數位典藏缺字問題的解決

周亞民
景文技術學院資訊管理學系
milesymchou@yahoo.com.tw

黃居仁
中央研究院語言學研究所
churen@gate.sinica.edu.tw

摘要

以計算機處理中文，長久以來都必需要面對缺字問題，在數位典藏計劃中，尤其是古籍的數位化，更突顯這個問題的重要性。本研究以漢字知識本體(Hantology)解決缺字，利用後設語言描述缺字，並將缺字的知識以交換系統傳送給不同的系統，解決缺字的問題，比較過去缺字解決方法，這個方法能夠解決缺字交換和檢索兩個關鍵問題。

關鍵詞：知識本體、缺字問題、資訊交換。

1. 前言

由於編碼系統採用一個符號對應一個字符碼(character code)，能夠被放到計算機的數量，受限於編碼空間的限制，必須挑選那些文字或符號要放到計算機，這些被挑出來的形成字符集(character set)，再將字符集的每一個文字符號加以編碼，那些沒有被放到字符集的文字，如果使用者需要在計算機使用，將無法找到對映的字碼，無法順利的在計算機中表達，這些不在字符集中的文字稱為缺字。

因為缺字無所不在，又經常發生，缺字成為中文資訊處理最顯著的問題之一，為了讓缺字如字符集中的文字或符號能夠被計算機處理，造字是最常用的方法，但造字卻產生缺字交換和檢索的問題。

在數位典藏計劃中，許多的資訊科技的問題，並不會特別的提高其問題的顯著性，但是缺字的問題，在數位典藏計劃因為有許多重要的典藏是古籍，由於古今文字的差異，為了完整保存古籍原貌，現有的編碼系統無法用來表達所需要的文字，缺字成為顯著而必需要解決的問題。

我們對於如何認為缺字是計算機少了一個漢字的知識，因此，缺字發生時，必需將缺字的知識加到計算機中，並將缺字與其它漢字建立關係，計算機有了缺字的知識，在交換缺字的文件時，就可以將缺字的知識也交換給對方作適當的缺字處理，因此我們設計了一個缺字交換語言和交換系統，讓接收端的計算機根據缺字交換語言自動選擇適當的字形表達缺字，同時可以被檢索，即使字形不同也可以找到。

2. 相關研究

缺字問題的相關研究，可以區分為下列四類主要的方法：

(1)保留備用的字碼空間，作為造字使用

由於中文字碼的字集可能沒有使用者需要的字，因此，大部份的編碼系統都保留了一部份的字碼，使用者可以自行造字後，指定這個區域的字碼給所造的字，這個方法就是造字法。每個編碼系統所保留的字碼空間都是有限的，早期保留的較少，大約在數千個左右，目前的編碼系統保留的較多。保留區的字碼原來的用途是只能在同一部計算機使用，才能確保資料的正確性和檢索的正確性，因為這個區域的編碼，並不是統一編碼，一旦交換給其它的計算機，就可能發生編碼的衝突，無法正確的被處理和檢索。

(2)增加字符集大小

由於缺字是因為中文字碼的字符集中沒有所需要的字，如果增加字符集的大小就可以減少缺字的機會，因此，中文字碼的字符集大小，是很多使用者在選擇中文字碼的主要考量，Big-5，CNS11643，Unicode 等各種字碼也持續的增加字集的大小，以滿足使用者的需求。

(3)集中式管理

由於使用者自行利用保留空間造字，會發生缺字交換和檢索的問題，因此，若改為集中式的造字，可以避免重複造字而編碼不同的問題，日本的文字鏡[1]和台灣的全字庫[5]都是採用這個想法所提出缺字的處理方法。

(4)利用字形結構

目前不依賴字碼處理缺字的方法，主要是中研院資訊所文獻處理實驗室的構字式、漢字構形資料庫和電子佛典協會的組字式，兩種共同點都是由漢字二維結構關係著手[2, 3, 4, 7, 8]。

