異體字語境關係的分析與建立

周亞民
台灣大學資訊管理研究所
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-利用計算機處理漢語，實際上是透過漢語的書寫形式，而異體字是漢語書寫形式的特性，但是長久以來異體字的關係沒有被適當的表達，而且將異體字關係過度的簡化爲全同異體字，而實際上不同的異體字在使用上並不完全相同，本研究之目的是表達異體字的關係，並提出詞義、聲韻、構詞、時間、空間、構詞的語境(context)模型，根據此模型分析和建立異體字的關係，作爲中文資訊處理的基礎資源。

1. 簡介

以計算機處理自然語言，需要解決語言形式與語意之間的關係，對於資訊檢索和機器翻譯皆是如此，WordNet 的重要性就是因爲它建立了不同的詞彙形式和語意之間的關係，不過，異體字和異體詞的關係，並不是藉由 WordNet 可以建立的，主要原因是漢語的書寫形式不同。漢語的書寫形式是表意書寫系統(Ideographic Writing System)[Miller 1991][Coulmas 2003]，表意文字是概念的具體表徵，由於漢字並不是特定的人所創造，因此不同字形表示相同的概念，或不同的概念使用相同的字形，是很普遍的現象，而且隨著時間的變遷，使用的範圍廣，字形所表示的意義擴大或縮小，最後交織成複雜的一詞多形關係，這些關係將不同的字形連結成爲異體字。

異體字關係是文字學的主要議題，異體字可以分爲全同異體和部份異體，全同異體指的是音義完全相同而字形不同的字，而部份異體只需有部份用法相同即可，用法完全相同的稱爲狹義異體字，廣義異體字則包含部份異體字和全同異體字[裘錫圭 1995]。也有文字學家認爲只有全同異體字才能被認定是異體字 [董琨 1993][洪成玉 1995] 文字學中討論的正字、俗字、通假字、假借字、古今字、重文等，都是與異體字有關，但是都只由一個面向討論，缺乏整體的架構可以分析和比較異體字的關係，因此，也使得要分析異體字的關係不容易。因爲計算機的編碼系統採用一個字形一個字碼，因此對於計算機而言，只要字碼不同就是不同的字，例如：群≠羣、說≠説，而造成中文資訊處理上的問題，尤其是檢索，但是它們都是同一個字的異形或異寫的異體字。爲了要在計算機中表達異體字的關係，最早的是謝清俊教授設計的 CCCII(Chinese Character Code for Information Interchange)[謝清俊、黃克東 1989]，其後則是漢字構形資料庫[莊德明 1999][莊德明、謝清俊 2005]，以及中央研究院歷史語言研究所袁國華教授提出的建立 UNICODE 漢字異體字表與異體字辭典之相關研究[袁國華、曾黎明 2005]。這些研究所建立的異體字關係，大部份都將異體字關係簡化爲全同異體字，但是真正用法完全相同的異體字不多，其它都是只有部份用法相同的部份異體字[裘錫圭 1995]，因此，我們提出一個能夠表達部份異體字關係的模型，並且據以表達異體字的關係，作爲中文資訊處理的基本資源。本研究是漢字知識本體研究中的一個部份，漢字知識本體除了描述異體字關係，還描述書寫形式、語音、字義、變異、詞彙衍生，以 OWL-DL 描述漢字的知識，並且與 IEEE SUMO(Suggested Upper Merged Ontology)和

GOLD(General Ontology for Linguistic Description)整合，以提供計算機在自然語言處理所需的書寫、構詞、語法知識[周亞民 2005] 。

2. 異體字的語境模型

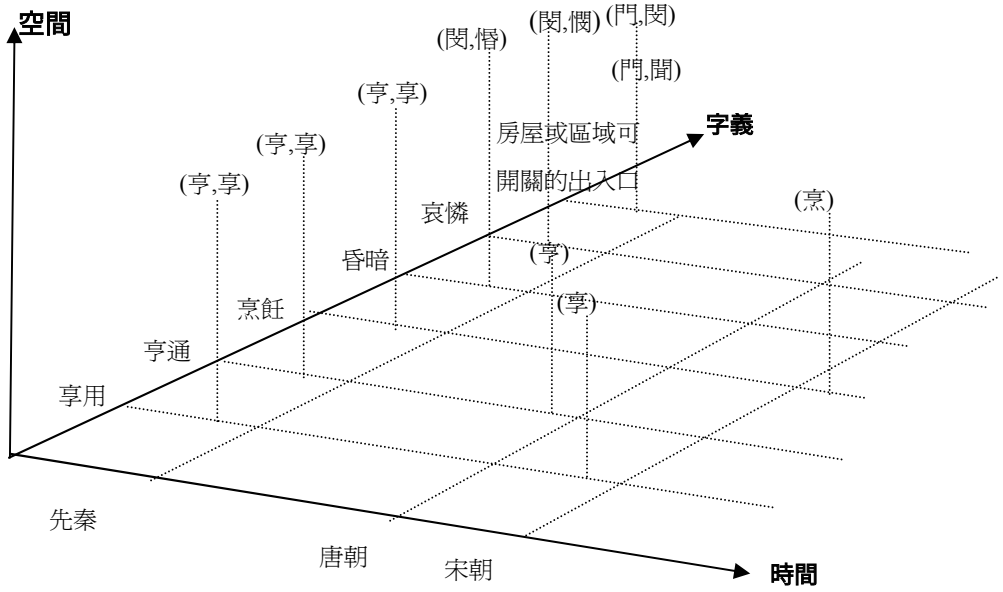
本研究對異體字的定義是音義全部或部份相同而字形不同即是異體字，甲骨、金、籀、篆、隸、楷等不同的文字發展不是異體字，因爲其改變是全面性的，不是發生在某一個字形的改變。我們將異體字的語境區分爲意義、時間、空間和構詞，分述如下：

(1)詞義

第一個重要的語境是兩個字形在什麼詞義時可以互相使用，如「蛇」古作「它」,「它」的本義爲蛇，但因常常假借爲第三人稱，因此，另加意符「虫」表示其本義，這兩個字只有「蛇」的意義爲古今字，第三人稱則不是古今字。又如陣戰的「陣」古作「陳」，論語衛靈公：「衛靈公問陳於孔子」，後改「東」爲「車」，說文沒有「陣」但收「陳」,「陳」與「陣」只有在陣戰的意義爲古今字。又如「戚」的本義爲兵器，假借爲憂慼的「慼」，論語：「小人長戚戚」，後造本字「慼」，故「慼」與「戚」在憂慼是假借關係的異體字。又如「拾」與「十」在數字的意義爲異體，其它的意義並沒有異體字關係，在「路不拾遺」這個詞彙中就不能寫「路不十遺」，因爲「拾」的意思是撿取而非數字，因此，必需區別「拾」與「十」在那些意義爲異體，以判斷兩個字形是否有異體字關係。

(2)時間

異體字關係的第二個重要的語境是時間。現代漢語的書寫形式中不存在著異體字關係，並不表示古代漢語不是異體字，例如「亨」與「享」，現在不是異體字，但是漢朝以前兩字可以通用，說文解字注：「亨，隸書作亨作享，小篆之變也」,「亨」與「享」是隸變後寫法不同，但意思相同的字，干祿字書：「亨」是亨通、亨宰,「享」是祭亨，因此，到了唐朝仍然以「亨」表「享」，而「享」已經沒有「亨」的意思，而「烹」爲亨飪的俗字，說文、玉篇和廣韻都沒有「烹」字，直到類篇和集韻才有，因此，可能到了宋朝才用「烹」表示烹飪的意思(圖一)。



圖一 異體字之意義、時間與空間的關係

再以正俗字爲例，正俗字大部份的詞義原來都是相通的，但是後來可能變成用法不同的字，如「邪」與「耶」原爲用法相同的異體字，干祿字書說：「耶邪：上通下正」，「邪」本義爲地名，後假借爲疑問詞「耶」，史記：「羽豈其苗裔邪？」又因爲「耳」與「牙」在漢朝時兩個字形接近，成爲沒有區別的異體字，但是現在兩個字形變成用法不同，疑問詞的「耶」不作「邪」，正邪不作「耶」。

每個漢字所表達的意義，會隨著時間而改變，什麼時候開始有某個意義，會影響與其它字形之間是否爲異體字的判斷，要知道某個字形何時有此意義並不容易，如果要進一步知道兩個字形在特定的時間，有那些共同的意義就更加困難，但也不是不可能，由於本研究關心的是異體字之間的關係，所以並不需要將造字以來曾經有的意義都加以分析整理，最重要的異體字之間有交集的意義，盡可能的由字書和例證中去找尋證據。

