. 一二三笔顺检字手册[M]. 北京：语文出版社. 2013.

[6] 崔世安. 中华人民共和国澳门特别行政区政府二〇一二年财政年度施政报告[R]. http://portal.gov.mo/web/guest/info_detail?infoid=134838. 2011.

[7] The Unicode Consortium (2012a). The Unicode Standard, Version 6.2.0 [S], Mountain View, CA: The Unicode Consortium, http://www.unicode.org/versions/Unicode6.2.0/

[8] The Unicode Consortium (2012b). CJK Radicals, the Unicode Standard 6.2.0 [S]. http://www.unicode.org/charts/PDF/U2F00.pdf

[9] The Unicode Consortium (2012c). CJK Radicals Supplement, the Unicode Standard 6.2.0 [S]. http://www.unicode.org/charts/PDF/U2E80.pdf

[10] The Unicode Consortium (2012d). CJK Strokes, the Unicode Standard 6.2.0 [S]. http://www.unicode.org/charts/PDF/U31C0.pdf.

[11] Zhang, X. Computer Input of Non-ASCII Non-Hanzi Chinese Characters [J]. The Journal of Modernization of Chinese Language Education (中文教学现代化学报), 2012(2).

[12] 傅永和. 汉字规范化 60 年[J]. 语言文字应用. 2009(4).

[13] 张小衡. 一个支持人工校对的中文简繁体转换工具[C]. In 孙茂松，陈群秀编，中国计算语言学研究前沿进展 (2009-2011). 北京：清华大学出版社，2011：569-575.

[14] The Unicode Consortium (2012f). Superscripts and Subscripts, the Unicode Standard 6.2.0 [S]. http://www.unicode.org/charts/PDF/U2070.pdf.

[15] Whistler K. On the Encoding of Latin, Greek, Cyrillic, and Han. Unicode Technical Note #26 [R]. http://www.unicode.org/notes/tn26/tn26-2.html. 2010.

[16] The Unicode Consortium (2012e). Latin-Extended B, the Unicode Standard 6.2.0 [S]. http://www.unicode.org/charts/PDF/U0180.pdf.

张小衡（1958—），助理教授、博士，主要研究领域为计算语言学、计算机辅助汉语教学和现代汉字学。

E-mail：ctxzhang@polyu.edu.hk