Computational Linguistics and Chinese Language Processing
vol.2, no.1, February 1997, pp. 131-145

建構一個以共時與歷時語言研究爲導向的歷史語料庫

Historical Corpora for Synchronic and Diachronic Linguistics Studies

魏培泉 * 譚樸森 + 劉承慧 ** 黃居仁 * 孫朝奮 ++

Pei-chuan Wei*, P.M. Thompson+, Cheng-hui Liu, Chu-Ren Huang*, Chaofen Sun++**

摘要

Abstract

中央研究院古漢語語料庫是為古漢語語言研究而構建的。這個語料庫不但具有大量的可作為古漢語語法及詞彙研究的電子文獻，而且擁有可以對文獻的語詞進行檢索、統計、搭配的多功能程式。以語法的發展為準，這個語料庫又分作上古漢語、中古漢語、近代漢語等三個次語料庫，相信這樣的劃分對古漢語的共時或歷時的研究都是頗為便益的。

現在上古漢語語料庫中有相當數量的文獻已經依據其原典、作者、文體等等完成了分類及標注的工作，其中又有不少文獻已經做了斷詞，在已斷詞的文獻中又有幾部古籍已完成詞類的標記。這些斷詞以及詞類標記的成果現已構成我們上古漢語詞彙庫的基礎。

The Academia Sinica Ancient Chinese Corpus is designed for linguistic research. The corpus contains ancient texts that are selected because of their usefulness in grammatical and lexical studies, as well as an inspection program with keyword searching, statistics, and collocation functions. The corpus is divided into three sub-corpora according to stages of grammatical developments, thus both synchronic and diachronic studies can be performed on them. Their current sizes are as follows:

* 中央研究院歷史語言所。 E-Mail: Weipc@pluto.ihp.sinica.tw
+ 英國倫敦大學亞非學院。
** 國立中山大學中文系。
++ 美國史丹佛大學。

a. Old Chinese subcorpus (from pre-Qin to Pre-Han): 5,128,068 characters.
b. Middle Chinese subcorpus (from Late Han to the Six Dynasties): 8,101,662 characters.
c. Early Mandarin Chinese subcorpus (from Tang to Ching): 4,406,381 characters.

A great portion of the texts from the Old Chinese subcorpus (4,497,051 characters) has been textually classified and marked-up according to their source books , author, text genre etc. A substantive part (520,794 characters) of the same subcorpus has also been segmented into words, which are in turn given part-of-speech tagging. results of the above two tasks form the basis of our Old Chinese Lexical Database .

關鍵詞：語料庫，詞彙庫，詞類，標記，檢索，古代漢語，中古漢語，近代漢語

keyword: corpus, lexical database, part-of-speech, mark-up, tagging, Old Chinese, Middle Chinese, Early Mandarin Chinese.

我們想藉此報告來敘述我們的古代漢語語料庫及詞庫的發展概況。這語料庫和詞庫的建基是自蔣經國國際交流基金會所支助的兩個計劃開始的，這二個計劃分別是「古代及現代漢語機讀文獻資料之語言分析 -- 文獻語言學之奠基研究」（ 1991.7-1993.6 ）、「上古漢語詞彙之蒐集與詞彙庫之構建」（ 1994.7-1996.6 ）。其中語料庫部分在原計劃中也還只限於蒐集上古漢語語料，如今已經有所擴展，已從上古漢語延伸到中古漢語以及近代漢語了。其所以嘗試建立一個古漢語的電子語料庫與詞彙庫，無非是希望它能作爲古漢語研究的一個良好工具。由於計劃參與者的興趣所偏，我們建構語料庫主要是針對古漢語語法的研究，但我們認爲，如果以此爲基再作進一步的發展，不但對古代的音韻、訓詁的研究也可以有所提昇，對傳統國學諸如版本校勘、辨僞之學也都會有很大的幫助。語料庫基本上包括一批電子的古文獻文檔以及檢索的工具；詞彙庫則包括一批詞項以及附屬於這些詞項的屬性資訊。後者雖然在表面上和前者可以分而觀之，但就理想與實際上，二者的關係還是很密切的。首先，建立一個較完整的斷代語料庫是建立斷代詞庫的先決條件，它可以讓詞彙的蒐集工作更能涵蓋及更趨精準。同時，語料庫中的文檔既可以提供作爲斷詞檔，在載入詞彙資訊時，利用語料來進行搜檢也可用來輔助對資訊的確認；其次，在實際上，我們就是直接在電子文檔上斷詞以蒐集詞彙的，離開語料庫也就沒有今日的詞彙庫；再其次，詞彙的屬性資訊在現階段的作業中也都是附標於文檔上的詞彙的，我們詞庫中這些屬性的資訊就直接從這樣的文檔中抽取的，因此屬性資訊的建立也是離不開電子文檔的。就我們看來，詞彙庫事實上也是語料庫的一個延伸，所以這次的報告也就把詞彙庫的簡介包括了進來。