(3)空間

由於漢字在不同地區的使用差異，也造成了異體字的現象，尤其是國家處於分裂的情形，異體字的現象更爲顯著，以春秋戰國時期來說，當時處於分裂的國家，文字的差異很大，說文：「田疇異畮，車涂異軌，律令異法，衣冠異制，言語異聲，文字異形」，文字異形指的就是戰國的字形因地而異的現象，同一個概念，有的國家用本字，其它國家用假借字，或不同國家用不同的假借字，例如「門」本作「門」，但是齊國假借「聞」作「門」，燕國假借「閔」作「門」，另外，還有部件使用不同，如：秦國的「廚」字爲从广从尌，而三晉卻是从广朱聲[裘錫圭 1995]。中國大陸於 1956 年開始進行了漢字簡化方案後，使漢字的結構大幅的改變，也造成了兩岸在文字使用上的差異，產生簡繁字形的異體字。

(4)構詞

異體字雖然音義相同，但是在構詞上可能不同，例如「記」和「紀」爲部份異體，在「記載」這個概念時，兩字爲異體，但是古漢語在這個概念的使用上沒有區別，但是現代漢語卻有區別，例如: 「筆記本」不寫爲「筆紀本」，「記者」不寫「紀者」，「紀念日」不寫「記念日」，「記憶」不寫「紀憶」[裘錫圭 1995]。

又如「昇」與「升」雖然有異體字關係，但是在構詞上並不一定能夠相替換，例如當作上升之意時，「升高」可作「昇高」，「提升」也可寫「提昇」，但是「昇華」不作「升華」，又如字義爲登進時，「升學」不作「昇學」，「升官」不作「昇官」[洪嘉馡、巫宜靜、黃居仁 2005]。簡化字與繁體字也有相同的情形，例如「藉」簡化爲「借」，但是「狼藉」不可簡化爲「狼借」，「慰藉」也不可簡化爲「慰借」，又如「乾」簡化爲「干」，然而「乾坤」不簡化爲「干坤」，「乾隆」不簡化爲「干隆」[江藍生、陸尊梧 2004]

與異體字構詞有關係的還有異體詞，例如：「按語」和「案語」、「梅雨」和「霉雨」， 異體詞又可以分爲有異體字關係和沒有異體字關係，如「梅雨」和「霉雨」，「梅」和「霉」並不是異體字，兩個字的意義並不相同，因江南一帶梅子成熟季節連綿降雨而得「梅雨」一詞，「梅」的概念是梅子，而「霉雨」是因爲連日下雨東西容易發霉而得霉雨，「霉」的概念是發霉，故「梅

雨」與「霉雨」是不存在異體字關係的一詞異形，這個問題是中文詞網可以解決的，在本研究只要描述「雨」可以衍生「梅雨」和「霉雨」，在中文詞網將「梅雨」和「霉雨」放入同一個同義詞集，計算機就知道梅雨與霉雨同義，而本研究關心的異體字問題，因爲中文詞網並不處理異體字問題，而有異體字關係的一詞異形，本身就是異體字，當然可以交換使用，只要整理異體字關係就可以解決。

3. 異體字關係的建立

我們採用中文電腦基本用字的常用字集、說文解字和漢語大字典的異體字表的共同交集共二千七百組作爲分析的對象。分析的重點包括不同字形在那些字音、字義和時間有異體字關係、字書描述異體字關係的體例和異體字構詞的限制。字音依古音、中古音和現代音分別建立，字義則區分本義、引伸義和假借義，對於異體字使用的時間和字義，則以例證的根據。 異體字關係的建立分爲三個部份：

(1) 建立字書對異體字關係的描述

確認異體字關係最基本的證據來自於字書，因此，建立異體字關係的第一步就是必須能夠描述字書對異體字關係的解釋。不同字書對異體字的描述體例不同，我們整理出來字書對異體字的體例如下：

A. 古作

「古作」描述異體字的古今字關係，例如：

集韻：「嶽，古作  。」

龍龕手鑑：「巖，古作岩。」

玉篇：「巨，古作丘。」

B.古文

「古文」也是描述古今字關係，例如：

說文：「仅，古文奴。」

一切經音義：「氾，古文泛。」

龍龕手鑑：「，古文米字。」

C.今作

「今作」描述古今字關係，例如：

玉篇：「  ，導也，今作唱。」

玉篇：「  ，移也，今作徒，同。」

龍龕手鑑：「  ，今作艱。」

D.後作

「後作」描述的也是古今字關係，例如：

漢語大字典：「或，後作國。」

漢語大字典：「暴，後作曝。」

E.本作

「本作」大部份描述的是本字，例如：

正字通：「瀏，本作。」」

玉篇：「噚，本作粵。」

F.本字

「本字」大部份也是描述本字，例如：

字彙：「喪，喪本字，从哭从亡。」

正字通：「弘，彈本字。」

G.通

「通」在異體字描述的關係比較複雜，最早使用「通用字」或「通」的體例是干祿字書[章瓊2004]，干祿字書的體例關係是通用的俗體字，其它字書有的是同源字，也有近義字，最多是指通假字，漢語大字典的體例也是通假字。

集韻：「儃，讓也，通作禪、嬗。」

正字通：「疑，又與擬通。」

干祿字書：「虛，虛的通行體。」

漢語大字典：「卿，通慶。」

H.俗作

「俗作」描述的是正俗字的關係，例如：

龍龕手鑑：「峨，俗作峩。」

字彙：「兔，俗作兎。」

正字通：「健，俗作健。」

干祿字書：「突，突，上俗，下正。」

I.用同

「用同」在漢語大字典中所表示的異體字關係爲後起同音替代字，例如：

漢語大字典：「咨，用同齜。」

漢語大字典：「廢，用同費。」

J.也作

「也作」在漢語大字典用的很多，只能描述兩字有異體字關係，例如：

漢語大字典：「忍，也作肭。」

漢語大字典：「轏，也作棧。」

K.亦作

「亦作」只能描述兩字有異體字關係，但無法明確描述何種異體字關係，例如：

說文通訓定聲：「發，亦作發。」

集韻：「背，違也，亦作貲。」

古今韻會舉要：「膳，亦作善。」

L.或作

「或作」也是只能描述兩字有異體字關係，例如：

集韻：「剡，利耜也，或作覃。」

龍龕手鑑：「峭，或作峭，峭，正。」

M.或從

「或從」以說文和集韻使用較多，並不能明確描述何種異體字關係，例如：

集韻:「榜，進船也，或从手。」

說文：「𢓊，正行也，从辵，正聲，征，或从彳。」

N.同

「同」只能描述兩字有異體字關係，並不一定是全同異體字，例如：

龍龕手鑑：「弘，同弘。」

廣韻：「弘，同引。」

漢語大字典:「巖，同壧。」

字彙補:「恖，與憂同。」

正字通：「㑹，與嶺同。」

O.籀文

「籀文」描述異體字爲籀文，例如：

說文:「童，男有辠曰奴，奴曰童，女曰妾，从辛，重省聲，𠇍，籀文童。」

字彙補:「畬，籀文畬字。」

R.簡化字

簡化字是漢語大字典的體例，描述繁簡關係的異體字，例如：

漢語大字典：「发，發，髮的簡化字。」

漢語大字典：「欢，歡的簡化字。」

這些體例有些比較明確的表達異體字的關係，包括：古作、古文、今作、俗作、籀作、通、用同和簡化字，其它的則只能知道有異體字關係，包括：同、亦作、也作、或作、也作、或從，都無法判斷是什麼關係，對於我們判斷異體字關係，最好能找到比較明確的關係，而且相同的異體字，不同的字書描述內容也不一定全然相同，例如漢語大字典：「騱，也作奔」，玉篇：「騱，今作奔」，篇海類篇：「騱，與驎同，亦作奔」，所以，每一組異體字，都盡可能同時將多本字書的考證加入，可以幫助我們得到較完整的輪廓。

另外，不同字書使用相同的體例，可能表示的是不同的異體字關係，這種情形主要發生在「通」這個體例，例如干祿字書將異體字分爲正、俗、通，「通」指的是通用字，但是漢語大字典的「通某」指的是通假字，集韻的「通作某」則不一定是通假關係。

對計算機而言，古今字、正俗字、假借字和通假字等都只能提供非常有限的異體字描述，應該將異體字的關係做更詳細的描述，才能提供計算機處理異體字的根據，字書對異體字的描述通常都不會很詳細，但是字書提供了漢字關係的基礎。

(2)異體字的時間面向

此面向描述什麼時間那些漢字有異體字關係，彼此又有那些共同的意義，而我們主要的就是字書和例證。字書對於漢字的使用情形，與例證所提供的資訊不同，前者只能確定若某字爲字書所收，則此字應出現在該字書成書之前，但是此字於字書成書時是否仍在使用，則必需根據例證。特定朝代的文字書寫習慣，會表現在該時代的文獻中，如果能夠找到出更多在該時代的文獻作爲例證，則更有充分的證據可以確定這些文字的使用並非源自於版本相異或作者個人的風格。例如「歸」的初義是女子出嫁，說文：「歸，女嫁也」，引申爲返回和歸依，但在古文獻中多假借爲「饋(贈送)」，書經：「唐叔得禾……王命唐叔歸於東」，又如論語：「陽貨欲見孔子，孔子不見，歸孔子豚」，又詩經：「自牧歸荑，洵美且異」。由這些先秦文獻中，可以得知在秦以前，「饋」有贈送

