漢字知識表達的幾個層面：字，詞，與詞義關係概論

中央研究院語言學研究所 黃居仁

Knowledge Representation with Hanzi: The relationship among characters, words, and senses

Chu-Ren Huang, Academia Sinica

摘要

語言是人類表達知識，與彼此溝通的最主要工具。在口口相傳的媒介下，語言使知識得以跨越時空、交換與傳承。經文字的書寫紀錄，人類的知識因約定俗成而降低失真率，更可以超越時空距離與個人記憶的限制而累積。

人類累積知識系統的能力，經由文字的表徵而跨越了面對面接觸的限制後，又經過了兩次的媒材革命。印刷術的發明，主要是數量級的改變，使得知識的傳播不再限於少數菁英份子，而回復到所有人們所共有共享。數位媒材的興起，則完全改變了知識累積的方式與速度。理論上，所有已紀錄的知識都可以永久保存，任何知識都可以互相結合，知識整理的方式也不再限於其原有的形式或系統。

在這個知識流通的時代，其實我們更應該思考知識表達的基本機制。不論知識如何傳播或重整，知識的源點與終點都是人。而人類知識的單元，呈現出來的就是語言的單元。語言中字詞與詞義的關係，事實上就表現了說話者所體認的知識單元。文字與意義間的系統關係，就是人類知識建構的最基本關係。這個知識的基本架構，在不同的語言中，有其必然的共通性，卻又有極富意義的變異。本文便嘗試由漢字與漢語這個系統的特性，來探討一些語言與文字表達知識的基本問題，重點將放在中文內部表達的知識與其系統。

大綱

1. 前言：資訊因語言所賦予的結構才成為知識
2. 基本單位：字與詞
3. 如何定義中文的詞（Word）：詞形，詞音，詞類，與詞義的互動
4. 中文知識系統的表達
5. 結語

1. 前言：資訊因語言所賦予的結構才成為知識

語言是人類表達知識，與彼此溝通的最主要工具。在口口相傳的媒介下，語言使知識得以跨越時空交換與傳承。經文字的書寫紀錄，人類的知識因約定俗成而降低失真率，更可以超越時空距離與個人記憶的限制而累積。人類累積知識系統的能力，經由文字的表徵而跨越了面對面接觸的限制後，又經過了兩次的媒材革命。印刷術的發明，主要是數量級的改變，使得知識的傳播不再限於少數菁英份子，而回復到所有人們所共有共享。數位媒材的興起，則完全改變了知識累積的方式與速度。理論上，所有已紀錄的知識都可以永久保存，任何知識都可以互相結合，知識整理的方式也不再限於其原有的形式或系統。

在這個知識流通的時代，其實我們更應該思考知識表達的基本機制。不論知識如何傳播或重整，知識的源點與終點都是人。而人類表達與處理知識的單元，就是語言的單元。語言中字詞與詞義的關係，事實上就表現了說話者所體認的知識單元。文字與意義間的系統關係，就是人類知識建構的最基本關係。這個知識的基本架構，在不同的語言中，有其必然的共通性，卻又有極富意義的變異。

另一方面，我們必須體認資訊因語言所賦予的結構才成為知識；這是從網路時代與數位媒材的觀點得到的結論。在網路上的資訊量成對數級數暴漲，從 giga 到 tera。大量的資訊，隨時上網，唾手可得。但是，這些資訊，到底多少是可用的，是使用者可以隨時抽取，在思考，談話，或報告中引用的呢？如同把一架噴射客機放在沒有跑道的荒山野地中，他毫無作用一樣；資訊如果不放在正確的知識體系裡，也發生不了作用。而我們在上文中剛提到的；人們的知識體系，是建立在他們的語言上的。如何讓流通的資訊變成有價值的知識，必須由人類知識體系的最基本單元—字詞—著手。

在世界幾千種語言中[1]，漢語是相當特殊的一個。不單單是因為漢字書寫系統的獨特性。他的表意基礎，更使得不同語族的語言，可以用同一個文字系統來書寫。本文便嘗試由漢字與漢語這個系統的特性，來探討一些語言與文字表達知識的基本問題，著重於所呈現的知識系統性。