壹、語料庫建構與功能概說

一、語料庫的內容與規模

就長期而言，我們想建立一個可以用來考察整個漢語語法史的語料庫。這個語料庫可以以漢語語法的發展爲準分爲三期而建立如下的三個次語料庫：

（一）上古漢語語料庫：先秦到西漢

1. 傳世文獻
2. 出土文獻

（二）中古漢語語料庫：東漢六朝

（三）近代漢語語料庫：唐以後

上古漢語語料庫（傳世文獻部分）是計劃中的初步目標，目前已大致完成（其現有語料請參考附錄一）。基於是否能反映句法的考量，這個次語料庫原本只以戰國及西漢的散文文獻爲蒐集目標，後來因爲欲進一步建構詞庫，因此也把《詩》《書》《楚辭》.... 等納入。這個語料庫主要是在中研院史語所的支援下逐步完成的（因爲基金會支助的計劃中輸入經費甚少），而且其中也還有不少語料是來自史語所的漢籍全文資料庫計劃的。出土文獻這一部分，史語所漢簡小組已輸入相當數量的文獻，但因涉及造字整合問題，要納入語料庫中仍頗受限制。

上古漢語的語料較有限，雖然其中有不少是難以斷定作者或著作時期的，但我們是不計真僞儘可能都蒐羅進來，可是這並不代表我們在實際的比較或統計上是不加別擇的。事實上我們自己在其中是作了不少區別，並視需求來選用它們。通常在考察先秦或西漢的語言時，那些信度有疑慮的作品便排除在外。那爲什麼我們要把那些可疑的作品也包括進來呢？其實出發點並不只在於古籍有限，棄之可惜；還在於語言的研究也並不一定只能限制在可信的作品中，那些不可靠的作品也可以用作參考比較的對象，同樣也可以用來深化古漢語的知識。

中古漢語語料庫和近代漢語語料庫雖非上述兩個計劃的對象，但因爲長期研究的需要，也是視機會而隨時蒐羅的。現階段也有了相當數量的語料，其主要來源爲中研院史語所或與其他機構合作所得（預期完成之書單參看附錄二、三）。到目前爲止，在輸入文獻方面和我們有合作關係的其他機構已有中山大學、中正大學、美國史丹福大學、香港中文大學等，其中有部分已進入我們的語料庫中，也有部分是配合合作單位的需求而輸入的，並不大適合放在我們的語料庫中，在此就略而不談了。與中正大學所進行的合作是交換一些六朝的佛經文獻，因爲輸入格式不同，還需要一番整理，所以這一次的交換所得尚未列入我們的語料庫中。

中古漢語語料庫現有的語料規模還算差強人意，語料的主要來源爲史語所，小部分爲和其他機構合作所得，只是校對費時，已上線的部分還不算多。近代漢語語料庫是目前發展最弱的一環，這一部分還需要作相當的努力。我們希望藉著合作而逐步充實其內容。此處劃分的三個次語料庫應該各自擁有相當數量的比較能反映當時語法狀況的古代文獻，不過要達到這個目標實際上並不容易，因爲文獻很有限，因此語料庫中難免要保留一些不甚理想的材料。

由於我們是以語法的發展爲考量來蒐集文獻，因此就得依文獻在反映當代語言的程度來加以別擇去取，也就必然會排除一些非此領域的重要歷史文獻。例如醫書類的古籍如果都蒐集進來，對醫學詞彙及醫療史的研究會很有用，但我們寧可把這個工作讓給別人來做，語料庫中只蒐集了一些比較能反映當期語言的醫書，承襲性高的同類著作就常常置於排除之列。

我們在這裡要爲語料庫目前公開的程度還不盡愜人意說一些話。主要原因是在建構語料庫的過程中，有許多困難需要逐一克服，所以每一語料要等到正式上檔往往需要相當長的時間。此外，目前電子文獻在公開上也還有一些其他的問題，也不是短時間就可以解決的。使用者在看到本語料庫的書單所列古籍相當多（參看附錄），而且已輸入的成績也相當豐厚，但實際上能看到的還很有限，想必不免會以爲我們吝惜分享。其實我們自己現在可以使用的語料也只是那些列爲已進行或已完成的書單中的小部分，其他則都是還在進行而遲早可以放在語料庫中的。因爲這個緣故，希望有機會接觸到本語料庫而深感有興趣者再忍耐一些時候。

原則上，在建立每一個次語料庫時，應還須按照著作的時地和文體，來選擇文獻以構建之，也就是說我們的目標是要建立一個古漢語的平衡語料庫。我們應可以隨時把這些文獻區劃出各種次類，並以之爲條件，來作進一步的整理與研究。只不過因古文獻保存的限制以及著者往往不明，這個要求在實際上並不容易滿足。

即便在古代文獻中有不少是難以精確繫年的，爲了可以對古漢語的共時與歷時獲致較精確的認識，在上述三個時期中還是應該把語料再進一步細分階段的，也就是應還可以作進一步的斷代，且每一代應各擁有足堪體現該期語法諸面的語料量，既可以作共時的描寫，同時也足以作異時的比較。我們在建構語料庫時，是隨時注意到這點的。例如在上古漢語中，儘管有不少作品著作時期不明，我們還是儘量選擇較可靠的文獻依其語法表現再細分階段。至少從計劃初始，我們就已把其中時間相對較確定的部分語料分成戰國以及西漢兩個時期，並把這種分期作爲斷詞以及標詞類的主要基準。因爲我們相信，要能發掘語言異時發展的真相，也是得建基在每一時代良好的共時描寫上。依斷代原則來建立語料庫，才能爲每一時代的共時描寫以及異代演變的考察奠定一個良好的研究基礎。