之義，而說文皆收有「歸」和「饋」，而依廣韻的記載，「饋」是求位切，「歸」依集韻記載亦爲求位切，有了這些證據就可以充分說明在先秦時代，「歸」和「饋」爲部份異體，當作贈送時兩字有異體字關係。

要找出異體字的使用情形，必須從大量的古文獻中尋找，我們可考慮使用數位化的古籍資料庫，但是處理古籍資料時，以今字取代古字是很常見的現象，而且古今皆是如此，檢索數位化的古籍時，對於所出現的文字，很難判斷是否真的是原來的用字，例如以中研院漢籍電子文獻檢索「茶」，可以在史記/列傳/卷九十三/韓信盧綰列傳第三十三找到：「漢十一年秋，陳豨反代地，高祖如邯鄲擊豨兵，燕王綰亦擊其東北。當是時，陳豨使王黃求救匈奴。燕王綰亦使其臣張勝於匈奴，言豨等軍破。張勝至胡，故燕王臧荼」子衍出亡在胡……，但是「荼」字最早應該出現在隋唐之間，但利用資料庫檢索，卻可以在西漢找到「茶」字，主要的原因也可能以今字代古字，因此，如果要找出異體字在不同朝代的使用，利用現有的資料庫雖然有效率，但是所使用的版本是非常關鍵的問題。

因爲兩漢以前的文獻由於能夠留下來的較少，現在我們可以看到的大部份都是隋唐以後所抄寫或刻印，很多文字都已經改用隋唐的習用字，因此，引用兩漢以前的古籍，不一定能夠完全反應先秦和兩漢的文字使用情形，但是傳抄和刻印仍有相當多的古文字被留下來，因此，還是有很好的參考價值，只是我們在引用時，常會懷疑字書所引用的正確性，必須要多持保留的態度，書證的引用必須要找到好的例子，參考來源的版本要經過校對。

目前台灣所編的字典對例證方面並不重視，除了字義的解說外，少數的字典有例句，但是對於我們建立異體字的時間關係沒有任何的幫助，因爲這些例句大部份都是出版社自行造句，最多只能反映現代漢語的使用情形，而沒有其它時代的使用情形。相較之下，康熙字典與漢語大字典則收錄了大量古籍書證。如果將康熙字典與漢語大字典比較，我們可以發現最主要的差異之一，後者有較多的考證資料，更重要的是包括近百年來的許多考古發現，例如敦煌莫高窟藏經的發現、山東銀雀山西漢墓漢簡、馬王堆帛書、睡虎地秦墓竹簡等考古發現，對於我們建立異體字的時間關係非常的重要，漢語大字典也搜集了部份的新證據，例如「復」字，古代借爲「腹(肚子)」，什麼時侯開始「復」有腹義，從睡虎地秦墓竹簡中可以找到使用的例子：早到室即病復痛，因此，我們可以說至少在先秦以前，「復」假借爲「腹」，最晚什麼時侯仍借「復」表「腹」，漢書中可以找到「復」心弘道，惟賢聖兮，所以至少到了東漢「復」仍然借爲「腹」，現在則已經沒有此義。

(3)異體字字義面向

異體字字義面向描述漢字之間有那些意義存在著異體字關係，目前異體字整理的較好的是漢語大字典與教育部的異體字資料庫，但是對於異體字之間究竟有那些共同的意義，並沒有清楚的描述。如果將漢語大字典與教育部的異體字資料庫比較，前者的異體字表雖然簡略，但是可以在字典中找出它們的關係，而且除了有字書的書證還有例證，不過異體之間有那些共同的意義仍然不夠清楚，必須要仔細比較兩個漢字之間的字形結構、字音、本義、引申、通假和假借關係，或考證其它資料，才能確認是否有異體字關係，以及有那些共同意義。

字書中對異體字關係的描述，究竟兩字爲通用或部份異體判斷上並不容易，例如龍龕手鑑：「㑹，古文，音財」，字彙補：「㑹」，古文財字」，即「㑹」和「財」爲古今字，但是這兩字只有「財物」和「財產」基本意義是相同，而財有很多通假義，包括「才能」、「剛剛」、「材料」等，

無法證明「㕣」也有「財」的通假義[趙振鐸 2003]，因此，我們將「財」與「㕣」的關係描述只有在解釋爲「財物」和「財產」爲部份異體字。

我們在分析異體字的共同字義時，同時會描述字義是本義、引申或假借，這些描述對於異體字的關係，是很重要的依據，例如甲乙兩字本義不同，乙字出現的字書較早，而甲字的引申義與乙字的本義相同，乙字現在已經失去本義，而前者的引申義仍在使用中，我們就可以判斷乙爲古字，甲爲今字，共同的意義亦非本義，則可知道甲乙應該不是一字異體。

那些意義是引申義，必須要先確認本義，但是要探求漢字的初義並不容易，因爲很多漢字都是數千年前所創造的，再加上字形經過了幾次重大的改變，更增加了考證的困難。但是掌握漢字的本義非常重要的，因爲本義與字形之間的關係密切，掌握本義則能夠找出詞義引申的擴展和脈絡，以及漢字彼此之間的關係，進一步確定形成異體字的原因。

由於確認漢字的本義有相當的困難，所以同一個漢字其本義存在多種不同的看法，主要的因素端看所發現的證據，以及對這些證據所作的詮釋。我們針對這個問題的解決方法，以說文的解釋爲主，說文以小篆作爲研究本義的基礎，難免會有錯誤，但是到目前爲止，要探求本義，說文仍是非常重要的參考。

漢字除了本義之外，還有很多意義，這些意義大部份都是由本義所引申而來的。在古漢語中，由於溝通的需要，因而產生新的詞彙，這些詞彙除了可以創造新字來表達，另外就是直接透過字義的引申，當然也可以利用產生雙字詞或多字詞的方式表達，不過古漢語使用單字詞較現代漢語普徧，引申是成爲表達概念的重要方式，這種方式最重要的好處是可以控制新字產生的數量。引申義有時候較本義更爲常用，因此失去了本義，就只好另外造字，或假借其它的字表達本義，例如「北」的小篆爲兩人背對背，後來引申爲北方(與中原相背)之後，本義逐漸失去，所以又另外造了「背」字表示本義，「北」與「背」成爲古今字。如果不描述它們的字義是本義、引申或假借，就無法描述這個層面關係。

(4)異體字字音面向

建構異體字關係必需要考慮字音，尤其是通假與假借關係，都是建立在字音相同或相近的基礎。漢字有一字多音和義隨音轉的特性，因此，異體字的共同意義的部份，一定要找出究竟字音爲何，才能確定異體字關係。不過很多異體字音已經改變，因此，不能以現代音作爲判斷的基礎，而必須從中古音和古音著手，才能確認異體字的關係。例如說文：「容，盛也，从宀，谷」。依說文的解釋「容」的本義爲容納盛載，說文：「庸，用也，从用，从庚，庚，更事也」，意思是採用或需要，但釋名：「容，用也」。依釋名的解釋，「庸」與「容」有共同的意義，但是從字形分析來看，兩字皆爲會意字，彼此之間沒有增減形符或改變形符，從金文和小篆字形也找不到共同的形符與演變關係，因此，如果釋名的解釋是正確的，那麼只有一個可能就是假借或是通假，但是「容」的現代音爲ㄖㄨㄥ 2，而「庸」的現代音爲ㄩㄥ，似乎沒有假借或通假的可能，然而依廣韻兩字的中古音皆爲餘封切，又古韻皆爲東韻，由此，可以確定兩者之間應爲假借或通假關係。