漢字的特性之一，就是可以分解為較小的部件，不同的字形有著共同的部件，這個特性被廣泛用在輸入法中，也是是構字式的基礎。謝清俊教授和莊德明先生提出的漢字的部件集，並分解超過五萬個字形，不僅證明這些部件集對漢字的良好表達能力，更發現以該部件集所表達的構字式只要字形不同，則分解出來的部件或字不相同，如此一來，這個特性就成為區別不同字形的好方法，因為遇到缺字時，也要讓計算機知道缺的是那一個字形。

構字式的字形呈現，則是利用 Big-5 原有的字碼空間，分別另造八套標楷體和細明體字，將構字式和新的字體放入漢字構形資料庫，只要使用者變換字體，即可顯示超過五萬個不同的字形。利用構字式和字體變換解決缺字問題的方法，是將使用者輸入的構字式，由漢字構形資料庫中找出字碼及對映的字體，因此，如果將缺字複製到其它不支援字體選擇的軟體，即可知道它重覆用那一個字碼。

3. 缺字問題分析

從我們對缺字問題的分析，以及過去解決缺字研究的缺失，可以提出幾個結論：

(1)缺字必然會發生，真正問題在於缺字交換和檢索

由上述缺字發生的原因所作的分析，我們知道是漢字的特性造成缺字，只要漢字的特性沒有改變，缺字就必然會發生，任何方法想要阻撓缺字的發生是不可能的，因此，希望能夠讓缺字不再發生是不正確的想法，真正要解決的問題是當缺字發生時，如何表達缺字，以及如何交換和檢索缺字，而缺字的表達過去的方法就解決了，例如造字或利用構字式將缺字表達在計算機，但是如何交換和檢索缺字尚未解決，這才是當前缺字問題的核心，只要此問題解決了，缺字的問題就被根本的解決了。

(2)缺字只是計算機進行中文處理的問題之一

計算機進行中文處理的有許多的問題，缺字只是其中一個問題，如果要解決缺字問題，應該從更宏觀的角度著手，才能找出問題的根本，並徹底解決問題，而且，我們相信如果缺字問題徹底解決，還有很多其它的問題也會跟著被解決。

(3)必須考慮缺字處理的環境，並與應用程式獨立

如果要解決缺字的交換和檢索問題，必須要考慮現在缺字處理的環境，因為我們不應該要求應用程式做任何修改，因為已經累積了大量的缺字文件和應用程式，如果需要修改程式才能進行交換和檢索，並不是好的方法，也不容易被接受。

(4)造字是缺字問題的主因，但應該接受它的存在

在各種缺字的表達缺字的方法中，大部份都是採用造字的方法[6]，造字也是造成缺字交換和檢索問題的主要原因，除非要求大家都不准造字，例如行政院推行全字庫多年，要求政府機關不能任意造字，但是根據行政院對政府機關所做的調查，只有百分之十三的受訪者遇到全字庫沒有的字，會提出申請並要求建立新的字碼，百分之三十八的受訪者仍然採用造字的方式，而實際上的比率應該更高，因為近百分之五十的受訪者未答或是暫不處理[5]，因此，要求不能造字是不可行的方法，但是造字的方法又是造成缺字問題的主因，而且造字的方法已經用了近三十年，累積這麼多的資料都是用造字的方法，要求全面改用其它缺字的表達方法和修改程式更加不可行，真正可行的方式是接受以造字表達缺字的事實，允許使用造字的方法，但是必需使用共同的方式交換，如果能夠解決這個部份的問題，缺字的交換和檢索問題大部份也解決了。

(5)必須增加交換的資訊，並制定交換的規格

如上所述缺字交換的問題之所以沒有被解決，原因之一是交換的資訊太少，如果解決缺字交換問題，必需要增加交換的資訊，而且還要制定交換的規格，才能讓交換的雙方都能正確的處理交換的內容，交換的標準則應該要符合目前網路的標準。

(6)必須增加計算機的漢字知識

計算機必須具備漢字的知識，才能在缺字交換和檢索時，提供所需的漢字知識，可惜的是計算機所擁有的漢字知識乏善可陳。謝清俊教授曾說:「雖然在今天，幾乎沒有電腦不能做的中文資料處理工作，可是，似乎也沒有電腦中文資料處理能力可以讓我們完全滿意的，造成這種現象的根本原因，是目前電腦擁有的中文知識不足」[10]，確實如此，因為在缺字就是因為計算機缺乏文字的知識，而造成無法正確的交換和檢索，例如很多缺字的問題，實際上是缺異體字，以佛研所建立大正藏的個案為例，有百分之七十都是缺異體字[9]，而計算機如果有異體字的知識，知道實際上有其它的字形可以用，交換或是檢索都應該可以使用異體字。