因爲我們的計劃都是先鎖定上古漢語，而且目前也只有這部分比較有些具體成績可說，所以以下所述仍以上古漢語的發展情況爲主。

二、語料庫的結構化

（一）、初步的構想

在我們初步的構想中，語料庫應至少可以把原典中所有對語言研究有用的結構及相關訊息呈現出來，這些訊息包括原典、文體、作者、年代、地區、主題等。我們可以以此爲條件來整理、搜尋以及統計語料庫中的語法與詞彙。

由於是以探討古代語言的真貌爲目的，所以我們對古文獻就得依其時間地域及其反映語言的程度來加以分門別類。如我們曾依據時期、主要的文體形式、傳承方式、可信度等等方面的訊息來將上古漢語的原典分作幾個類群，一個結果是把先秦（目前只限於戰國）及西漢分爲兩時期，裡頭還分出核心與非核心兩部分，底下又分出基準與參較兩部分（參較部份爲真僞或時間難定者）。由於我們是以原典中一篇篇的文章爲基本單位，因而把這些訊息附加在每一篇文章上更是免不了的工作，目前每一篇文章的出典以及文體訊息的標示在實際發展上已有相當具體的成績。

語料的結構需要在電子文檔上加上標記使得電腦可以據以檢索，我們語料庫現有的重要的結構標記系統大致如下：

A 、原始篇章結構之標記

這是根據所採用典籍現有的結構訊息而擬的標記系統，這些訊息包括原典、部、卷、篇、章節、段落以及頁碼、標題以及注文間的相互配置關係等訊息。

B 、文章結構之標記

這是自訂的文章結構標記。我們認爲，應以一篇篇獨立完整的文章作爲研究的基礎單位，但古籍的編排方式常有和現代所謂的文章單位不相吻合的，所以有時需要將文章的單位重新劃定。這些重新劃定的文章單位或大或小於原有的篇章。例如《莊子》內七篇中的每一篇依我們的標準都可以各分作數篇文章。不過我們的文章結構並非只是把原始篇章結構的終端節點再往下更細分節點而已，我們是希望利用這個結構來使得原典的組織表達得更有意義，例如所謂《墨子》十論是許多文章的集合體，原典傳統的篇卷組織並無此此訊息，但它終究是一個可以獨立而且頗具意義的單位。我們把本文所謂的文章結構視爲一種近於現代篇章組織而有別於傳統原典組織的結構，也把上述這種意義的單位包含在文章結構內，並把它作爲原典與文章單位間的一個節點。

文章結構所以不合併到原始篇章結構中，是因爲兩者互有扞格，難以統一，否則我們也沒有必要如此費事的分作兩套。例如有的文章跨卷，要維持傳統的分卷，就難以在同一結構下給與這類文章一個合適的位置。又如《韓非子．說林》跨七、八兩卷，而卷七又有《喻老》一篇，我們如果想保持傳統篇卷的訊息，要讓每卷各擁有一個節點，就很難同時爲「說林」立定一個節點，理所當然的原來應受「說林」管轄的各文章也就無從統合於其下了。又譬如，我們希望把《墨子》的十論建爲一個次目錄，作爲一個有意義的研究單

位，但它是跨了許多卷的，這和傳統的篇卷組織也很難在一個結構下並存。

我們劃定的文章既是一基本單位，就應該加上單位標記。文章單位一經加上標記確認，附屬於這些文章的種種屬性資訊才能接著附添上去。

C 、其他原典及文章資訊之標記

要將文章視爲研究的基本單位，又要按其作者、年代、地區、文體、主題等來加以整理、比較的話，就需要把這些屬性資訊加到我們劃分出來的每一篇文章上。

（二）、改進現有建檔方式的展望

我們現在建檔的方式是在電子文檔上加標記，並利用一套文獻分析建檔程式來把附加了結構標記的文檔轉換爲語料庫的內部結構檔，使得語料可以建在線上，既可供檢索，也可作爲斷詞之用。不過這個辦法既較不人性化，同時也很難讓研究者有自行設定需求的空間。目前，中研院資訊所謝清俊教授所領導的研究群已發展出一種可以直接在視窗環境上以連結的方式來構建文檔關係的工具（這部分可參考謝清俊【 1997 】第五節），這種建檔方式不需如我們目前得在文檔上加上繁複而且易出錯的結構標記，建檔者也不必先得學會一套特別的標記符號系統才可以進行建檔，也可以依建檔者個人的需求隨時自建結構與節點屬性。將來時機一旦成熟，我們的建檔將會改在這種新環境中進行。

三、語料庫的用戶介面和工具與對古代語法分析的功用

有了內容充足的電子語料庫，再加上一些爲語言研究而設計的工具，對古漢語語言的研究即可有莫大的助益。使用者透過用戶介面可以選擇各種工具，以達到查詢、引得、統計等功能，程式依其功能大致可分爲如下數類：