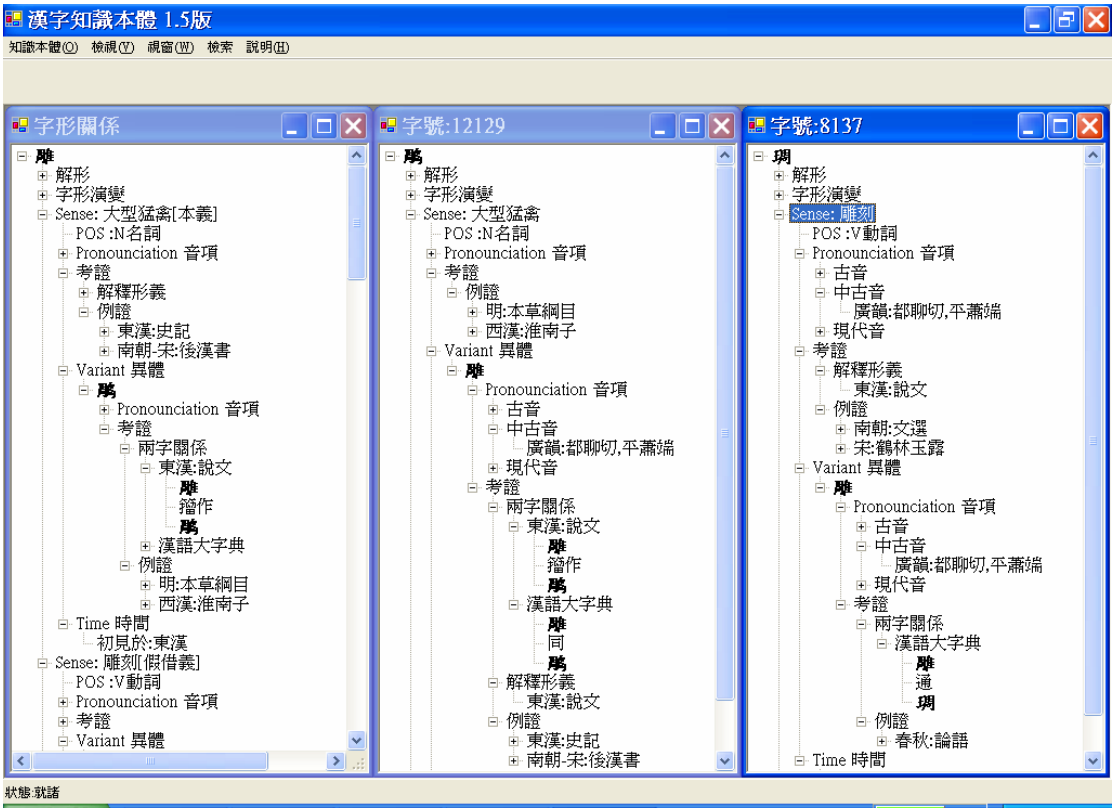
古音反應的是先秦的語音系統，由於時間久遠，研究上較爲困難，因此意見較分歧，如果仔細比較，古音韻部的研究較聲類有較爲一致的研究成果，故我們古音部份只先描述韻部。雖然古韻的研究較古聲類有一致的結果，但是仍然有不同的分部方法，例如顧炎武分爲十部，王念孫分爲二十一部[董同龢 1979]，王力則分爲三十部，漢語大字典也分爲三十部。由於漢語大字典有相

當的權威性，且收字較多，比較容易找出古韻的資料，因此異體字字音知識的部份，我們以漢語大字典的古韻三十部作爲古韻的分部系統。中古音最有代表性的字書是廣韻和集韻，其注音方式採用反切音，因爲同韻或同聲紐可以使用不同的反切下字或上字，確認異體字字音不能只依賴反切上字與下字是否相同，而是應該要依賴聲紐和韻調，因此，我們也將廣韻的反切上字與下字整理，做爲確認異體字的參考。

4. 研究結果與應用

爲了驗證本研究的異體字語境模式，我們以兩年的時間，描述了三千個異體字關係，並且以本研究提出的模型進行表達，以雕爲例說明異體字關係描述的結果(圖二)，對於「雕」與「鵰」的描述是「鵰」爲「雕」的籀文，又根據本研究所對於意符「鳥」與「隹」都表達鳥類的概念，而漢字在使用時經常會更換其同表達相同概念的意符，因此，不僅知道「鵰」爲「雕」的籀文，更可以知道爲什麼它們形成不同的字形結構。另外，再依描述「鵰」與「雕」爲部份異體字，只有當作「大型猛禽」爲異體關係，因爲「雕」有「用彩畫裝飾」和「雕刻」的意義，而「鵰」沒有這些意義，另外「雕」與「鵰」在東漢以前就有「大型猛禽」的意義。

而「雕」與「琱」的關係，根據異體字關係的描述，「雕」假借爲「琱」，且爲有本字假借，而且早在春秋時「雕」即有「雕刻」的意義，「琱」的意符是「玉」，表達的概念是「玉石」，而「雕」的意符是「隹」，表達的概念是「鳥」，兩個字的意符概念差異表示兩字的造字本義應該不同，古韻皆爲幽部，因此，兩字應爲假借關係，雕與琱只有在「雕刻」這個意義有異體關係，因此爲部份異體關係。



圖二　雕的異體字關係

爲了說明本研究對於異體字關係表達的優點，我們將本研究與電腦漢字字形與詞彙整合知識庫比較，該計劃是是數位典藏國家型計劃技術分項計劃－建立 Unicode 漢字異體字表與異體字辭典之相關研究的成果之一，研究動機是因爲 Unicode 的異體字很多，對於計算機的應用造成處理上的困擾，並且嘗試要表達部份異體字關係[袁國華、曾黎明 2005]。此研究是文獻中與本研究最相關的研究，缺乏其它相關的研究，也顯示出異體字關係在計算機長久以來不受重視的事實．我們以「雕」、「鵰」、「琱」三個異體字爲例進行比較，電腦漢字字形與詞彙整合知識庫對於這三個字的關係描述代碼是 H，即它們都是漢語大字典所收的異體字，這個描述計算機只能知道「雕」、「鵰」、「琱」有異體字關係，但是究竟是什麼關係，就沒有任何的描述。相同的異體字，本研究能夠充分的表達在那些意義和什麼時間兩個字形是異體字，以及在當時的字音是什麼，以了解是否有假借關係，本研究的異體字描述架構優於電腦漢字字形與詞彙整合知識庫的異體字描述。

圖三　電腦漢字字形與詞彙整合知識庫的異體字關係[袁國華、曾黎明 2005]

我們將本研究所建立的異體字關係應用在異體字的檢索，且將重點放在部份異體字的檢索。本研究將異體字檢索分爲共時(synchronically)與歷時(diachronically)，共時檢索對於不同年代的文獻，均視爲同一個斷代，而歷時檢索則會根據被檢索文獻的時代產生不同的異體字。歷時異體字檢索考慮的是異體字關係，如果不是異體字關係的差異則不是本研究能夠提供的知識，也不是歷時異體字檢索，例如現代漢語用「衣架」，古漢語用「桁」，如樂府詩集中有「還視桁上無懸衣」，那麼如果檢索詞是「衣架」，就必需以「桁」進行古漢語文獻的檢索，又如現代漢語用「黑馬」，古漢語用「驪」，這些詞彙關係並非異體字。無論是共時檢索或歷時檢索，如果根據本研究所建立的異體字關係知道檢索詞有部份異體字，檢索系統會要求檢索者選擇檢索詞的詞義，根據詞義決定適當的候選異體字，如果被檢索文件的年代是候選異體字使用的時段，就會將後選異體字一起加入檢索字，反之如果並不在異體字的使用時段，就不會加入檢索字。如果檢索詞是雙字詞或多字詞，則會分別找出異體字，加入檢索詞彙後進行檢索。

4.1 共時異體字檢索

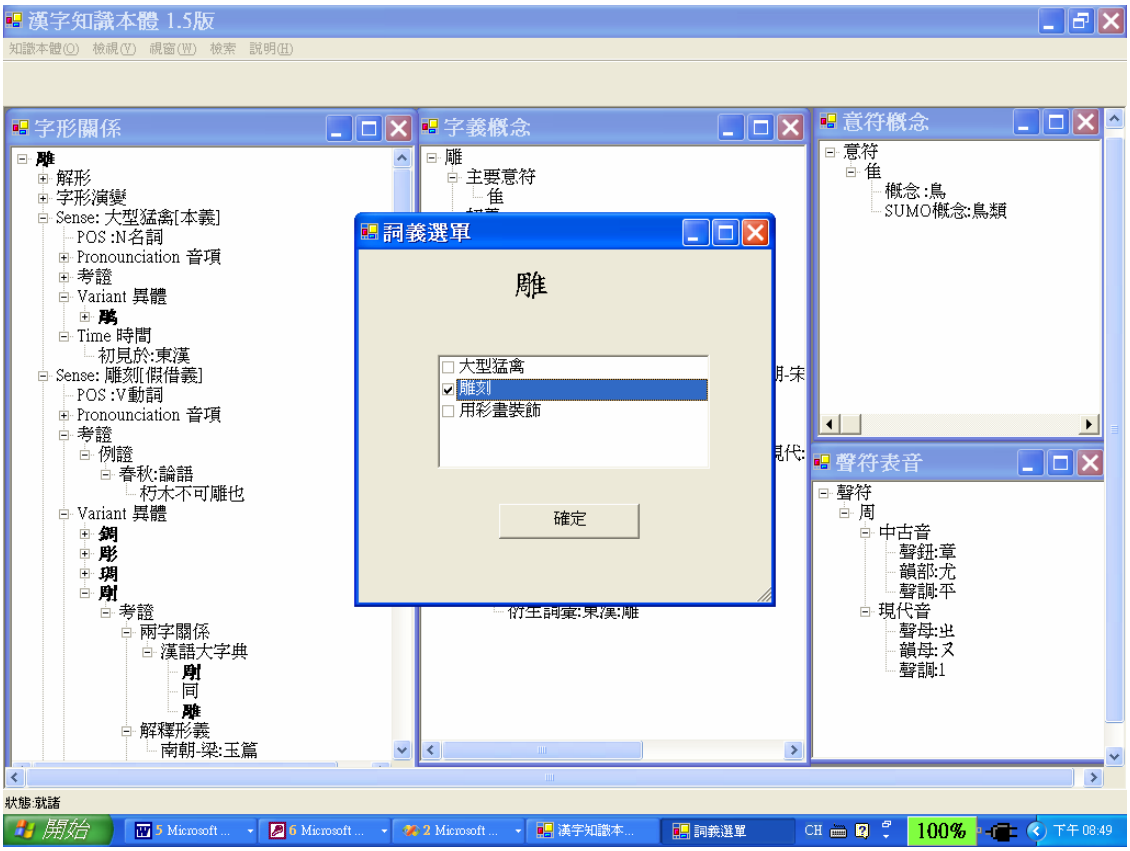
我們以「雕」作爲共時異體字檢索的例子，根據漢語大字典「雕」的意義有:一種大型凶猛的鳥、兇猛、雕刻和用彩繪裝飾等，根據我們所建立的異體字關係描述了「雕」分別在不同的意義有異體字，包括「鵰」、「彫」、「琱」、「剛」、「鍋」，異體字中的「鵰」只有當大型猛禽時與「雕」爲異體，「彫」與「雕」則是在雕刻和用彩繪裝飾是異體字，「琱」、「剛」、「鍋」等皆與「雕」在雕刻的意義爲異體。如果以「雕」進行檢索，由於「雕」有部份異體字，檢索系統會要求確認檢索字義，若以雕刻爲字義，檢索詞會加入異體字「彫」、「琱」、「剛」和「鍋」。表一是作爲雕的異體字檢索的文件集合，這些文件分別用了不同的異體字，來源包括中央研究院漢籍電子文獻的二十五史資料庫、中央研究院平衡語料庫和聯合報聯合知識庫。