4. 解決缺字交換和檢索問題的方法

我們認為缺字的問題，應該要改變過去對缺字的看法，不應該只是將缺字視為計算機缺了一個字形，而是缺少一段知識片斷，這個知識片斷是缺字有關的知識，以及這個缺字與其它漢字之間的關係，為什麼應該將缺字問題視為計算機少了有關缺字的知識，而不是如大部份的人將缺字當作缺少字形，因為如果只將缺字視為缺字形，這個看法將會影響計算機未來對這個缺字的處理。如果只是缺字形，只要造個字形問題不就解決了，為什麼到至今還要面對缺字的問題，所以要解決這個問題，必需打破長久以來根深柢固的想法，想法必需要改變，才有可能徹底解決缺字的問題。

若缺字只是缺字形，搜尋缺字就找不到它的異體字，或是檢索詞與缺字不同也找不到，缺字交換時如果有其它字形，也不會知道其實是同一個字，對於自然語言處理也會形成未知詞，因為詞典沒有它，語音處理也無法進行文字轉語音處理(TTS)，只能跳過缺字，因為語音資料庫中沒有它，這些都是將缺字過於簡化為缺字形的結果，但是如果將缺字的形音義關係，以及它的異體字關係放入計算機，缺字交換有了異體字字形的知識，就可以使用異體字處理缺字，搜尋不會因為檢索詞與缺字不同而找不到，如果缺字的異體字是詞典已收的詞或詞素，就不是未知詞，有字音的知識就可以語音合成，或是利用字形結構的聲符合成，所以，將缺字視為缺字形，只能解決當下的顯示問題，所有後續這些缺字造成的連鎖問題都無法解決，但是將缺字視為缺少缺字的知識，而將缺字的知識放入計算機，上述問題都可以獲得解決。

為了徹底解決缺字，計算機必需有豐富的漢字知識，我們以漢字知識本體(Hantology)提供漢字的知識[12,13]，當發生缺字時，使用者先描述缺字的知識，再將它加入到 Hantology，建立缺字與其它漢字的關係，如果使用者需要將缺字的文件交換給其它單位，則利用後設語言(meta-language)描述缺字和相關的異體字，再與缺字文件一起傳送，接收端收到後，再利用缺字的描述，將缺字加入接收端的 Hantology 中，讓接收端的資訊系統能夠正確的處理缺字。

為了驗證這個方法，我們設計並實作了一個缺字交換系統(Missing Character Interchange System, 以下簡稱 MCIS)和缺字描述語言(Missing Character Description Language, 以下簡稱 MCDL)[11]，這個系統可以找出文件中的缺字後，利用 Hantology 自動產生缺字的描述。此交換系統，其架構中主要的模組為 MCDL Parser 、MCDL Generator、Packer、unPacker。MCDL Parser 主要的功能是分析 MCDL，將 MCDL 描述的缺字加入到 Hantology 和，MCDL Generator 由缺字文件找出缺字並搜尋 Hantology 是否有缺字的知識存在，如果有缺字的知識則將缺字的知識取出，產生 MCDL 描述缺字的知識，Packer 則是將缺字文件、MCDL 和缺字的字型封裝成為一個 MCDL package，unPacker 的功能是將收到的 MCDL package 分解後交給 MCDL Parser。

當傳送端的使用者需要交換缺字的文件時，必須先利用缺字交換系統打開缺字的文件，MCDL Generator 將會找出文件中的缺字，檢查缺字是否已經在 Hantology 中，不在 Hantology 中的缺字，將會要求先將缺字放入知識本體中。如果文件中的每一個缺字都確認完成，MCDL Generator 會對每個缺字產生對應的缺字描述，使用者可以要求產生通用異體字，MCDL Generator 同時會產生缺字的通用異體字描述，若使用者需要產生部份異體字，缺字交換系統必須詢問使用者缺字的字義，才能從 Hantology 中過濾出適當的部份異體字。