（一）、目錄與正文查詢

可以查詢文獻的路徑或結構以及頁碼等。語料庫中的目錄又分爲原典的原始篇章目錄以及自訂的文章目錄，在實際檢索中可以隨時在兩個目錄系統間切換或者在不同的目錄層次上下移動，並且可以隨時選擇所要檢索的範圍。

（二）、引得檢索

這個工具可以在語料庫中搜尋特定的詞彙或例句。由於是以語言研究爲目的，計劃目前採用的語料庫檢索系統的檢視方式是採用「語境伴隨關鍵詞」（ key word in context; KWIC ），主要是由中研院資訊所發展出來的，不過這個工具中還有一個指示詞條在原典位置的功能，卻是靠中研院計算中心的幫助而完成的。目前本工具仍在繼續進行發展更有助於語法研究的其他功能。這個檢索工具現今也已合併在史語所的檢索工具中，使用者可選取該工具中的「格式」項，其中的「引得檢索」便是本計劃所開發出來的。

我們的檢索程式一般檢得的關鍵詞在文獻中每出現一次在畫面上就占一行，其前後顯

示上下文若干字（字數可自定），同時又可看到每一條的出典以及頁碼的訊息。若是覺得哪一條還不夠清楚，仍可在那一條上直接按鍵以查詢更完整的上下文及所出篇章。以上說的是畫面呈現的方式，其檢索方式則大略可分爲如下數種，其對語法研究的功用則略舉一二例於後。

A 、一般詞項檢索

可一次檢索一詞或多詞，並依需要來選擇如何排序。此外，檢索時也可以排除關鍵詞左邊或右邊不欲包括的詞。

1. 右邊排序：從關健詞右邊第一字起按筆劃繁簡依次排序。
   例之一：可用來觀察動詞的及物性或搭配關係，也可藉以對動詞進行小類的劃分。
   例之二：可用來考察介詞省賓的狀況。
2. 左邊排序：從關健詞左邊第一字起按筆劃依次排序。
   例之一：可藉以觀察動詞修飾成分的分布狀況。
   例之二：可用來考察代詞「之」前可能出現的動詞，對動詞小類的劃分也有幫助。
3. 依文排序：按詞項在原典出現的次序排列。
   例之一：可用以觀察詞項在一部或多部原典中內部分布的狀況，可以藉此觀察該詞項在不同作者、時期或文體上表現的異同。

B 、不連續詞項的檢索：可以檢索在語段中並不連續但具互倚關係的語詞。
例之一：如可用來考察被動式「爲 ... 所」式的出現情況。

例之二：如可用來考察連詞的互倚關係。

C 、重疊字檢索：這項功能可以把檢索範圍中的 AA 、 AABB 、 ABAB 等幾種疊字形式都檢索出來。
例之一：古漢語的疊字狀詞可以很容易藉 AA 的檢索而尋得。

例之二：古漢語中以疊字來表示逐指的用例也可以很容易藉 AA 的檢索而弄清楚。

D 、詞類檢索

文檔上的詞彙在加上詞類的訊息以後，我們也可以藉以考察詞項本身詞類的分布狀況，也可以藉以觀察和此詞相搭配的詞在詞類上的分布狀況爲何。這個加詞類訊息的工作也是我們目前正在進行的一個工作。

（三）、字數及字頻的統計

這個功能可以把古漢語某個階段中各個單詞在整個語言中所占的地位以量化的方式呈現出來，把這種統計用之於異時的比較對語言變化的考察可能會有很大的幫助。

（四）、語詞搭配 (collocation)

這個工具可顯示語詞的共現關係。字詞可能的搭配狀況對詞彙的研究頗有助益，一個好的語料庫是求出這個結果的重要基礎。我們現有的工具已可做出語詞共現的統計表，而且不論順序或逆序的搭配都不成問題。

利用以上數項工具來研究斷代或歷史語法是我們的主要目標。將來也可延伸到對其他人文學的研究，例如對傳統國學諸如訓詁、版本校勘、辨僞等即會有很大的幫助。

貳、語料上詞資訊的標注與詞庫的構建

我們詞庫長期的目標是將可能獲致的各種語言資訊建在古代詞彙資料庫上，這些資訊包括古代漢語每個字或詞的古今音、語法、語義以及其他古代文獻記載的相關訓詁、語法用例及出處、音韻、頻率等知識。希望不但可以利用詞庫來研究古代的詞彙並建立斷代的詞庫以及詞典，而且可提供作爲多學科的研究。不過此計劃限於時間與人力，現在對詞庫各詞項屬性資訊的賦予就只限於詞構和詞類兩項。以下是我們工作現今之概況與預期在近期內達成的事項。

一、詞項的蒐集與語詞的內部結構

目前已進行斷詞的主要是在上古漢語部分。由於上古漢語缺乏良好的現成詞典，我們無法藉助它來蒐集詞項；而且基於詞彙最好直接自語料抽取的信念，我們採取在電子文檔上直接斷詞以蒐集詞彙的方式。進行的步驟是：自定一定的標準，在電子原典上直接進行半人工化的斷詞，同時賦予每一語詞的構詞訊息。這個斷詞的成果可由電腦程式依我們研究的需求自動整理出來，且可將之轉入一個有特定架構的詞庫中（但還預留載入別類新資訊的空間）。