檢索結果可以發現出現「雕」、「彫」、「琱」、「剛」和「鍋」的文件都被找到，但是「睽違 12 年白尾海鵰現蹤」、「新校本隋書/紀/卷四　帝紀第四/煬帝下/大業十二年」、「新校本舊唐書/列傳/卷八十八　列傳第三十八」、「新校本宋史/志/卷一百四十九　志第一百二/輿服一/五輅」、「平衡語料庫」等有異體字「鵰」的文件不會被當作通用異體字而被檢索出來，因爲「鵰」與「雕」只在大型猛禽爲異體字。

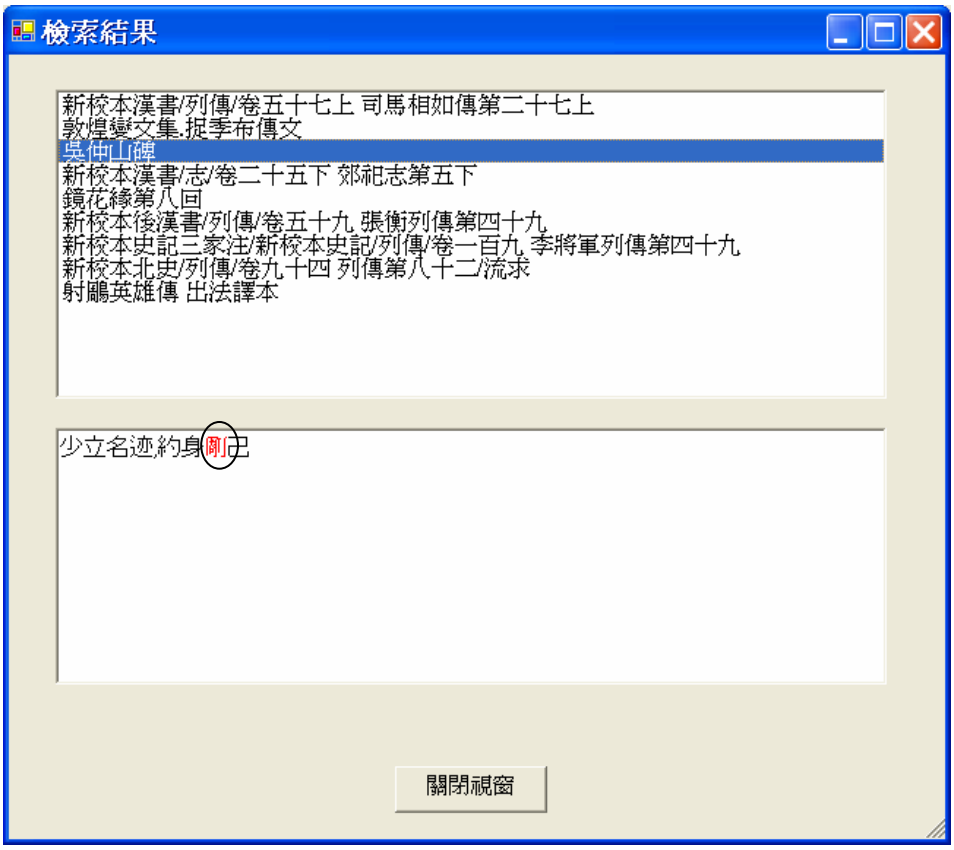
表一　檢索雕的異體字文件集合

編號	來源	標題	文獻部份內容
1	中央研究院 廿五史資料庫	新校本史記三家注/新校本史記/列傳/卷一百九　李將軍列傳第四十九	匈奴大入上郡，天子使中貴人從廣[一]勒習兵擊匈奴。中貴人將騎數十縱，[二]見匈奴三人，與戰。三人還射，[三]傷中貴人，殺其騎且盡。中貴人走廣。廣曰：「是必射雕者也。」
2	中央研究院 廿五史資料庫	新校本漢書/志/卷二十五下　郊祀志第五下	則一梁豐鎬之間周舊居也，固宜有宗廟壇場祭祀之臧。今鼎出於一東，中有刻書曰：『王命尸臣：「官此栒邑，[六]賜爾旂鸞黼黻琱戈。」[七]尸臣拜手稽
3	中央研究院 廿五史資料庫	新校本後漢書/列傳/卷五十九　張衡列傳第四十九	願竭力以守義兮，雖貧窮而不改。執雕虎而試象兮，阽焦原而跟止。[五]庶斯奉以周旋兮，要既死而後已。[六]俗遷渝而事化兮，泯規矩之圜方。
4	中央研究院 廿五史資料庫	新校本北史/列傳/卷九十四　列傳第八十二/流求	所居曰波羅檀洞，塹柵三重，環以流水，樹棘爲藩。王所居舍，其大一十六間，琱刻禽獸。多裸鏤樹，似橘而葉密，條纖如髮之下垂
5	中央研究院 廿五史資料庫	新校本隋書/紀/卷四帝紀第四/煬帝下/大業十二年	二月己未，真臘國遣使貢方物。甲子夜，有二大鳥似鵰，飛入大業殿，止于御幄，至明而去。癸亥，[一七]東海賊盧公暹率众萬餘，保于蒼山
6	中央研究院 廿五史資料庫	新校本舊唐書/列傳/卷八十八　列傳第三十八	辭辯縱橫，音旨明暢，高宗深納之。思謙在憲司，每見王公，未嘗行拜禮。或勸之，答曰：「鵰鶚鷹鸇，豈众禽之偶，奈何設拜以狎之？且耳目之官，固當獨立也。」
7	中央研究院 廿五史資料庫	新校本宋史/志/卷一百四十九　志第一百二/輿服一/五輅	駕六青馬，馬有金面，插鵰羽，鞶纓，攀胸鈴拂，青繡屜，錦包尾。又誕馬二，在輅前，飾同駕馬。餘輅及副輅皆有之。駕士六十四人。金輅色以赤，駕六赤馬
8	中央研究院 廿五史資料庫	新校本元史/列傳/卷一百三十一　列傳第十八/完者都	完者都許以爲副元帥，凡征蠻之事，一以問之。且慮其姦詐莫測，因大獵以耀武，適有一鵰翔空，完者都仰射之，應弦而落，遂大獵，所獲山積，華大悅服
9	漢語大字典	吳仲山碑	少立名迹,約身剛己

