

知識的投射與解讀

兼論數位典藏中之語言座標

Projection and Interpretation of Knowledge: On Linguistic Anchoring of Digital Archives

黃 居 仁 (Chu-Ren Huang) 83-89

中央研究院語言學研究所

Institute of Linguistics, Academia Sinica

壹◎前言

人文及社會科學的研究本質可以視為知識的投射與解讀。任何素材，不管是文本、器物、書畫、或調查統計資料，對學者而言，都是由某個特定時空投射出來的知識綜合體。研究的目的，便是在跨越時空與使用者的藩籬，得到最完整有效的知識投射，導出最精闢深入的解讀。

人文知識投射及解讀的基本碼 (Code) 是什麼呢？生命科學知識的基本碼是基因。生命的知識與奧秘，都由基因承載。找出所有可能基因之後，所有生命的知識奧秘的解答，便在基因如何排序，組合的限制，及基因 (排序組合) 與其功能間的關係。

人本知識的表達系統，最基本的還是人類的語言；而語言的基本碼單位，便是字詞。病理學家王乃三博士，便曾以基因為喻，來看語言的解碼。單一的字詞，有如基因。可盡數的字詞，一如有限的基因。語言知識的建構與解析，其奧秘就在字詞的組成，及組成後的無限語境及相對功能上。

從知識內容來看，語言是人類溝通、表達情感、提供資訊、傳承文化共同且必備的工具。語言，使得人得

以傳承並累積文化知識，異於禽獸。而語言的流通與演變，造成不同的語言的差異，更是人類族群與文化認同的重要指標。在新世紀裡，由於科技縮短了時空，使跨語言的知識流通變成必然；這種多語資訊環境，不但豐富了人類的文化資產與知識內容，更突顯了語言作為知識基底架構的重要性。

本文便是知識投射的觀點出發，以語言文字為重要的投射與解讀媒介，來討論投射解讀的方式、方法及可用參考資源。重點在數位知識時代的投射與解讀。

貳◎知識的投射

嚴格說來，從創作，到複製，到閱讀，文本或語言的每次使用，都是一個單次的投射。當然，每次投射的完整性，及是否有效傳達，取決於投射對象及使用環境等多重因素。本節中由投射的形式及過程兩個角度，來分析知識的投射。

一、投射的形式

投射的形式，與知識組合後的自然單位有密切不可分的關係。我們先依投射單位區為文本，熟語，語言詞彙，及語言隱喻，再來分析其投射原則的異同。

1. 文本

文本應是人文社會者最熟悉的研究單位之一。文本作為知識投射形成的特點是其自由度，也就是說，文本由作者意志決定的程度最高，約定俗成的成分最少。因此文本的長短、呈現方式，或者體裁都可有很大的變化。短自兩句對話，20字的五言絕句，都可算是一個文本，一次演講。一齣話劇的演出，也是一個文本。文本的另一個特色，是與語境（context）間的互動。“Text and Context”一向是西方人文研究的重要題旨。一個簡單的文本，常因其使用的語境，而投射不同的知識，得到不同解讀。當然，text本身也可以是context。這是另一個文本不同於其他投射形式的特色。也就是文本中可包含其他文本，如章回小說中引用的詩詞（如《紅樓夢》中的「好了歌」）。「好了歌」本身是一個文本，但他出現在《紅樓》這個大文本中，有了不同（更豐富）的意義。

2. 熟語：成語、歇後語

除了文本之外，其他三種投射基本上都有固定的形式，只是形式限制各自不同。先是熟語（Conventionalized expression），顧名思義是因熟用之後有了固定形式的一句話。熟語是把一個小文本（故事／敘述典故）濃縮到一句話裡，因而可以用一句話的形式投射一個文本的訊息。但因傳達的訊息已約定俗成，故不會隨語境而改變。倒是熟語的形式有時可容許小規模的變化，在中文環境裡，熟語的形式最主要有（1）成語，如「有口皆碑」，如「二桃殺三士」；（2）歇後語，如「騎驢看唱本——走著瞧」。歇後語是中文特有的形式，其投射的特點之一是歇後的部分可以省略。

3. 詞彙

以詞彙投射知識，是約定俗成成分最高的。基本上，詞彙所表達知識之間的關係是固定的，改變詞彙的形式，就會改變意義。舉例來說，「慢」與「惰」是語意相關的字，但「怠慢」與「怠惰」是完全不同的詞。