由於漢語歷史語言學至目前爲止的研究對於古漢語所謂詞的界定仍相當模糊，所以我們必須在建立詞彙庫的同時從語料的分析中汲取經驗，以歸納來檢測預設的準則。因此這個蒐集詞彙的過程同時也是對古漢語的一個研究過程。

上古漢語除了有一些不可分解的複音詞外，無疑還有些可確定爲詞的複合結構，可是還有許多由詞素結合的複合結構很難判斷爲詞或詞組，主要原因是上古漢語詞的界線甚難取得一個憑準。爲了作業以及將來共時比較的方便，所以我們採取較寬的斷詞標準，所標選的不一定是嚴格定義的詞，它也可能還是詞組，我們也只以「語詞」而不以「詞」來稱呼這些標選出來的詞彙。從另一個角度看，這樣做可以預留更大的空間以供更進一步的學術研究。這些語詞在構詞上的分類細節可參考我們的斷詞手冊上的說明。

對於上古漢語雙字以上的複音詞或語詞，我們把它分作六類以供斷詞（如下），在實

際的斷詞時就依此分類來給各詞項定類。

1. 並列語詞、2. 偏正語詞、3. 聯綿詞（是不可分析的偶字詞，主要是雙聲疊韻詞）、4. 複音詞（包括加詞綴或疊字之語詞，類似派生詞）、5. 專名、6. 其他結構（不在以上諸類中的暫歸此類）。

我們所標記的另有述賓一類，不過在上古漢語中此類鮮少爲詞，這一類只是臨時性的標記，目前標記的目的是作爲特定研究之用，所以並不正式列爲一類，統計詞彙時就把它排除在外。除了專名以外，基本上這些類是以語詞的結構來區別的。專名也還包括單字的，其所以和這些類混在一起，是因爲專名在程式的自動計算上是屬於困難的一部分，最好能趁著逐文標記時先行標示清楚，以免將來統計時造成許多錯誤。至於不在上述六類中的，我們就歸作單字詞。

至於中古漢語部分，也進行了少量的語詞結構標記。爲了適應新的需求，結構的種類增加「音譯詞」一項。

現代漢語的述補式動詞暫時歸在並列式中，一方面是因爲它至少是動詞的連用，另一方面是因爲在上古漢語中，它即使不能算是並列，但也不一定是述補而可能是偏正。它由偏正轉爲述補的時間並不易斷定（可能是中古，但實際轉變的過程與時間則不清楚），暫時把它歸在並列式中，一方面既可避免在無可靠的証據下任憑主觀的依違而歸類到偏正或述補；另一方面也因爲如果把它一體放在並列式中，也就可以方便將來作較完整的考察。

目前已有人利用我們所標選的並列式來做研究而發表論文的。以下試舉兩例以說明斷詞的結果對古漢語研究所帶來的便利性（分別爲章明德【 1995 】、劉承慧【 1993 】）：

其一例爲較全面的考察先秦散文中並列結構的組成語素彼此在語義和語音上的關係（其中語音關係特別著墨在兩個語素的相對語序和聲、韻、調的相關性），方法是抽取已斷過詞的先秦語料中所有的並列語詞來加以分析整理。

又一例爲述補式的研究，利用現代漢語詞彙庫中的述補式來和上古漢語並列語詞中的相同形式作比較，以考察上古漢語中此式的性質與地位。

二、詞項屬性資訊的綴加

目前只進行如下兩項：

（一）、詞構：見上述。

（二）、詞類

古漢語的詞類是爭議頗大的一個論題，現在則是到了可以利用語料庫來幫助我們重新思考的時候了。爲達此目的，我們設計了一個詞類標記手冊，目前已經有論語、孟子、左

傳、莊子、韓非子等五部書做了詞類標注的工作。

由於詞類的爭議很大，所以我們的標類手冊並不妄圖構擬一個體系龐大而又精密完整的系統，而根據兩個原則來設計：其一，設計用來標注文檔的詞類系統儘可能保留重要的分別，以便留下較寬廣的空間來供學者作進一步的研究；其二，標類作業要求既簡易而又能客觀一致，儘量避免易涉主觀的分類。該標類手冊目前仍在作局部修改以及添加更詳細的說明，該設計的詳細內容請參考計劃現有的《詞類標注手冊》。

三、詞頻統計

我們較早的統計程式主要是以字爲本，如總字數、平均句長、段落平均字數、各平均值的標準差，以及使用字集、字頻等。現階段還完成了一些詞的統計程式，可以計算並表列古籍中各種語詞的訊息，如某一個或多個構詞類的詞頻、一部或多部古籍的詞數及詞頻等，表列時也還可以依不同的選擇作不同的排序。

