談簡繁轉換的幾個關鍵問題

教育部語言文字應用研究所 王曉明（大陸）

財團法人中文數位化技術推廣基金會 魏林梅（臺灣）

一、引言

信息技術的迅猛發展，為信息交換構築了良好的平臺，網絡技術的發展，徹底改變了以往的時間、空間概念，使得信息交換日漸便捷與高效，因此備受人們的親賴。然而，中文的信息交換在此卻遇到了障礙，長期以來，由于大陸普遍使用簡體中文(GB 碼)而港、澳、臺使用的是繁體中文(BIG5 碼)，用戶平臺也分別使用不同的中文系統，兩岸四地中文信息不能直接互換，面臨著中文平臺和簡繁轉換兩大問題。隨著國際標準 ISO/IEC 10646 在信息技術領域裏的廣泛應用，中文平臺問題隨之而解，简繁漢字分離的狀況宣告結束。但這只解決了信息交換問題，信息交流問題還是沒有得到解決。本文將在國際標準 ISO/IEC 10646 框架內，談一談簡繁轉換涉及的幾個關鍵問題。

二、簡繁轉換的現狀

從上世紀九十年代中起，隨著國際標準 ISO/IEC 10646-1:1993 的發布以及兩岸四地交往的日益頻繁，大陸、臺灣和海外的一些研究機構開始研發簡繁漢字轉換工具。比如：大陸的中科院軟件所、四通利方公司、新天地公司；IBM(臺灣)公司、臺灣的倚天公司等。但是，由于缺乏文字學研究的支持，其中的某些系統尚達不到實用化的程

度；另外那些付諸應用的系統，雖然都能够實現一對一簡繁漢字的轉換功能，但是仍然解决不了一對多的轉換瓶頸問題。

網上泛濫的在綫簡繁轉換系統以及隨處可以下載的簡繁轉換軟件只能完成漢字一對一的轉換功能，一對多的簡體字被固定地轉換為同一個繁體字，而且這些系統中所用的簡繁漢字對照表也不全，大多數隻包含一千多個簡體字和繁體字的對照關係，而《簡化字總表》中所列簡體字一共有 2235 個，因此在轉換的過程中，很多漢字都沒有做轉換，轉換效果極差。

目前，比較通用的辦公套件，如：微軟 Office、Sun 的 OpenOffice、永中 Office、金山 WPS 等，基本上都提供了簡繁轉換功能，但是水平參差不齊。這些辦公套件的簡繁轉換功能仍然未實現對 CJK 20902 中所有簡繁漢字的轉換支持，對于一對多的漢字也經常轉換錯誤。

雖然過去的十餘年裏，很多研發機構和公司都努力研發實用化的簡繁轉換工具，但是目前還沒有一個真正的精密轉換系統被研發出來，同樣的內容，不同的工具轉出不同的結果來，導致結果的不可信，經常是計算機轉一遍，還得人工校一遍；而且在方便用戶使用上還存在諸多問題。究其原因，解决問題的辦法沒有找准，長期以來，都企圖依靠技術解决問題，結果事倍功半，解决了這個問題，又引起了新的問題。由于主攻方向性的錯誤，導致核心問題遲遲得不到很好地解决。仔細研究就會發現，除了文字本身的問題，更主要的是對簡繁轉換的認識問題。

三、簡繁轉換的幾個關鍵問題

簡繁轉換的核心問題是語言文字應用問題，即漢字在不同應用環境(如，大陸、臺灣、香港、澳門以及海外)下的使用問題。為了更好地解決簡繁轉換問題，必須打破傳統的條框局限，以一個全新的視角來審視這個問題。

首先，避免錯誤地以標準代替實際應用。標準是起規範作用的，而且它只規範了字與字之間的關係，并不能涵蓋實際應用，不能徹底解決漢字的實際應用問題，尤其是涉及簡繁用字關係問題。在實際應用中，容易出現的問題主要有以下三類：