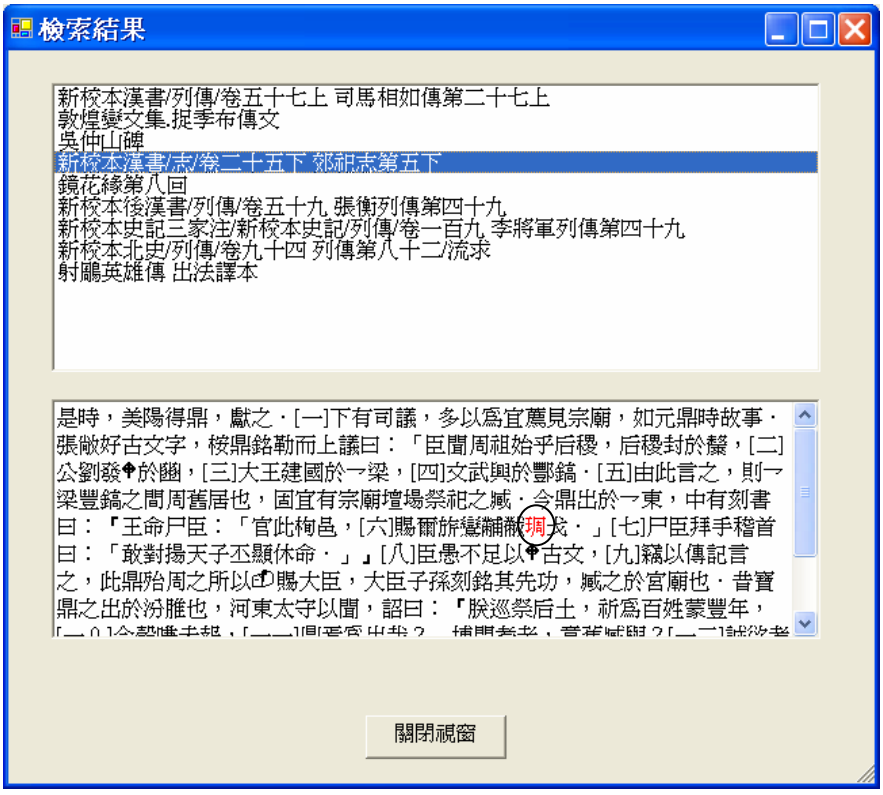
10	漢語大字典	敦煌變文集捉季布傳文	駿馬剚鞍穿鏁甲,旗下依依認得真
11	中央研究院廿五史資料庫	新校本漢書/列傳/卷五十七上　司馬相如傳第二十七上	於是乎乃使剸諸之倫，手格此獸。[一]楚王乃駕馴駮之駟，[二]乘彫玉之輿，[三]靡魚須之橈旃，[四]曳明月之珠旗，[五]建干將之雄戟，[六]左烏號之
12	鏡花緣	鏡花緣第八回	忽見山旁又走出一只小虎，行至山坡，把虎皮揭去，卻是一個美貌少女。身穿白布箭衣，頭上束著白布漁婆巾，臂上跨著一張雕弓。走至大蟲跟前
13	平衡語料庫	平衡語料庫	地域整備法」，其內容乃是想達成一箭三鵰之目的：1依民間活力來擴大國內需求
14	聯合知識庫(20050311)	睽違 12 年 白尾海鵰現蹤	高雄市野鳥學會會員在鳳山水庫發現一隻猛禽類，經與鳥類相關資料比對，確認是台灣罕見的白尾海鵰，鳥會紀錄中，上一次在鳳山水庫發現白尾海鵰，已是12 年前的事
15	聯合知識庫(20041211)	射鵰英雄傳 出法譯本	備受華人愛戴的著名武俠小說作家金庸（查良鏞），其作品「射鵰英雄傳」首次被翻譯爲法文；這部翻譯作品早先已面世。據稱，翻譯者用了三年時間將四冊的「射雕英雄傳」全部翻譯成爲兩冊的法文版



圖四 輸入檢索詞並確認「雕」的意義

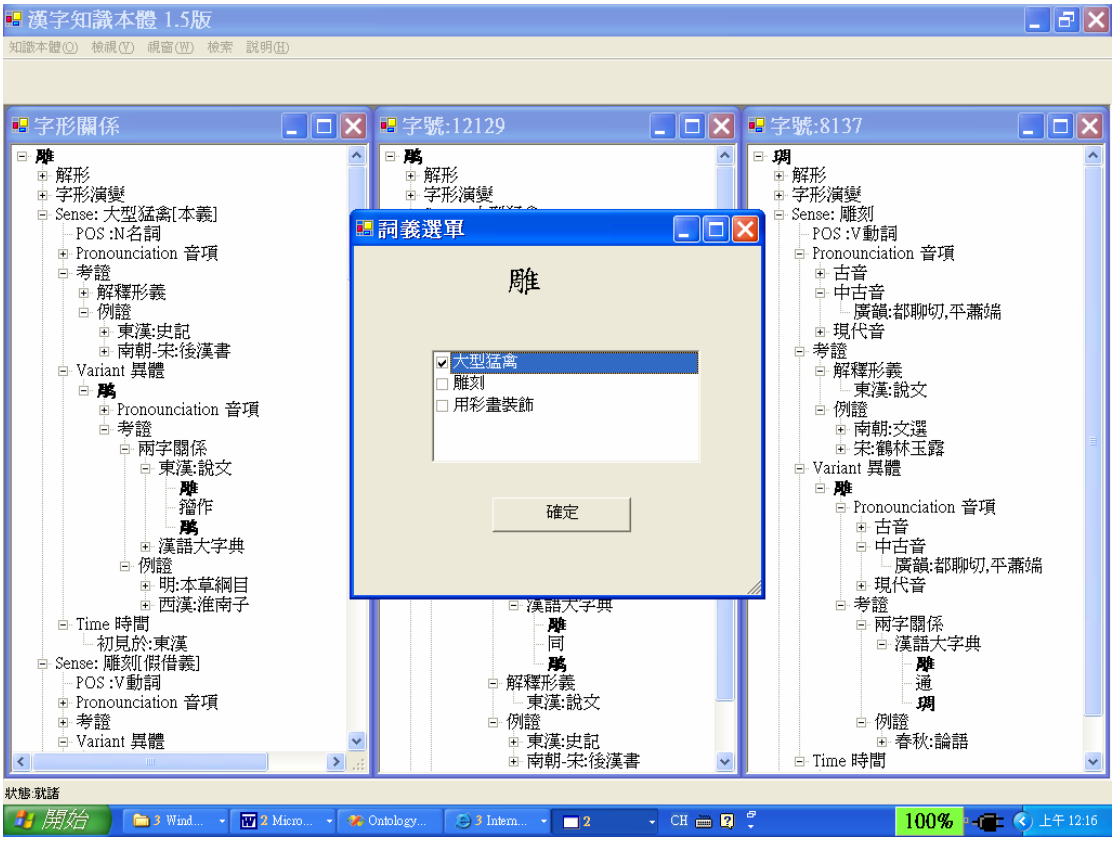


圖五 檢索「雕」(雕刻)可查到部份異體「剛」的文件

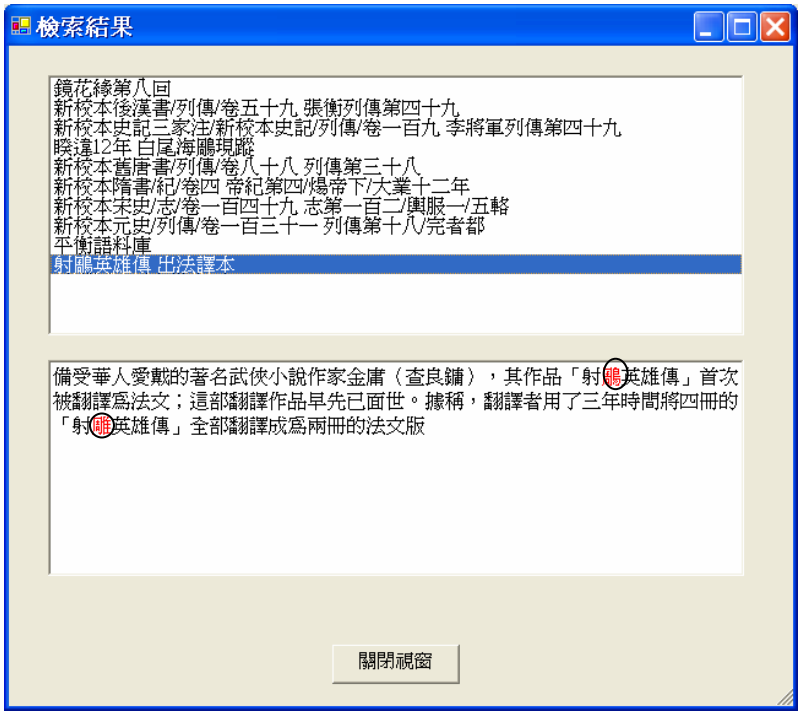


圖六 檢索「雕」(雕刻)可查到部份異體「琱」的文件

如果將檢索的字義條件改爲大型猛禽，異體字中只有「鵰」會被加入檢索詞，而其它的異體字「彫」、「琱」、「剛」和「鋼」都不會被加入，所以檢索的結果沒有出現「彫」、「琱」、「剛」和「鋼」的文件。



圖七　檢索「雕」(大型猛禽)