四、從屬詞彙庫

基於詞彙會因文體、主題、時代、作者的不同而呈現不同的面貌，因此在我們的構想中，詞彙依據不同的條件來分類整理是必須要做的事情。也就是說在整個上古漢語詞彙庫底下，我們打算再以主題、原典等爲分類標準，建立一些從屬的詞彙庫（ sub-lexica ）。我們不但預計對一些重要的原典做各別的以及綜合的詞彙排比統計，並且也計劃根據一定的條件對同類的一些古典做同樣的工作。例如可做先秦詞彙庫、史記詞彙庫、古醫書詞彙庫等。將來隨著語料庫的擴增，我們還要再構建一系列斷代的詞庫，以作爲漢語詞彙史研究的一項利器。

計劃參與者相關著作

劉承慧，1993，《述補式的早期發展》，國科會 1992-1993 年度計劃報告。

劉承慧，1994，《先秦動詞的性質與類別》，中山大學中國文學系所第四十八次學術討論會。

劉承慧，1995，《文獻語料庫與漢語史研究》，國科會 1994-1995 年度計劃報告。

劉承慧，1995，《先秦漢語實詞的分類問題》，國科會 1994-1995 年度計劃報告。

劉承慧，1996，《先秦實詞與句型》，中國文學的多層面探討國際學術會議，臺灣大學。

魏培泉，1994，《先秦主謂間的連詞「之」的分布與演變》，國際中國先秦語法研討會論文。

謝清俊， 1997 ，《中央研究院古籍全文資料庫的發展概要》，文見本刊同期。

章明德， 1995 ，《先秦漢語詞彙並列結構的研究》，國立政治大學碩士論文。

周玫慧， 1996 ，《上古漢語疑問句研究》，國立臺灣大學碩士論文。

Thompson, P.M. （譚樸森） 1991. Chinese text input and corpus linguistics, in V.H. Mair and Y. Liu (eds). Characters and Computers, IOS Press.

網址： http://www.sinica.edu.tw/ftms-bin/ftmsw3
目前在這個網址上的語料因爲改爲視窗版，還不及將我們的檢索功能都轉換過來。

附錄一、上古漢語語料庫（傳世文獻部分）目錄

本語料庫下列這些典籍凡未標示尚未輸入者都已輸入，但小部分仍在校對中。有的著作要放在這個語料庫或者中古漢語語料庫頗難決定。其中有些典籍爲僞作，有些是真僞難辨或著作時間難定的，但因仍是可以研究的對象，所以我們並沒有把它排除在外。例如有的明知是僞，斟酌利弊，仍然列入此間，如《列子》《尚書孔傳》。也有雖知是東漢以後的著作，但因記錄的是先秦或西漢的事情，且其中蒐集的材料不能排除有可能直錄自過去的文獻，所以暫時仍放在上古漢語語料庫中，如《吳越春秋》《孔子家語》《孔叢子》《漢書》《前漢紀》《西京雜記》等。

爲了讓語料庫能配合古漢語語言研究者的便利，我們把書籍的排列作了一番安排，有助於減少選取的時間。這個排序以及類聚的主要考量點大致是這些典籍的時代、真僞、文體，多少也還考慮了傳統的類舉習慣。大體而言，從《左傳》到《呂氏春秋》，可作爲先秦語言研究的主體；從《史記》到《列女傳》，是西漢語言研究的主體，只是其材料及語言多有承襲先秦的，在使用時還需要注意如何別擇。以《史記》爲例，其中記錄秦漢時期的部分和記錄先秦的部分就可以先分開來，再來互爲比較其語言的異同。

尚書，詩經，周易，儀禮，周禮，禮記，春秋公羊傳，春秋穀梁傳，春秋左氏傳，國語，戰國策，論語，孟子，墨子，莊子，荀子，韓非子，呂氏春秋，老子，商君書，管子，晏子春秋，孫子兵法，大戴禮記，韓詩外傳，吳子，尉繚子，太公六韜，司馬法，愼子，文子，關尹子，鶡冠子，公孫龍子，鄧析子，尹文子，鬼谷子，燕丹子，列子，孝經，爾雅，周髀算經，九章算術，黃帝內經素問，黃帝內經靈樞，難經，本草經（尚未輸入），古本竹書紀年，逸周書，穆天子傳，孔子家語，孔叢子，吳越春秋，越絕書，山海經，楚辭

史記，漢書，賈誼新書，桓譚新論（尚未輸入），陸賈新語，春秋繁露，淮南子，新序，說苑，列女傳，鹽鐵論，法言，西京雜記，前漢紀，方言，尚書孔傳，詩經毛傳（此二種尚未上檔）

附錄二、中古漢語語料庫目錄

以下書單包括已輸入以及期望能在三五年內完成輸入的著作，但這並不表示我們想要完成的中古漢語語料庫就僅止於此。其中部分尚未輸入的口語化程度較弱，現在我們仍在考慮是否一定要完成。雖然其中有的爲散文，有的是含有口語詞彙的韻文，但就反映當時語言而言往往還不如佛經，所以將來也可能放棄收入而尋求再多蒐集其他佛經。中古的文獻自然很多，但緣於中古以後承襲文言的習慣以及部分作者的藻飾之好，有許多文獻在擬定書單時就不列入考慮之內。至於注解這一類材料在取作語料上還有若干問題，所以我們尚未處理。其中部分可以取自史語所漢籍全文資料庫計劃輸入的《十三經注疏》，但該文檔未作標點，也是暫時擱置的理由之一。