MCDL Generator 產生的 MCDL 交給 Packer，Packer 到系統中的造字檔中，將缺字的字型抽出來，最後將 MCDL、缺字文件和缺字字型一起封裝後，用來作為傳送的檔案。接收端收到缺字的封裝檔後，unPacker 會將它還原為 MCDL、缺字文件和缺字字型，一起交給 MCDL Parser，根據 MCDL 的描述，MCDL Parser 將缺字與接收端的 Hantology 整合，針對異體字的描述找出異體字後將它們的關係建立起來，每個缺字都會利用構字式[3,7,8]檢查接收端是否已經存在，如果已經存在，根據接收端的字碼將缺字文件中的缺字字碼替換，若不存在，則會找出空的字碼給這個缺字，並將缺字字型加入接收端的造字檔和造字表中，但是，如果使用者選擇可以使用異體字，接收端對映的缺字不存在時，則會檢查是否有異體字，若有異體字可以使用，就不會找空的字碼，而是使用異體字取代缺字。

我們以中華電子佛典協會數位化的大正藏的長阿含經，說明如何應用漢字知識本體和缺字交換系統，解決缺字的交換問題。這份經文中有三個缺字分別是「蕆」、「驈」和「鐵」(圖 1)，使用者以造字的方式在計算機中表達「蕆」、「驈」、「鐵」，同時指定字碼為 FA46(U+E004)、FA45(U+E005)、FA46(U+E005)，描述缺字並加入到 Hantology 中，我們以缺字「鐵」為例，「鐵」不僅在說文未收，漢語大字典也沒有收，因此，這個字只能從文字的演變規律來分析，「鐵」應該是「戜」的增繁字，「戜」的本義為兵器，可能為了強調所使用的材質，因此加了表示金屬的意符「金」而成為「鐵」，增加或改變意符的現象，在漢字的發展中，是常見的現象，「戜」也出現在其它的經文中，而異體字在佛經是很普遍的，再從本經文的前後文判斷，應該是兵器無誤。因此，我們利用異體字知識管理系統描述「鐵」和「戜」兩字的異體關係後(圖 3)，自動將這些知識與漢字知識本體連結(圖 4)，便可以在他的計算機中處理這部經文(圖 5)。

諸末羅即共入城。供辦葬具已。還到天冠寺。以淨香湯洗浴佛身。以新劫貝周匝纏身。五百張疊次如纏之。內身金棺。灌以香油。奉舉金棺置於第二大鐵槨中。栴檀木槨重衣其外。以眾名香而蕆其上時。有末羅大臣名曰路夷。……大迦葉將五百弟子從波婆國來。在道而行。遇一尼乾子手執文陀羅花。……充足生快樂如羅漢遊法如象被深鈎而猶不肯伏驈突難禁制放逸不自止猶如清涼池眾花覆水上……讚歎稱說本所誦習……彼諸大仙頗駕乘寶車。持鐵導引。白蓋自覆。手執寶拂。著雜色寶屣。又著全白疊。如汝師徒今所服不答曰。不也摩納。汝自卑微。不識真偽。而便誹謗。輕慢釋子。自種罪根。長地獄本。云何。摩納。如彼諸大……