圖八 檢索「雕」(大型猛禽)可找到出現「鵰」的文件

4.2 歷時異體字檢索

共時異體字檢索並不考慮異體字使用的時間和檢索文件的時間，而歷時異體字檢索則會依據已經建立的異體字關係提供歷代異體字的差異，根據被檢索文件的時間進行檢索詞的修正，例如「獅」本假借「師」，後來才增加意符「犬」造了分化字「獅」以明確其假借義[裘錫圭 1995]，「師」的初義爲軍隊編制單位，且爲最高編成單位[宋子然 2002]，在漢以前就被借假爲「獅」，如漢書：「鉅象、師子、猛犬、大雀之仓食於外囿」，到了北宋新唐書都還是借「師」表示「獅」，而本字「獅」較早出現在玉篇，因此，以「獅子」爲檢索詞，如果被檢索文件是北宋以前的，就應該改以「師子」進行檢索。

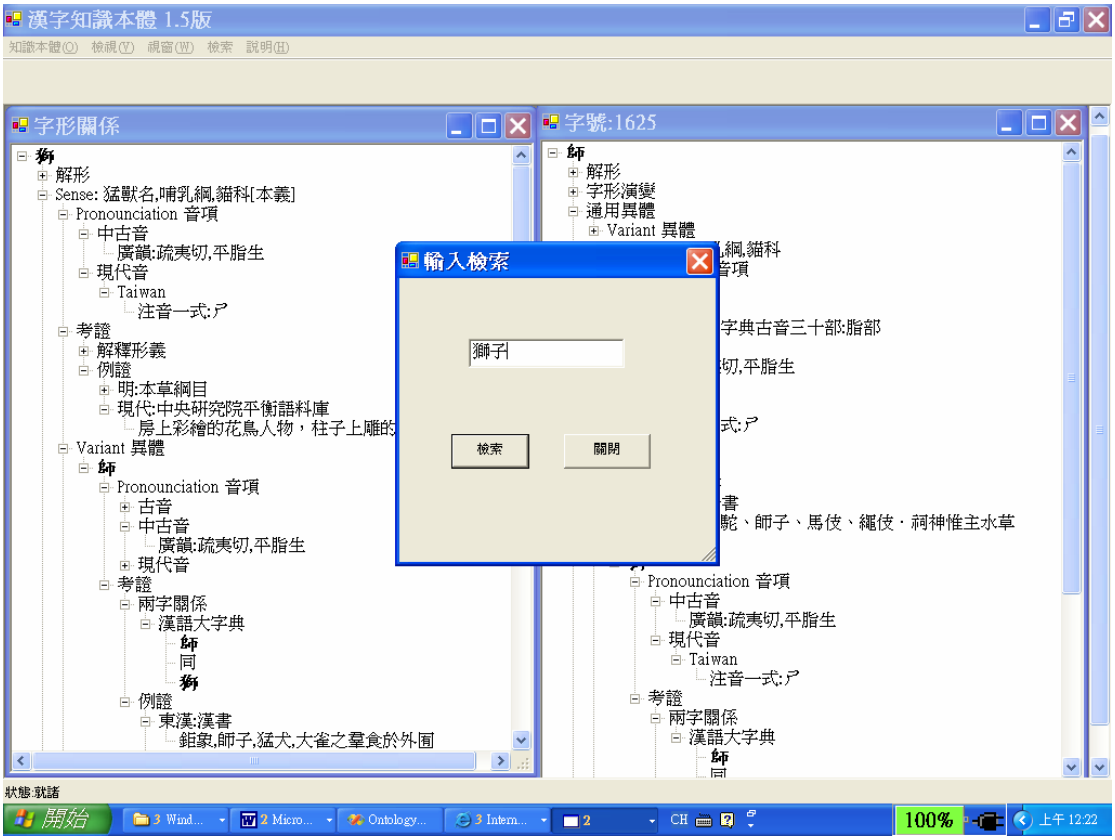
表二 是作爲「獅」的異體字檢索的文件集合，來源是中央研究院漢籍電子文獻的二十五史資料庫、中央研究院平衡語料庫和聯合報聯合知識庫，由於歷時檢索會根據文件的時間和異體字的使用時間決定檢索詞的擴展，因此文件的時間必需被放入計算機，這些文件的時間範圍由東漢至現代。

表二　獅的異體字檢索文件集合

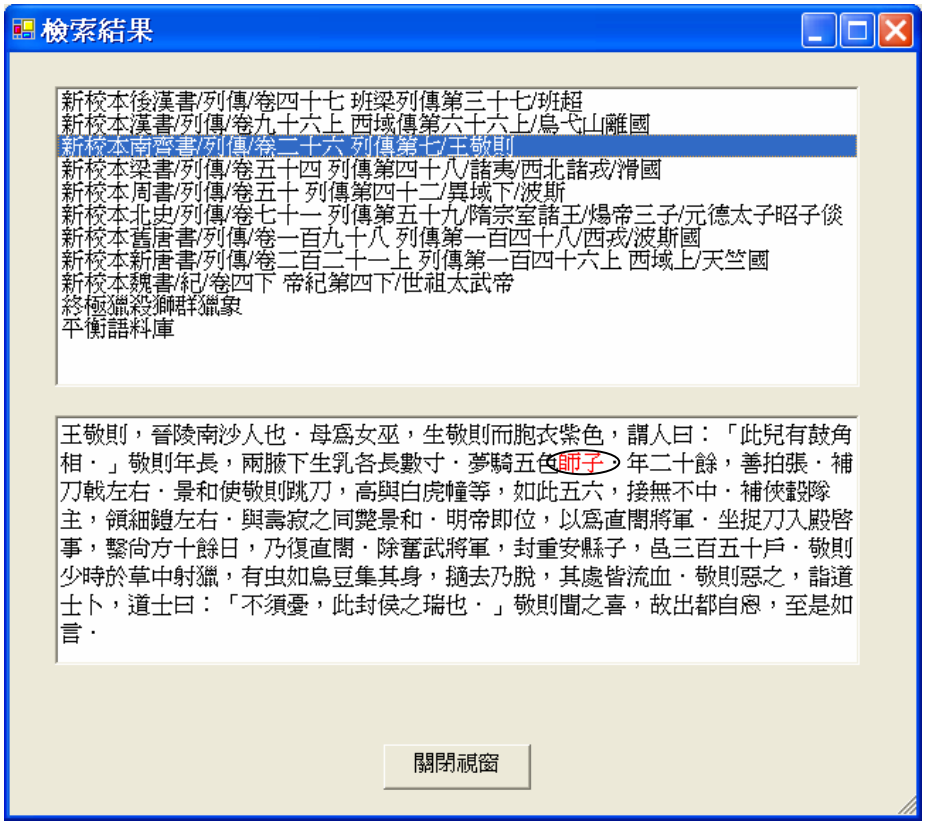
編號	來源	標題	時間	文獻部分內容
1	中央研究院廿五史資料庫	新校本漢書/列傳/卷九十六上 西域傳第六十六上/烏弋山離國	東漢	果菜、食飲、宮室、市列、錢貨、兵器、金珠之屬皆與罽賓同，而有桃拔、師子、犀牛。[二]俗重妄殺。[三]其錢獨文爲人頭。幕爲騎馬。以金
2	中央研究院廿五史資料庫	新校本後漢書/列傳/卷四十七 班梁列傳第三十七/班超	南朝宋	初，月氏嘗助漢擊車師有功，是歲貢奉珍寶、符拔、師子，[一]因求漢公主。超拒還其使，由是怨恨。永元二年，月氏遣其副王謝將兵七萬攻超。超兵少，皆大恐。超譬軍士曰：「月氏兵雖
3	中央研究院廿五史資料庫	新校本魏書/紀/卷四下 帝紀第四下/世祖太武帝	北朝北齊	十有一月辛卯，至于鄒山，劉義隆魯郡太守崔邪利率屬城降。使使者以太牢祀孔子。壬子，次于彭城，遂趨盱眙。頍盾國獻師子一。十有二月丁卯，車駕至淮。詔刈藿葦，汎筏數萬而濟。
4	中央研究院廿五史資料庫	新校本南齊書/列傳/卷二十六 列傳第七/王敬則	南朝齊	王敬則，晉陵南沙人也。母爲女巫，生敬則而胞衣紫色，謂人曰：「此兒有鼓角相。」敬則年長，兩腋下生乳各長數寸。夢騎五色師子。年二十餘，善拍張。補刀戟左右。
5	中央研究院廿五史資料庫	新校本梁書/列傳/卷五十四列傳第四十八/諸夷/西北諸戎/滑國	唐	漢永建元年，八滑從班勇擊北虜有功，勇上八滑爲後部親漢侯。自魏、晉以來，不通中國，至天監十五年，其王厭帶夷栗　始遣使獻方物。普通元年，又遣使獻黃師子、白貂裘、波斯錦等物
6	中央研究院廿五史資料庫	新校本周書/列傳/卷五十 列傳第四十二/異域下/波斯	唐	富室至有數千頭者。又出白象、師子、大鳥卵、珍珠、離珠、頗黎、珊瑚、琥珀、笔璃、馬瑙、水晶、瑟瑟、金、銀、　石、金剛、火齊、鑌鐵、銅、錫、朱沙、水銀、綾、錦、白疊、氍、氈毹
7	中央研究院廿五史資料庫	新校本舊唐書/列傳/卷一百九十八列傳第一百四十八/西戎/波斯國	五代	大驢、師子、白象、珊瑚樹高一二尺、琥珀、車渠、瑪瑙、火珠、玻张、琉璃、無食子、香附子、訶黎勒、胡椒、華撥、石蜜、千年棗、甘露桃
8	中央研究院廿五史資料庫	新校本新唐書/列傳/卷二百二十一上 列傳第一百四十六上 西域上/天竺國	北宋	天竺國，漢身毒國也，或曰摩伽陀，曰婆羅門。去京師九千六百里，都護治所二千八百里，居葱嶺南，幅圓三萬里，分東、西、南、北、中五天竺，皆城邑數百。南天竺瀕海，出師子、豹、龜、
9	中央研究院廿五史資料庫	新校本北史/列傳/卷七十一列傳第五十九/隋宗室諸王/	唐	三歲時，於玄武門弄石師子，文帝與文獻皇后至其所。文帝適患腰痛，舉手馮后，昭因避去，如此者再三。文