2. 基本單位：字與詞

談論中文書寫系統的基本單位時，不能不由「字」與「詞」這兩個單位的區分開始。在語言的日常使用中，「字」與「詞」兩個概念是不分的。從我們常用的「英文單字」或「英文新詞」可互相代換這個現象看出。但是，中文裡，「字」還是比較基本的用語。口語中談到語言的最小單位，幾乎都是用「字」；如「一個字也不認識」，「生字」等。但「字」的概念，來自漢字的書寫系統，與後來

[1] 根據 SIL 出版的民族語言(Ethnologue)這個最權威的資料庫，目前世界上有六千多種語言。這還不包括歷史上存在，但已滅亡的語言。也不考慮因語言演變的變化太大，當代語言與古代語言無法相通的許多例子。

翻譯西方「詞」(word)的概念，並不相同。這是為什麼趙元任提出「社會詞」(sociological word)這個名詞，來解釋中文中「字」的觀念，以有別於語言學家定義的「詞」。他的意思，就是在使用中文的社會中，一般是以「字」作為語言文字的基本單位。這與語言結構中的基本單位—詞—並不相同，但作為約定俗成單位的功能是類似的。趙元任這個說法，對字詞的社會語言學，提出了很好的解釋。但對於我們今天區分知識表達基本單元的議題，則嫌不夠明確。因此，我們底下還是採語言學理論中的定義，把「字」與「詞」作明確的區分。

「字」或「漢字」(Chinese characters）是中文書寫的基本單位。其概念區分很清楚，就是所謂的「方塊字」。在一個書寫的方塊裡，只會有一個字。比如說，「三個字」這個括弧中的字串，含有三個字，任何會中文的人都不會算錯。

「詞」(word）是一個嚴謹定義的語言學名詞，他的定義就是「語言中表達意義的最小獨立單位」。這個定義適用於所有的語言，因此詞的抽象概念，對語言的對比研究非常有用。但是，定義中的「最小單位」「獨立單位」等定義，往往在不同語言中有不同的測定方法。在中文裡，字與詞的最大差別，在每個詞可能帶有一個或一個以上的字。「葡萄」也許是個最明顯的例子。意義上，「葡萄」這兩個字是一個單位，分開後「葡」字與「萄」字都沒有可區辨的意義。因此，「葡萄」是一個詞，一個雙字詞。這也是語言學上常常用來判定詞的辦法。將一個字串往下切分，如果切分後的單位沒有意義，那麼切分前的最小單位就是詞。

儘管在概念上，可以區辨獨立的詞。中文「字」與「詞」分辨的最大困擾在於書寫的傳統。傳統書寫的自然單位是方塊字，而詞與詞之間並沒有約定俗成的標記，像西方文字的空格。因此像「雷達」這樣的詞，雖然很明確的是一個不可分割的意義單位，但是「雷」與「達」本身各有分別的意義。[2] 因此在文本中，不靠上下文，不能馬上斷定某個漢字該獨立成詞，或應該與其他字合成一個詞。這就是中文自然語言處理中，吸引了相當多研究的「斷詞」(segmentation)問題。

我們在討論表達知識的基本單元時，可以暫時不去擔心文本中字與詞有時不易切分的問題。我們需要更關心的，是如何決定意義表達的最小單位。在知識體系（或概念系統）上，也就是要找出最小的概念單位來。這就是「詞」這個定義單位非常重要的原因。因為每個詞有形與義兩個成分，詞義的成分，就代表了這個語言中表達的最小概念單位。在心理與腦神經學研究中，是假設概念儲存於心理詞彙(mental lexicon)中，而概念提取的動作，是藉著「詞彙提取」(lexical access）的動作完成的。換句話說，要找出知識的基本單元，不論是心理或資訊處理，都要由「詞」著手。

3. 如何定義中文的詞（Word）：詞形，詞音，詞類，與詞義的互動

3.1. 詞與中文裡詞的相關定義一例

[2] 比如在「天空打雷」句中，「雷」就是一個詞。

語言在知識流通上扮演的關鍵角色，已引起了國際標準組織(ISO)的重視，並開始研擬與語言資源管理有關的國際標準(ISO TC37 SC4)。但不容諱言的，這些標準的初稿當然還是以西方語言的特色與經驗出發。一個相當清楚的例子，是這份初稿在定義同形詞(homograph) 時，引用了 ISO 在其他標準中已採用的定義[3]：「同形詞是一個與另一個詞拼法相同的詞。」這個定義，在西方語言中沒有疑義。比如說，最有名的一組英文同形詞，是 *bank* 這個拼法，可以表示「銀行」這個詞義，也可以表示「河岸」這個詞義。[4] 這兩個意義不同的 *bank*，拼法相同，因此是同形詞。但是，中文並非拼音文字，如何採用同一定義呢？若採用這個定義，如何定義中文裡的「拼法相同」呢？