圖1 大正藏長阿含經的部份經文

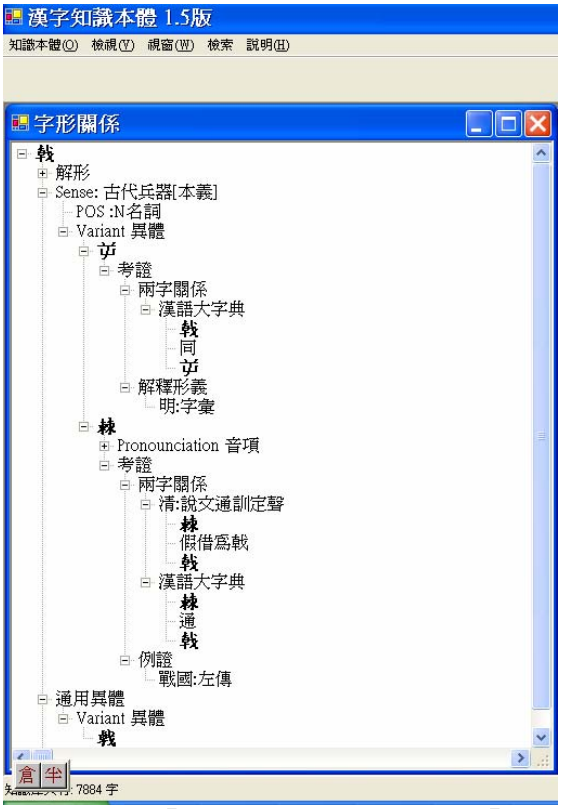

圖 2 傳送端「戜」的字形關係(加入缺字「鐵」前)

將這些缺字的知識加入到 Hantology 後，再利用缺字交換系統打開這部經文，缺字交換系統將根據 Hantology 產生 MCDL 的缺字描述，交換系統先詢問使用者是否要利用 Dublin core 描述缺字文件，使用 Dublin core 可以幫助未來檢索這份文件時，決定適當的異體字進行檢索(圖 6)。

圖 3　利用漢字知識本體管理系統描述缺字「鐵」和「戟」的關係

圖 4 傳送端「戟」的字形關係(加入缺字「鐵」後)

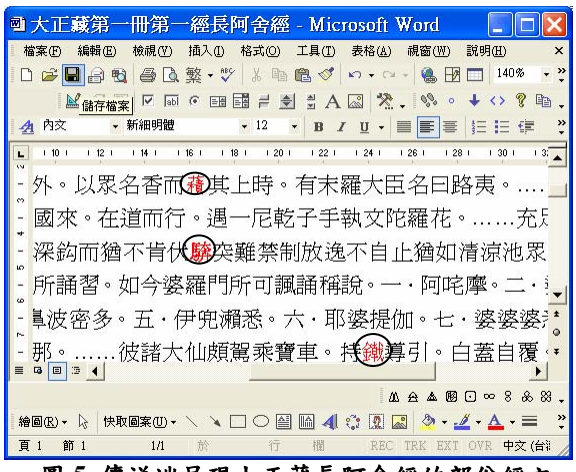

圖 5 傳送端呈現大正藏長阿含經的部份經文

```
<?xml version="1.0" encoding="utf-8"
standalone="yes" ?>
<MCDL
        xmlns=http://www.im.ntu.edu.
        tw/mcdl-syntax-20050101
        xmlns:DC="http://purl.org/dc/
        elements/1.1">
<Character pos="87">
<Apply-glyph>
  <Glyph-expression>𧦝
</Glyph-expression>
  <Code charset="big5">FA46</Code>
  <Font>FA46.gly</Font>
  </Apply-glyph>
  <Semantic-symbol>艸
</Semantic-symbol>
  <Variants />
  </Character>
<Character pos="157">
<Apply-glyph>
  <Glyph-expression>驎
</Glyph-expression>
  <Code charset="big5">FA45</Code>
  <Font>FA45.gly</Font>
  </Apply-glyph>
  <Semantic-symbol>馬
</Semantic-symbol>
  <Phonetic-symbol>奔
</Phonetic-symbol>
<Variants>
  <Glyph-expression class="general">驥
</Glyph-expression>
  </Variants>
  </Character>
<Character pos="271">
<Apply-glyph>
  <Glyph-expression>金𠦝戟
</Glyph-expression>
  <Code charset="big5">FA44</Code>
  <Font>FA44.gly</Font>
  </Apply-glyph>
  <Semantic-symbol>金戈𠦝
</Semantic-symbol>
<Variants>
  <Glyph-expression class="partial"
meaning="古代兵器">戟
</Glyph-expression>
  </Variants>
  </Character>
  <Document>大正藏第一冊第一經長阿含
經.txt</Document>
<Meta><DC:identifier>大正藏第一冊第一經
長阿含經</DC:identifier> DC:title>長阿含經
</DC:title>
<DC:date>317/420</DC:date>  </Meta>
  </MCDL>
```