4. 語言隱喻

把語言隱喻做為知識投射形式的特色是以知識領域間的相互關係為基礎；也就是說，投射的形式其實就是知識。「當代隱喻理論」（Contemporary Theory of Metaphor）主要是由雷克夫（G. Lakoff）及強生（M. Johnson）所發展出來，主要就是關切知識與認知領域的投射。比如說，本文開場白用了基因的隱喻，就是把有關基因的

領域知識投射到本文的主題來。另一個大家熟知的例子是時間隱喻。因為時間是抽象概念，因此各語言通常無法直接描述時間，故借用不同隱喻。中文的基本隱喻是「時間是流水」，便是利用一般人熟知有關流水的知識（向前流、不回頭、可測量長短等）來描述時間。因此，語言隱喻做為知識投射形式，主要特色在其目的並不傳遞知識，而利用現有知識架構，描述不熟悉（或較抽象）的知識領域。

二、投射的過程：起點與終點

知識投射，除了形式外，也因其過程而會產生本質上的差異。投射過程的變化，主要可由比對其起點與終點來分析：

1. 同語同時（Monolingual）

當創造／使用文本者與解讀者基本上使用同一個語言、活在同一個時代時，便是「同語同時」的知識投射，這是最基本的過程。當然，即使同語同時，也有各人用語、社會階層、專業領域、甚至方言等的差異。因此投射與解讀，仍需費周章。但同語同時的投射，起碼可以假設同樣的（基本）詞彙頻率與分布等，可做為其他較複雜投射／解讀形式之化簡模式。

2. 異語同時（Synchronous）

時代相同，但創造／發表者與解讀者使用的語言不同。這可能是當今全球經濟最關鍵的知識投射方式之一。請注意到，經過翻譯與否並不是知識投射的重點，只要投射知識用以表達的語言與解讀者使用的語言不同，就會產生表達形式與解讀形式的差異，就會產生多語知識投射的問題。

3. 異時（Diachronous）

讀者也許已開始納悶為何本節不分同語與異語。其實道理很簡單。時代拉開以後，嚴格說來，所用的語言已是不同的語言（用「語言」最簡單的定義原則，彼此若不能聽懂彼此談的話，說的便是兩種語言）。漢字書寫系統，使我們有「看得懂古文」的部分錯覺，但若起古人於地下，他們說的話，我們絕對是聽不懂的。而我們所看的古文，其實和不會說日文的人看日文漢字，或現代學者看甲骨文，看楔形文字一樣，只不過利用其他知識盡力猜測而已。準確程度或有不同，但與真正使用該語言為母語者，總是有一段距離。因此，只要拉開時間，不但背景知識不同，使用語言內含的知識，也一定

會有變異。若幸而使用相同的書寫系統（如當代中文與古漢語），只是表達較方便，其實知識表達方式的差異，是一樣複雜的。

參◎ 知識的解讀，受到投射形式、起點、終點的影響

知識投射的目標是要傳遞訊息，得到正確的解讀。上節已分析投射的形式及過程，本節將要先分析投射知識的內容與本質，再由上述四個變數看如何解讀。

一、知識內容

這裡所指的知識內容，是指的投射的知識綜合體中，包含了那些類的知識。上文中提到了表達人本知識的基本媒介是語言，我們也因此由語文知識出發來看其內容。以語文表達的知識綜合體，很自然的承載了兩大類的知識內容。一類是語文本身承載的訊息，我們稱之為語文知識本體；另一類是描述所使用語言文字的基本訊息，我們稱之為語文知識的座標（引用謝清俊教授的說法）。以下分述之。

1. 語文知識的本體

這裡指的本體，是語文表達知識時所必要同時表現出來訊息。這些知識的本體，雖然不見得是傳承/解讀的（主要）對象，卻是用以表達訊息的必要手段，或是表達某些訊息所必須留下的蛛絲馬跡。因此是知識投射與解讀過程中，不可或缺的工具，其內容可羅列如下：（1）字形（2）讀音（3）詞類（4）語意（5）頻率（6）解釋/用例（7）語境，篇幅所限，不一一贅述。

2. 語文知識的座標

語文知識的座標，把所投射的知識內容及其表達方式，放在一個正確的位置上，使其能得到正確的解讀；可說是由語文這個表達形式產生的 Metadata（後設/詮釋資料）。內容羅列如下：