這個次語料庫分作一般及佛經兩部分，理由是佛經數量龐大而且也已有個現成的編目，另列比較容易尋取；另外佛經以外的作品能反映口語的也比較有限，藉此分別也可以比較佛經與非佛經語言上的異同。佛經部分只列出目前已輸入者，但將來還可望繼續擴展下去。書單中也有些作品雜採了唐代及六朝的材料，如《古小說鉤沉》《太平廣記》《法苑珠林》之類，原則上固然可以兼列入近代漢語語料庫中，但我們目前暫時不作重複列舉。在只列一次的條件下，我們寧可把它放到中古漢語語料庫中，這是因爲這些書中的唐代作品對唐代語言的研究所助有限，而那些可定作六朝作品的對六朝語言的研究卻是很重要的參考資料。至於《全上古三代秦漢三國六朝文》中的三國六朝部分比起秦漢以前，可以補充我們不足的材料也比較多，所以也只列在這個次語料庫中。

（一般）

爲助尋檢之便，以下書單依其性質粗分爲四個部分。其中劃底線的尚未輸入。

（一）、論衡，白虎通，風俗通義，潛夫論，申鑒，中論，獨斷，忠經，釋名，傷寒論，金匱要略，太平經，典論，博物志，抱朴子，神仙傳，世說新語，齊民要術，搜神記，洛陽伽藍記，荊楚歲時記，華陽國志，顏氏家訓，真誥，毛詩草木鳥獸蟲魚疏，金樓子，曹操集，陶淵明集，庾子山集

（二）、後漢書，三國志，晉書，宋書，南齊書，梁書，陳書，魏書，周書，北齊書，南史，北史，隋書，後漢紀，東觀漢紀，八家後漢書，九家舊晉書，

（三）、樂府詩集，全上古三代秦漢三國六朝文，古小說鉤沉（魯迅），太平廣記

（四）、詩經鄭箋，儀禮鄭玄注，周禮鄭玄注，禮記鄭玄注，論語何晏集解，孟子趙岐注，楚辭王逸注，公羊傳何休注，呂氏春秋高誘注，淮南子高誘注，戰國策高誘注，國語韋昭注，周易王弼注，老子王弼注，列子張湛注，左傳杜預注，穀梁傳范寧注，莊子郭象注，爾雅郭璞注，方言郭璞注，山海經郭璞注，文選注，水經注，論語皇侃集解

（佛經部分（大正藏））

東漢到西晉的作品中有少數爲作者及著作年代較可疑的，如果以（梁）僧佑的《出三藏記集》是否有登錄爲準，則這類作品只占已輸入作品中很小的一個部分，但使用者在使用這個語料庫時仍請善自別

擇。

(東漢)迦葉摩騰共法蘭： 784. 四十二章經

(東漢)安世高： 13. 長阿含十報法經， 14. 佛說人本欲生經， 31. 佛說一切流攝守因經， 32. 佛說四諦經， 36. 佛說本相猗致經， 48. 佛說是法非經， 57. 佛說漏分布經， 91. 佛說婆羅門子命終愛念不離經， 92. 佛說十支居士八城人經， 98. 佛說普法義經， 105. 五陰譬喻經， 109. 佛說轉法輪經， 112. 佛說八 正道經， 131. 佛說婆羅門避死經， 140. 阿那邸化七子經， 149. 佛說阿難同學經， 150A. 佛說七處三觀經， 150B. 佛說九橫經， 151. 佛說阿含正行經， 167. 佛說太子慕魄經， 348. 佛說大乘方等要慧經， 356. 佛說寶積三昧文殊師利菩薩問法身經， 492. 佛說阿難問事佛吉凶經， 506. 犍陀國王經， 525. 佛說長者子懊惱三處經， 526. 佛說長者子制經， 551. 佛說摩鄧女經， 553. 佛說●女祇域因緣經， 554. 佛說柰女耆婆經， 602. 佛說大安般守意經， 603. 陰持入經， 604. 佛說禪行三十七品經， 605. 禪行法想經， 607. 道地經， 621. 佛說佛印三昧經， 622. 佛說自誓三昧經， 684. 佛說父母恩難報經， 701. 佛說溫室洗浴眾僧經， 724. 佛說罪業應報教化地獄經， 729. 佛說分別善惡所起經， 730. 佛說處處經， 731. 佛說十八泥犁經， 732. 佛說罵意經， 733. 佛說堅意經， 735. 佛說鬼問目連經， 779. 佛說八大人覺經， 791. 佛說出家緣經， 792. 佛說法受塵經， 1467. 佛說犯戒罪報輕重經， 1470. 大比丘三千威儀， 1492. 佛說舍利弗悔過經， 1557. 阿毘曇五法行經， 2027. 迦葉結經