圖 6　缺字交換系統產生的 MCDL 缺字描述

接收端的缺字交換系統收到缺字交換 MCDL 封裝檔後，讀取 MCDL 缺字描述，將缺字的知識與接收端的 Hantology 整合，圖 7 為接收端整合缺字的知識前的漢字知識本體，圖 8 則為整合缺字的知識後的漢字知識本體，最後，缺字交換系統會對缺字文件作處理，如果使用者設定可以使用異體字，有符合條件的部份異體字或通用異體字將會被使用，而不使用原來的缺字，例如接收端的「驂」已改用「驏」(圖 9)。

圖 7　接收端整合缺字「鐵」前的漢字知識本體

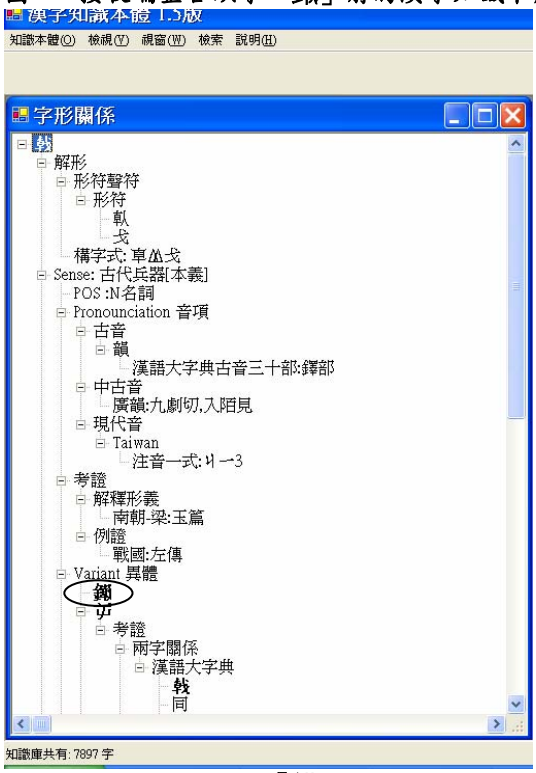

圖 8　接收端整合缺字「鐵」後的漢字知識本體

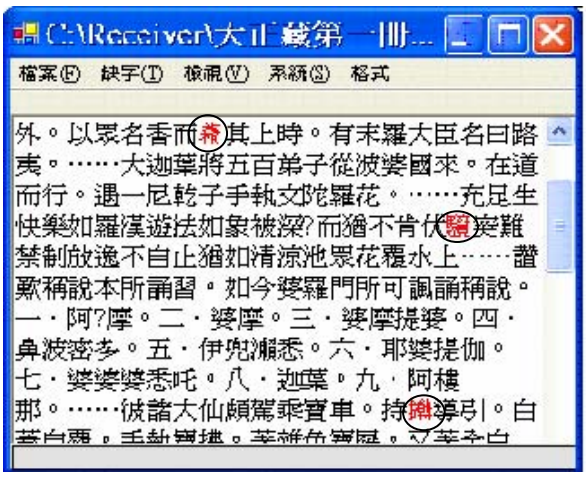

圖 9 接收端正確的呈現長阿含經的內容

5. 結論

本研究所提出的方法可以成功的讓過去我們提出的缺字交換系統，也是使用 XML 描述和交換缺字，成功的在不同的應用程式之間交換缺字，但是缺字與其它漢字之間的關係並沒有描述，而是將缺字與獨立於其它漢字，也沒有交換字形結構以外的描述，更沒有漢字知識本體可以提供漢字的知識和關係．本研究則是以我們所建立的漢字知識本體為基礎，將缺字的描述加入漢字知識本體，建立缺字與其它漢字之間的關係，尤其是部份異體字的關係，缺字的描述和交換皆增加字形以外的描述，使得傳送和接收端能夠對缺字做更適當的處理，也能將缺字與接收端的漢字知識本體整合，建立與接收端漢字之間的關係．我們將各種缺字交換的方法的結果彙整於表 1。

表 1　本研究與其它缺字交換的比較
(V: 完全解決 O: 部份解決 X: 不支援)

	造字檔交換	可攜式文件	構字式與字體變換	本研究
與應用程式獨立	V	V	X	V
可同時處理多個缺字文件	X	V	V	V
可檢索缺字	X	X	O	V
可編輯缺字	X	X	O	V
可處理有通用和部份異體的缺字交換	X	X	X	V
可處理有通用和部份異體的缺字檢索	X	X	X	V