（1）時代（2）地域（3）階層（4）領域（5）作者/使用者（6）出處（7）文獻，限於篇幅，不另細述。

二、知識的本質（Ontology）

討論過規範投射知識的外顯條件，包括形式、過程、及內容的外顯分類，我們可以更進一步探討知識的

本質。我們認為，四種知識投射形式的本質可概述如下：

1. 文本是主觀組織的複雜知識系統

我們在討論知識投射的形式時，談到創作者在建構文本時，有很大的自由度。雖然不同文體各有其限制，但創作者有選擇文體的自由（包括混合建構不同文本，組成大文本；如高行健的《靈山》在小說文本的大架構下，融合了散文、民歌、戲劇等文體的小文本）；而引導整個架構的，其實就是創作者想主觀投射出來（即先經建構，再予表達）的知識體系。當然，知識體系，因主觀需求而有不同的層面，不同的風貌。比如說，學生週記，大概多是一個流水帳；詩/詞通常是經營一個意象或情境；而愛因斯坦的《相對論》，則是對所有物理現象的基本解釋；達爾文的《物種起源》，則是對生物演化的解釋。愈豐富的知識體系，在解讀時有愈多的可能性，及更多新知識的衍生。

2. 熟語是一個情境

相對於文本的複雜結構，使其可以表達種種不同的知識系統；熟語的知識是一種情境（Situation），或是一個場景（Scene）。正因熟語投射的知識本質是一個固定典故造成的情境，其目的是要引用投射起點（說話者）及終點（解讀者）所共同熟知的知識，因此，並無真正解讀上的困難。知識投射的內容最穩定、有效。

3. 詞彙投射的知識是概念的群組合（cluster of concepts）

詞彙的使用也是約定俗成的。其與熟語的對比在於表達的知識最基本、最接近思想的基本單位——概念。因為所有知識系統都必須由不同概念建構，因此詞彙可說是知識投射的基本單位。

其實，詞彙又是人類使用語言的基本單位，他與真正的概念又有所區分。簡單來說，語言並不把概念做邏輯式的切分，而是因認知或人類經驗的需求，把有關係的概念組合在一起。而這個概念群組投射的時候是不可區分的。因此我們主張詞彙投射的知識是概念群組。這個群組可能只含單一概念，也可能包括一組認知上相關的概念。

以下試舉兩個常見的例子：「書」是一個詞義明確的詞，可是在概念上可以指其物理特性（外觀、重量等等），可以指其組成材質（這本塑膠書有 22 頁），或其內容（書有趣），甚至其創作過程（這本書不好寫）等等，這些是有明確關聯的一組概念，但也可以一起使用

(這本書好極了)。類似的，「門」既可以指其建築物的入口（進門），也可指關閉該入口所用的設施（門闕著）；重要的是投射的知識是整組概念（門很寬），只是某些語境需求才會有所篩選。

4. 語言隱喻投射的知識是概念系統

上一節已稍提到使用語言隱喻時是把概念系統由一個熟知的領域，投射到另一個領域。與詞彙不同的是，隱喻投射的是概念的系統，不是概念本身；與熟語不同的是，投射的過程有規律限制。因此，解讀必須受系統知識與概念的規範（這是安可思教授之國科會研究計畫研究中的課題）。以「人生是一場戲」為例，有主角，有開幕，有劇本，有悲喜等概念，都可以系統性的轉移，用以描述人生。但如要賣票，要憑票入場，演出時打字幕等概念則是用不著的。

三、由知識本質到知識特質：組織、複雜度，與變異性

1. 概念主導VS.語言主導

上一個小節未對隱喻的討論突顯了語言隱喻作為知識投射形式的一個特色：那就是說其知識表達的組織是由概念主導的。相對於隱喻，文本（篇章結構），熟語（句型），及詞彙（構詞、構字）的知識表達組織是取決於不同語言的結構規律。這個組織上的對比將影響解讀的策略。

2. 知識的複雜度及系統的完整性

文本 > 熟語 > 詞彙

以知識的複雜度而言，雖然隱喻以概念主導組織，與其他三個形式不同，但隱喻具有完整的系統。與文本一樣，系統最完整，知識最複雜。相對的，詞彙是最小的知識投射單位，為概念群組，群組中並隱含認知的結構，單元性高，複雜度低。最後，熟語的知識單位內容大於詞彙，其本質為「引用」情境，故其知識內容是一個具體而微，且無變動性的系統。故知識複雜度及系統完整性均居中。