(東漢)支婁迦讖： 204. 雜譬喻經， 224. 道行般若經， 280. 佛說兜沙經， 313. 阿●佛國經 350. 佛說遺日摩尼寶經， 361. 佛說無量清淨平等覺經， 417. 佛說般舟三昧經， 418. 般舟三昧經， 458. 文殊師利問菩薩署經， 624. 佛說伅真陀羅所問如來三昧經， 626. 佛說阿闍世王經， 807. 佛說內藏百寶經

(東漢)嚴佛調： 778. 佛說菩薩內習六波羅蜜經

(東漢)安玄共嚴佛調： 322. 法鏡經， 1508. 阿含口解十二因緣經

(東漢)支曜： 46. 佛說阿那律八念經， 114. 佛說馬有三相經， 115. 佛說馬有八態譬人經 608. 小道地經， 630. 佛說成具光明定意經

(東漢)康孟詳： 137. 舍利弗摩訶目連遊四衢經， 197. 佛說興起行經

(東漢)曇果共康孟詳： 196. 中本起經

（東漢）康孟詳共竺大力： 184. 修行本起經

（吳）支謙： 153. 菩薩本緣經， 185. 佛說太子瑞應本起經， 198. 佛說義足經， 200. 撰集百緣經， 225. 大明度經， 362. 佛說阿彌陀三耶三佛薩樓佛檀過度人道經， 632. 佛說慧印三昧經， 735. 佛說四願經， 790. 佛說孛經抄

（吳）維祇難： 210. 法句經

（吳）康僧會： 152. 六度集經

（西晉）竺法護： 154. 生經， 170. 德光太子經， 186. 佛說普耀經， 222. 光讚經， 263. 正法華經， 266. 佛說阿惟越致遮經， 285. 漸備一切智德經， 398. 大哀經

（西晉）無羅叉： 221. 放光般若經

（西晉）法矩共法立： 23. 大樓炭經， 211. 法句譬喻經

（西晉）聶承遠： 638 佛說超日明三昧經

（西晉）安法欽： 2042 阿育王傳

（東晉）僧伽提婆： 26. 中阿含經

（姚秦）竺佛念： 212. 出曜經

（姚秦）佛陀耶舍共竺佛念： 1428. 四分律

（後秦）鳩摩羅什： 201. 大莊嚴論經， 208. 眾經雜撰譬喻， 223. 摩訶般若波羅蜜經

（後秦）弗若多羅共羅什： 1435. 十誦律

（北涼）曇無讖： 157. 悲華經

（元魏）慧覺： 202. 賢愚經

（梁）寶唱： 2063. 比丘尼傳， 2121. 經律異相

（梁）僧佑： 2145. 出三藏記集

（蕭齊）求那毘地： 209. 百喻經

（隋）闍那崛多： 190. 佛本行集經

（唐）道世： 2122. 法苑珠林

附錄三、近代漢語語料庫預定目錄

在以下所列書籍中，已輸入或進行中者劃底線，但大多數只是才開始輸入或仍在校對中，所以尚未能夠上線。這些劃有底線的書籍現在大部分納入史語所的漢籍輸入計劃，並由該所漢籍全文資料庫計劃負責輸入及校對，其中也有部分是與其他機構合作輸入的。

這裡所列的書單可說只是近代漢語語料庫第一階段的需求，將來隨著新的發展必然會有新的需求。例如唐宋也有不少筆記可以補充以下所列書單未能反映出來的語言現象，但因爲往往文白夾雜，免不了要費披沙揀金之功，所以在本階段我們只選擇整體上反映當代語言較多的文獻。

（唐五代）

遊仙窟，朝野僉載，王梵志詩，神會語錄，六祖壇經，入唐求法巡禮行記，敦煌變文集，敦煌曲子詞集，唐五代禪宗語錄（大正藏多種），全唐詩，祖堂集

（宋）

乙卯入國奏請（出自《續資治通鑑長編》265 卷），河南程氏遺書，景德傳燈錄，五燈會元，禪宗語錄（大慧普覺禪師書，虛堂和尚語錄，碧巖錄），三朝北盟會編，王俊首岳侯狀（出自王明清《揮麈錄》），大唐三藏取經詩話，朱子語類，全宋詞，劉知遠諸宮調（金），董解元西廂記諸宮調（金）

（元明清）包括宋元難分者

魯齋遺書（許衡）（直說大學要略，大學直解），吳文正集，孝經直解（貫雲石），元典章，元代白話碑，清平山堂話本，大宋宣和遺事，五代史平話，全相平話五種（武王伐紂平話、七國春秋平話後集、秦并六國平話、前漢書平話續集、三國志平話），元朝秘史，皇明詔令（部分），老乞大諺解，朴通事諺解，訓世平話，正統臨戎錄，元刊雜劇三十種，關漢卿戲曲集，永樂大典戲文三種（張協狀元、小孫屠、宦門子弟錯立身），宋元四大戲文（荊釵記、白兔記、拜月亭、殺狗記），水滸全傳，三遂平妖傳，西遊記，金瓶梅詞話，醒世姻緣，型世言，三言二拍，肉蒲團，儒林外史，紅樓夢，岐路燈，鏡花緣，品花寶鑑，兒女英雄傳，老殘遊記，桃花扇