表 1 呈現了三個缺字交換解決方案的里程碑，第一個里程碑是可攜式文件，它解決了缺字的呈現問題，第二個里程碑是構字式和字體變換，它不僅

決了缺字的呈現問題，而且可以進行檢索和編輯，但是限制是應用程式必需修改，且支援字體的變換，第三個里程碑是本研究所提出的方法，不僅可以沒有任何限制下呈現、檢索和編輯缺字，還可以自動使用異體字取代缺字，無論是通用或是部份異體關係，還可以檢索缺字的異體字，或是以異體字檢索缺字。有異體字的缺字檢索和下一節的異體字檢索的應用不太相同，即使計算機有異體字的知識，能夠進行異體字檢索，也不能進行有異體字的缺字檢索，因為缺少一個將缺字與其它漢字之間建立關係的機置，找缺字就一定找不到它的異體字，反之相同。本研究提出的方法，可以徹底解決長久以來使用造字所形成的缺字問題。若要能夠解決大部份的缺字問題，可以使用本研究所提出的方法，但所需要的安裝的軟體系統也是所有方法中最多的，因此解決缺字問題要考慮需求，如果只是要能夠顯示缺字，可以使用可攜式文件，若是考慮到編輯，則在支援字體的應用系統中，可以使用構字式和漢字構型資料庫，但是，若同時要解決缺字顯示、搜尋、編輯和其它後續中文資訊處理，則可採用本研究提出的方法。

誌謝

本研究感謝中央研究院資訊所謝清俊教授、簡立峰教授、何建民教授、莊德明先生，台灣大學資訊管理研究所吳玲玲教授，師範大學國文系季旭昇教授，京都大學 Christan Wittern 教授等給予的意見。

參考文獻

[1]古宇時雄、前寺正彥、野村英登、穀本玲大、谷田貝常夫(1999)， How 90000 Mojikyo fonts are working at present by the extension of UTF-16:文字鏡字庫與運用，電子古籍中的文字問題研討會，臺北，6 月 14-16 日。

[2]莊德明(2001)，中文電腦缺字解決方案，第一屆中國文字學國際學術研討會，天津， 8 月 22-25 日。

[3]莊德明、謝清俊(2005)，漢字構形資料庫的建置與應用，漢字與全球化國學術研討會，台北，1 月 28-30 日。

[4]周亞民(2003)，缺字文件的交換，智慧型漢字編碼會議，中央研究院，3 月 17-29 日。

[5]陳玉芳(1999)，「政府機關使用 CNS 11643 中文標準交換碼全字庫概況」調查，行政院主計處政府機關資訊通報 ，11 月，頁 13-17。

[6]邱菊梅(2001)，中文電腦缺字研究，玄奘人文社會學院中國語文研究所，碩士論文。

[7]謝清俊(2003)，缺字問題的回顧與前瞻，漢字智慧型編碼與應用研討會，台北，3 月 17-19 日。

[8]謝清俊、周亞民、莊德明、Christian Wittern, John Lehman(2001)，解決缺字問題結案報告，教育部委辦。

[9]謝清俊(1996)，電子古籍中的缺字問題，第一屆中國文字學會學術研討會，天津，8 月 25-30 日。

[10]謝清俊(1992)，談中國文字在電腦中的表達，中國文字的未來，海峽交流基金會，一版一刷。

[11]Ching-Chun Hsieh, Lin-Lin Wu, Ya-Min Chou(2004), A Missing Characters Description Language for Han Characters, ***Proceedings of International Computer Symposium***, Taipei, Taiwan, Dec. 15-17, pp.954-959.

[12]Chou, Y.M. and Huang, C.R.(2006), Hantology: Linguistic Resources for Chinese Language Processing and Studying, to appear in ***Proceedings of Language Resources and Evaluation***, Genoa , Italy, May, 24-26.

[13]Chou, Y.M. and Huang, C.R.(2005), Hantology: an Ontology based on Conventionalized Conceptualization, ***Proceedings of Ontologies and Lexical Knowledge Bases***, Oct. 15, pp.7-15.