3. 投射知識的變異性

文本 < 隱喻 < 熟語 < 詞彙

知識表達方式及知識間的關係是約定俗成的，因此可能因成俗（Convention）的改變（或語言使用環境的改變）而變異。一般而言，知識系統愈複雜，其投射的知識內容愈不易變化（當然，複雜系統的解讀具多樣

性。這是解讀的問題，而不是投射知識系統本身的變化）。因此，投射知識變化的可能，恰與知識系統複雜度成反比。詞彙與概念的連結最單純，任意性最高，受創作者主觀控制的程度最低，當然知識內容最容易改變。

肆◎知識解讀、解讀工具

由知識的投射到解讀，若要用一個簡單的比喻，便是作者和讀者間的關係。當然，閱讀的方式很多，包括許多跨媒體的閱讀，但是獲取知識的閱讀目的不變，則其閱讀的策略不能與知識投射的方式脫節。而人文社會學者對知識的解讀，當然希望深度與廣度都超過一般讀者。然而在解讀時的策略，則學者的解讀與一般讀者的策略應該相同；只是學者能否善用現有資源，能否善用（參考）資訊，在資訊時代更形重要。因為隨著文本數位化的普及，機讀文獻已是人文參考資料最完整的來源。但是我們仍需強調，無論使用了多少參考資料庫與工具，解讀的最後仲裁者仍然是人，只有讀者才能決定相關訊息資料之選用，才能擇定特別的解讀。

一、解讀的依據之一：詞彙庫

本文對知識投射的方式、內容、及本質的分析，強調了知識表達方式因著時空等多項指標而變化。因此解讀的重點在於找到正確的投射起點，然後依據投射起點的語文知識，給予投射內容正確的解讀。

因此，建立完整解讀環境的前提，便是在時間、地域、文體、主題、作者、背景等多重因素下，建立每個可能的語文知識環境的解讀參考資料。這是所謂的「語言座標」。以漢學研究而言，最顯而易見的例子也許就是斷代詞彙庫。我們稱之為「時間軸上的語言座標」。

每個知識綜合體都因其創造時代不同而使用不同的語彙。若不能了解當代的語彙，當然無法正確解讀。進一步說來，因為語彙代表了概念的群組，每個時代的詞彙庫，不但收集了所有當代的觀念，而其組合方式更代表了概念間的關係；其使用頻率則透露了概念間輕重緩急的訊息。因此每個斷代的詞彙庫，不但是該時代知識概念的基本集合，也是解讀該時代知識的重要訊息。中研院語言所文獻語料庫及合作單位已開發之斷代詞彙

庫，可為漢學研究之參考；如附表一。

二、解讀的依據之二：語料庫

「語料庫」的作法，是把某個特定時空背景的語料，作大量的取樣。取樣需大到具代表性，也就是說，對特定時空的語文知識，我們可以不用另定規律描述，只要對這個語料庫做研究、觀察即可。當然，觀察的方式，可以很直觀，也可以用複雜的統計式及工具。語料庫之組成及其用途，取決於其設計原則；而其功能上的多樣性，則來自標準化的標誌。所有語言表達方式均是語料，因此，出口下筆皆語料。但是，如何把大量生語料組成語料庫呢？「語料庫」(Corpus)的定義是，根據一套設計原則而蒐集完備的一組標準化電子文獻。因此他必須有(1)設計原則(Design Criteria)：語文知識座標的組合；及(2)標準化：標記(mark-up)與(詞類)標記(tag)。重要的中文語料庫請參見附錄。

三、解讀工具與資源的分類

除了上節的詞彙庫及語料庫外，其他解讀工具羅列如下，可參見附錄的實例。

I. 知識庫

1. 詞彙庫 / 辭典
2. 語料(庫)
3. 句結構資料庫 (TreeBank)
4. 語意資料庫 (如詞彙網絡WordNet)

II. 知識排比與抽取的工具

伍◎結語

語言典藏與語言座標是知識投射及解讀的基本架構。由知識的投射及解讀出發，我們發現語文知識的完整典藏及語文座標的建立，是典藏知識必備的基本架構。在目前的環境下，推動語言典藏及語言座標更有以下重要意義：