一是簡轉繁和繁轉簡時遇到的一對多問題。由于標準對其之間的關係闡釋不夠充分，可操作性不强，有些一對多的情況人是完全可以辨別清楚地，但計算機卻無法理解和處理，原因是：人可以借助很多自己掌握的標準之外的知識，而計算機卻做不到，它只能依靠人給他提供的信息來分析、處理問題，如果機械地採用標準對照表，轉換發生錯誤的幾率會很高，該用 A 的地方用了 B，該用 B 的地方却用了 A，這是司空見慣的事，比如：发、面、里、出等字。以下的幾個簡轉繁的例句（词）是利用 Microsoft WORD 2003 的轉換結果：

被轉換詞語	轉換結果
发毛	發毛
他想明天去剪发。	他想明天去剪發。
下面	下麵
他来时，我正要吃面。	他來時，我正要吃面。

下面穿着裤子。　　下麵穿著褲子。

二是不同應用環境中的慣用字(詞)問題。大陸、臺灣、香港、澳門及海外等不同應用環境下漢字的使用習慣存在著諸多差異，如，“跟着”(大陆)和“跟著”(臺湾)，這在標準中是沒有體現的。要使得簡繁轉換的結果符合目標應用環境的用字習慣，就必須對不同應用環境下的大量真實語料進行統計和分析。

三是實際應用與標準不一緻的問題。比如：《簡化字総表》一表中明確規定簡體“淀”對應繁體“澱”，可我們卻不能將“海淀”轉為“海澱”。

由于一般人員對漢字的字、詞使用習慣等知識缺乏瞭解，多數人都是機械地利用標準對照表，從而造成該轉的沒轉，不該轉的倒轉了。

其次，避免簡單地以簡繁關係代替了簡、繁兩种環境間的實際用字關係。簡、繁兩种環境間的用字關係既包括通常意義上的簡繁關係，同時還有所謂的新舊字形關係、正異體字關係、術語關係等等。比如：

簡體環境	對應	繁體環境
冲	⟺	沖
厢	⟺	廂
宫	⟺	宮
说	⟺	說
卧	⟺	臥
软件	⟺	軟體

数字化  數位化

第三，避免錯誤地以歷時關係代替共時關係。共時的簡與繁之間的關係是雙向的，既要闡釋清楚繁與簡的關係，同時也要清楚地闡釋出簡與繁的關係，兩者缺一不可。然而，《簡化字總表》、《第一批异體字整理表》等，這些規範都是單方面指導大陸現實漢字應用的，它是站在歷時的角度闡釋漢字的字際關係，更多地是考慮繁體與簡體、歷史與現實的字際關係，是單方面地；它不是規範共時的簡、繁用字關係的，因而也就不可能徹底解決大陸與海外之間簡繁漢字的實際應用問題。應該從實際出發，認真調查研究簡、繁環境的現實用字、用語狀況，這樣才能徹底解決簡繁轉換的瓶頸問題。

第四、明確範圍，避免遺漏。這也是目前簡繁轉換軟件普遍存在的問題。在建立簡繁對應關係前，首先要明確簡繁轉換基於的字符集範圍，比如：CJK 20902 字、CJK 加 CJK A 27484 字等等。其次，要考慮字符集自身的封閉性，比如：有了簡體字，與之對應的繁體字是否也在其中？反之亦然。在轉換結果中要對不封閉性做出明確提示，即：對於那些應該轉換而目標字又不在所框定的字符集內的字作出明顯標記，不要將問題隱藏，要讓用戶了解問題所在之处，避免不必要的再校對工作。順便說一下，對於轉換結果不確定之处也應做出明確標記，比如：一對多的情況。

第五，對漢字編碼標準要有詳盡的了解。ISO/IEC 10646 是國際編碼標準，它是按文种給字符編碼，而不是按語種。其中的中日韓統一漢字，既包括中國的簡、繁、異體漢字，也包含日文漢字、韓文漢