		煬帝三子/元德太子昭子倓		帝歎曰：「天生長者，誰復教乎！」由是大奇之。文帝嘗謂曰：「當爲爾娶婦。」應聲而泣。
10	中央研究院廿五史資料庫	新校本宋史/列傳/卷四百八十九列傳第二百四十八/外國五/占城	元	大中祥符三年，國主施離霞離鼻麻底遣使朱浮禮來貢。四年，遣使貢師子，詔畜于苑中。使者留二蠻人以給豢養，上憐其懷土
11	中央研究院平衡語料庫	平衡語料庫	現代	有一隻淘氣的小老鼠，看見獅子睡著了，就從獅子的爪子上，爬到他的鼻子上，把獅子弄醒了
12	聯合知識庫	終極獵殺獅群獵象	現代	由曾獲多項艾美獎的生態影片導演休貝爾夫婦，歷時八年時間追蹤拍攝的紀錄片「終極獵殺」，首度拍攝到獅群獵捕大象的珍貴畫面，也推翻科學界認爲獅子不會攻擊大象的既有觀點

若以「獅子」爲檢索詞，根據本研究建立的異體字關係可以知道自東漢至北宋期間，借「師」表示「獅」，因此，如果被檢索文獻的時間是介於這段期間，檢索詞會擴展爲「師子」，如果不在東漢至北宋期間，只會以「獅子」作爲檢索詞，所以由檢索的結果可以發現北宋以前的文獻都被找出來，北宋以後只有出現「獅子」的文獻被找到，但宋史(元朝)中出現「師子」的文獻卻未被找到，因爲它們不是本研究建立的異體字關係描述「師」作「獅」的時間。



圖九　輸入檢索詞「獅子」



圖十 檢索「獅子」可找到中古漢語文件中的「師子」

5. 結論

本研究的困難在於如何將複雜的異體字關係很有系統的方式加以表達，對於異體字的規範和整理，二千多年來是文字學研究之一，雖然已經累積了很多異體字的知識，但是卻一直無法在計算機表達，我們所提出異體字的模型和建立異體字關係的方法，可以將異體字的關係表達在計算機，不僅能夠表達全同異體字關係，更重要的是能夠表達部份異體字關係，改變過去計算機將異體字視爲全同關係的缺失，與過去異體字的描述方式有顯著的不同，我們分析和表達了三千個異體字，發現本研究提出的異體字描述架構能夠充分的表達異體字的關係。本研究可以表達異體字關係包括古字、今字、正字、俗字、有本字假借、無本字假借等，並且描述異體字在那些詞義、時間、構詞和聲韻可以互相使用。 WordNet 的詞彙關係只建立在詞義，而且沒有考慮語言的變遷，我們不僅分析詞義，還加入時間、構詞和聲韻對異體字關係的影響，同時也要分析字形和字音的變化，如此才能釐清異體字關係，再以系統化的方式描述，因爲需要分析的變數很多，而 WordNet 只由詞類和詞義建立詞彙的關係，本研究的分析更複雜，但是，利用 WordNet 建立的中文詞網不能夠表達漢字的特性，而異體字是非常漢字重要的特性，不過因爲無法利用 WordNet，只能逐字進行分析，因爲需要很長的時間。

異體字是漢語書寫系統的重要特性，不過卻沒有能夠將異體字的關係表達在計算機，造成許多中文資訊處理上的問題，本研究將建立的異體字關係應用在異體字檢索問題，目的是提供檢索系統檢索詞彙的不同形式，但是檢索詞彙在被檢索文件中的詞義，必需要由前後文決定，尤其是多義詞的歧義性，這是資訊檢索的問題，並不是本研究要解決的問題，但是透過本研究的實例，可以呈現異體字關係可能的應用，除此之外，我們還利用本研究成果，設計一個檢索介面，用來檢索 Google 的文件，但是提供異體字的檢索。本研究還可以應用在很多其它問題，例如中文網

域名稱(domain name)和缺字的問題。中文網域名稱目前還沒有完全解決異體字造成無法解析的問題，只能同時註冊多個可能異體字網域名稱，主要原因就是計算機沒有異體字關係的知識，而缺字問題實際上大部份都是缺異體字字形[謝清俊 1996]，如果有異體字字形可以使用，就不一定要使用缺字字形，也不需造字，而造字所產生的交換和檢索問題也同時可以被解決。

本研究只是開始，必定還有其它的異體字關係未能表達，如同 WordNet 仍有許多需要詞彙關係沒有被表達，但是即使如此，運用 WordNet 已經足夠改進很多自然語言處理的問題[Fellbaum 1998][Miller 1995]，我們也開始應用本研究建立的異體字關係，但是還需要更多的應用，才能確認本研究對於異體字關係的描述是否足夠。文字是語言的形式表達，文字整理和分析是非常基礎，因爲它是基礎的研究，其影響非常廣泛，但是願意投入這個基礎研究的很少，希望未來能夠有更多的研究資源投入。

誌謝

感謝吳玲玲教授、謝清俊教授、簡立峰教授、季旭昇教授、高照明教授和何瑁鎧教授給予本研究許多的意見。

參考文獻

1.江藍生、陸尊梧(2004)，簡化字繁體字對照字典，漢語大詞典出版社，第六次印刷。

2.宋子然(2002)，訓詁理論與應用，巴蜀書社，第一版。

3.林樹(1972)，中文電腦基本用字研究，國立交通大學工學院計算與控制學系出版。

4.周亞民(2005)，漢字知識本體－以字爲本的知識結構與其應用示例，國立台灣大學資訊管理學系博士論文。

5.段玉裁(1813)，說文解字注，黎明，1990 印刷。

6.洪成玉(1995)，古今字，北京，語文出版社。

7.洪嘉馡、巫宜靜、黃居仁(2005)，異體字與異體詞詞彙語意初探，第六屆漢語詞彙語意學研討會，廈門。

8.袁國華、曾黎明(2005)，建立 UNICODE 漢字異體字表與異體字辭典之相關研究，數位典藏國家型計劃技術分項計劃，NSC93-2422-H001-018，中央研究院歷史語言研究所。

9.許慎(121)，說文解字，徐鉉校定，北京，中華書局，2004，第一版，第二十二刷。

10.徐中舒(1992)，漢語大字典，建宏出版社。

11.張如瑩、黃居仁(2004)，中央研究院中英雙語知識本體詞網(Sinica BOW)：結合詞網、知識本體與領域標記的詞彙知識庫，ROCLING XVI: Conference on Computational Linguistics and Speech Processing, 台北，9 月 2-3 日。

12.莊德明(1999)，漢字印刷字形的整理，電子古籍中的文字問題研討會，臺北， 6 月 14-16 日。

13.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。

14.章瓊(2004)，現代漢語通用字對應異體字整理，巴蜀書社，第一版。

15.董同龢(1979)，漢語音韻學，文史哲出版社，第七版。

16.董琨(1993)，漢字發展史話，台灣商務。

17.裘錫圭(1995)，文字學概要，臺北，萬卷樓，4 月再版。
18.謝清俊(1996)，從缺字問題談漢字交換碼的重新設計－漢字的字形與編碼，漢字，字碼與資料庫國際研討會，京都，東京，10 月 4 日(修正版 1996 年 12 月 20 日)
19.謝清俊、黃克東(1989)，國字整理小組十年，十二月。
20. Coulmas, F.(2003), *Writing Systems:An Introduction to their Linguistics Analysis*, Cambridge University Press.
21.Fellbaum, C.(1998), *WordNet:an Electronic Lexical Database*, The MIT press.
22.Miller, G. A.(1991) *The Science of Words*, Scientific American Library.
23.Miller, G. A.(1995)"WordNet:A Lexical Database for English", *Communications of the ACM*, Vol.38, No.11, Nov., pp.39-41.