- 人文與本土的關懷：弱勢語言的典藏傳承是其文化在資訊時代薪火相傳的前提。
- 國際化與國際競爭力：跨語言的典藏解讀與知識處理能力是知識經濟時代國際競爭力的前提。
- 科技與人文的契合：語言文字所蘊涵的知識架構是

人文素養與數位化數據間唯一隨時暢通的陽關大道。

我們希望這篇不成熟的短文，能引起將來對語文知識座標及解讀方法更多深入的討論。

致謝：特別感謝多年來合作伙伴的啟發，尤其是謝清俊、陳克健、魏培泉、譚樸森、劉承慧、及羅鳳珠。感謝助理盧秋蓉的費心繕打。感謝《漢學研究通訊》鍥而不舍的邀稿，及多次對延遲交稿的容忍；要不然，這篇小文，儘管仍不完備，也沒「暫時完稿」的一天。當然，疏失錯漏之處，仍是我個人的文責。

附錄◎解讀參考工具及資料網站

這份資料雖在最近修訂，但大部分是在去年(2000年)搜集的。網路更新變異極快，故過時漏收之處在所難免。特別感謝臺大高照明、中華佛學研究所維習安，及暨南大學張景新提供部分資料。

Online Resources for Chinese Corpus Linguistics

Chinese Corpora 中文語料庫

「中央研究院現代漢語平衡語料庫 (Sinica Corpus)」：

<http://www.sinica.edu.tw/~tibe/2-words/modern-words/> OR

<http://www.sinica.edu.tw/ftms-bin/kiwi.sh>

「上古漢語語料庫」：

<http://www.sinica.edu.tw/ftms-bin/ftmsw3> OR

<http://www.sinica.edu.tw/~tibe/2-words/old-words/>

「香港城市大學語言資訊科學研究中心共時語料庫 (LIVAC)」：

<http://www.rcl.cityu.edu.hk/livac/>

Chinese TreeBank 中文句結構資料庫

The Chinese Penn TreeBank

<http://www ldc.upenn.edu/ctb/>

Sinica TreeBank

<http://godel.iis.sinica.edu.tw/CKIP/trees1000.htm>

Searchable On-line References 可檢索之數位參考室

搜文解字 (Chinese Linguistic KnowledgeNet)



<http://www.dmpo.sinica.edu.tw/~words>

搜文解字II: 生活語文與唐宋流行歌 (即「文國尋寶記」)

<http://www.sinica.edu.tw/wen>

KanjiBase on the Web (Chinese characters)

<http://www.gwdg.de/~cwitter/home.htm>

General Links 有用的連結

中研院資訊所中文詞知識庫小組

<http://godel.iis.sinica.edu.tw/CKIP/>

中華民國計算語言學會

http://rocling.iis.sinica.edu.tw/ROCLING/index_c.htm

教育部國語會

<http://www.edu.tw/mandr/index.htm>

Prof. Marjorie Chan (Ohio State U.): Chinese-Links

<http://deall.ohio-state.edu/chan.9/c-links.htm>

Chinese and Oriental Languages Information Processing Society (COLIPS)

Singapore site with on-line journal and other links

<http://www.iscs.nus.edu.sg/~colips/new.html>

Chinese Language Information Page

Commercial site with rich but incomplete information and not updated regularly

<http://www.webcom/bamboo/chinese/chinese.ht.ml>

Harvard's Chinese Linguistics Links

<http://www.fas.harvard.edu/~whu/China/linguist.html>

中華佛學研究所

<http://www.chibs.edu.tw/index.htm>

Tools 工具

<http://www.mandarintools.com>

Downloadable segmentation and tagging program

<http://godel.iis.sinica.edu.tw/CKIP/ws>

Concordances 建索引典的工具

By Christian Wittern (for DOS/MAC)

<http://www.ijnet.or.jp/iriz/iria.html/tools/concord.htm>

By Pregadio (For MAC)

<http://helios.unive.it/~pregadio/computing/conc/conc.html>

Word Lists, etc. 詞表

<http://www.wenlin.com>

Character Frequency and Stroke Number

<http://casper.beckman.uiuc.edu/~c-tsai4/chinese/charfreq.html>

Word List Survey

<http://www.geocities.com/hao510/wordlist/index.htm>

Misc. 其他語料庫

Bell Labs Text-To-Speech Demo

<http://www.bell-labs.com/project/tts/Madarin.html>

中英, 中日, 中西雙語語料庫

光華雜誌 <http://www.sinorama.com.tw>

Multilingual Resources 多語知識庫

EuroWordNet

<http://www.hum.uva.nl/~ewn/>

English 英文語料庫的資源與工具

英國國家語料庫 (British National Corpus)