字以及越南喃字，它吸納了多個國家、地區的需求，包括各自的特殊需求在内，因此，收入的字符情況有一定的複雜性。

對於簡繁轉換來説，需要注意的是字形問題。ISO/IEC 10646 在編碼過程中同時還作了漢字認同工作，認同所依據的原則與目前文字界的認識有一定的出入；此外，ISO/IEC 10646 作為信息處理用的國際編碼標準，出於兼容性方面的考慮，它覆蓋了各個國家、地區的現行標準，從而，也就避免不了其中有些例外情況的存在。因此，在國際標準 ISO/IEC 10646 框架内進行的簡繁轉換，不但包括編碼層面的轉換，同時還包含字形層面的轉換，同屬一類情況，可能採用的轉換方式會不同。比如新舊字形問題，其中有些字形差异是可以通過字庫互相轉換來解決的，只要換成本地風格的字庫就可以了，而有些字形差異必須通過代碼轉換，是要建立直接對應關係的，即：兩個形體同時出現在一個字符集内，各自擁有獨立的編碼。比如：宫廷的“宮”字，它的字形差异是不能通過字庫來轉換的，在 Windows 平臺中“宮（U+5BAB）”和“宮（U+5BAE）”是兩個編碼，前一個是大陸的用字，后一個是海外用字，轉換時它們之間必須建立明確的對應關係;而蜈蚣的“蜈”字，它的字形差異就可以通過字庫來解決，只要將其字體設置為港澳臺及境外的字庫，就顯示爲“蜈”，無需建立對應關係，也就是説它不需要進行代碼轉換。

第六、明確轉換的方向性。這一點很關鍵，原因是各自有各自的規範，遵從的標準也不盡相同，我們認為是異體關係的兩個字，很可能在對方同時作為正體存在，因此，針對不同方向的轉換應該建立不

同的對應關係，即：簡轉繁與繁轉簡分別建立不同的對照表，這樣既方便處理，同時也可以將複雜問題進行分解，便於問題的解決。比如：发（髮、發）、松（松、鬆）、干（干、乾、幹）等字，由簡轉繁是一對多，在轉換過程中常常出錯，但由繁轉簡是多對一，就不存在轉換錯誤的問題了；而著（著、着），由繁轉簡是一對多，在轉換過程中常常出錯，但由簡轉繁属多對一，就不存在轉換錯誤的問題了。

第七、轉換結果要盡量符合目標環境的用字規範。但是，這並不意味著轉換結果中的每個漢字都應該是規範字。有規範字的要用規範字，同時，與規範沒有衝突的字也是允許存在的。這一點強調的是，不要用被轉換環境的用字規範來約束轉換結果，應該用目標用字環境的規範來約束，即：由簡体轉換為繁體，轉換結果要符合繁體環境的用字規範；由繁體轉換為簡體，轉換結果要符合簡體環境的用字規範。這個道理誰都懂，但常常會被忽視，所以，在此贅述。比如：簡體“说”應該對應“說”，而不應該對“説”；簡體“为”應該對應“為”而不應該對“爲”，等等。

第八、被轉換的內容未必都是規範的，這一點常常被忽視，需要特別説明。一提到簡繁轉換，多數人就會認為是《現代漢語通用字表》的 7000 通用字或者 GB2312 的 6763 漢字與 BIG5 的 13000 字之間的轉換。事實並非如此，原因之一是，計算機中的編碼漢字量已大大超出了文字規範所圈定的範圍，這些字都是可以用的；原因之二是，規範並不能完全覆蓋社會實際用字；原因之三是，使用漢字的人並非都是文字、規範專家，更多的是普通百姓，因此，被轉換的內容未必是完

完全全地符合規範的，這一點在研究、開發轉換工具過程中要給予充分考慮。

四、結束語

有關簡繁轉換涉及的具體問題還很多，如：一對多的解決方案、專有名詞的轉換、名詞術語的轉換等等。除了文字本體的問題之外，簡繁轉換還涉及句法分析、自然語言理解等問題。無論是什麼問題、多少問題，都應該首先明確問題的歸屬，即：哪些是技術層面問題，哪些是文字使用層面的問題，這是解決問題的第一步。

解決簡繁轉換問題，應該從共時的角度出發，面向實際應用，避開學術爭端，注重實效。

期待在不遠的將來能夠徹底解決簡繁精密轉換問題，從而促進兩岸四地信息交流的便捷與通暢。