<http://thetis.bl.uk/>

ICE (International Corpus of English) [Sample]

<http://www.ucl.ac.uk/english-usage/>

線上機讀英文詞彙知識庫 (lexical knowledge database)

普林斯頓大學 WordNet

<http://www.cogsci.princeton.edu/~wn/>

Special Interest Group on Lexical Resources (SIGLEX) /ACL

<http://www.cres.com/siglex.html>

Linguistics Data Consortium (LDC)

<http://www ldc.upenn.edu/>

Encyclopedia Britannica (including Merriam-Webster Collegiate Dictionary)

<http://www.britannica.com>

Grammar 文法

<http://www.ucl.ac.uk/internet-grammar/>

Annotated list of resources 帶注釋的資源清單

Christopher Manning on statistical natural language processing and corpus-based

computational linguistics

<http://www.sultry.arts.usyd.edu.au/links/statnlp.html>

- Michael Barlow on Corpus Linguistics
<http://www.ruf.rice.edu/~barlow/corpus.html#Text>
 Linked list
<http://csmart.iis.sinica.edu.tw/~mjlee/links/NLP/>
 Association of Computational Linguistics
<http://www.cs.columbia.edu/~acl>
- Papers and Lecture Notes 網路上有關語料庫語言學的論文或講義
 Computation and Language Archive
<http://xxx.lanl.gov/>
 Collection of Corpus Linguistics Papers
<http://www.ilc.pi.cnr.it/EAGLES96/browse.html>
 Corpus Linguistics by Tony McEnery and Andrew Wilson
<http://www.ling.lancs.ac.uk/monkey/ihe/linguistics/contents.htm>
 Lectures on Tools for Corpus Linguistics by Daniel Hardt
<http://www.cst.ku.dk/dan/corpus.html>
 Data-Intensive Linguistics by Chris Brew and Marc Moens
<http://www.ltg.ed.ac.uk/~chrisbr/dilbook/>
 Catherine N. Ball 介紹如何利用關鍵字詞前後文查閱程式與語料庫
<http://www.georgetown.edu/cball/corpora/tutorial.html>
- 喬治城大學計算 Chi-Square 值講義
http://www.georgetown.edu/cball/webtools/web_chi_tut.html
- Tools 語料庫語言學的工具
 賓州大學
<http://www.cis.upenn.edu/~adwait/penntools.html>
 Mike Scott 的 Wordsmith (關鍵字詞前後文查閱程式 (Concordancer) 及詞彙分析)
<http://www.liv.ac.uk/~ms2928/>
 雪梨大學 Toke 英文分詞 (tokenization) 與構詞分析程式 (morphological analyzer) (UNIX 及 Window 版)
<http://www.sultry.arts.usyd.edu.au/SULTRY/toke/index.html>
 Eric Brill 詞類標記程式
<http://www.cs.jhu.edu/~brill/code.html>
 語法剖析器
 Davy Temperley, Daniel Sleator, John Lafferty 的聯繫語法 (Link Grammar) 剖析器
<http://www.link.cs.cmu.edu/link/>
 Steven Abney 的 Cass 部分語法剖析器 (partial parser)
http://www.sfs.nphil.uni-tuebingen.de/~abney/Steven_Abney.html
 Dan Melamed 統計工具程式
<http://www.cis.upenn.edu/~melamed/home.html>

附表一

- | | |
|--|--------------------------|
| 1. | 2. |
| 中央研究院中文詞知識庫小組技術報告 | —— 以下各專書詞表可由「搜文解字」I、II網站 |
| —— 以下各字詞頻表均有電子檔 | 直接查詢 |
| 技術報告 94-02 古漢語字頻表 | 唐詩三百首詞表 |
| 技術報告 95-01 注音檢索現代漢語字頻表 | 四書詞表 |
| 技術報告 97-01 古漢語詞頻表 (甲) | 老子詞表 |
| 技術報告 97-02 論語詞頻表 | 莊子詞表 |
| 技術報告 98-01 詞頻詞典 | 紅樓夢詞表 |
| 技術報告 98-02 Accumulated Word Frequency in | 水滸傳詞表 |
| CKIP Corpus | 宋詞三百首詞表 |
| | 國小課本詞表 |