我們如果假設拼法指的是任何一個拼音方式（包括注音符號，或任一個拼音方案），那麼我們會得到的所謂「同形詞」，其實是「同音詞」，就是與另一個詞語音相同的詞。如：

（1）形式，行事，刑事，形勢，型式，刑室。

「同音詞」(homophone), 如英文的 *night* 與 *knight*，已經有明確的定義了。因此，例（1）中的六個詞，應該是同音詞。在中文裡，如果說兩個詞的詞形（或字形）相同，大家一定會想到類似以下的例子：

（2.1） 長(chang2)頭髮
（2.2） 長(zhang3)見識

(3.1） 穿制服
（3.2） 制服敵人

換句話說，「同形詞」(homograph)的定義，應該取決於該語言書寫的方式(orthography)。沒有書寫，就沒有同形的問題。阿拉伯文字的書寫系統，與漢字或拉丁字母大不相同，我們也應該預期這些語言中「同形詞」的表現大不相同。因此，黃居仁在 2004 的 ISO 的相關研擬討論會上提出要把同形詞的定義修正為以下，也獲得通過。

(4) A homograph is a word that is written like another.

這個同形詞定義問題，清楚顯示了中文詞彙與意義間，有許多與西方語言學傳統理論假設不同的關係。我們最近的研究，致力於釐清中文詞的定義以及詞形與詞義間的關係，更進一步廓清中文意義表達的基本單元。

---

[3] 原文是 'A homograph is a word that is spelled like another.'
[4] 還有「飛機或船舶傾向一邊」，「以某人當金主」等其他 5 個以上詞義。

3.2. 詞形，詞音，詞類，與詞義的互動

詞為語言表達意義與知識的基本單元。至於詞這個單元的判定，在語言學分析方法上，因為意義抽象較難界定，多半由詞形，詞音，與詞類來協助。但是，中文裡這三個要素出現了複雜的互動。以下這個例子，正是我去年在 ISO TC37 SC4 研擬標準時提出，作為詞的定義標準檢測的一個例子。

（4） **背**

- *bei4* N. **背**向著太陽，就不會眼花了。——*方向* *1*
- *bei4* N. **背**上很癢。—*身體部位* *1*
- *bei1* V. **背**東西上山。—*動作* *1*
- *bei1* V. **背**了一屁股債。—*擔負責任或義務* *2*
- *bei4* V. 小孩很會**背**書。—*記憶* *3*
- *bei4* Prep. **背**著父母談戀愛。—*瞞著他人做事* *4*
- *bei4* ADJ. 最近很**背**，連喝水都嗆到。—*運氣不好* *5*
- *bei4* ADJ. 老了耳朵**背**，聽不清楚。—*重聽* *6*

以上「背」一個詞形，可能的八種意義中，以詞音區分兩組（四聲六個，一聲兩個），詞類區分四組。按傳統的語言學或辭典學的分析，可能會受以上兩個外在形式的影響，比如假設兩個第四聲屬於同一個同形詞，或名詞與動詞分別屬於兩個不同的同形詞等。但是，由深入的詞義與知識內容分，詞彙語意學家，大概會得到右邊編號的六類，六個同形詞。換句話說，詞形，詞音，與詞類，雖是好證據，但是不能取代詞義的基本判定。

其實，回歸到中文裡「異體字」與「異體詞」的對比，以相同詞形（符號的外觀）來代表不同意義的現象，又出現同中有異，異中有同的有趣現象。

（5）同中有異：異體字在不同的詞中，有時可以替代，有時不能替代

同：： **姐，姊** 如 〈姊姊，姐姐〉〈姊妹，姐妹〉

異： **小姐：***稱謂*

小姊：*姊姊中年紀最小的*

（6）異中有同：不是異體字，不能通用，但造成可通用的異體詞

異： **蘇，穌** 並非異體字

同： ＜耶蘇，耶穌＞ 為異體詞

（7）異中有同，同中又有異

（7.1.）異中有同：**升**有*動詞，名詞*兩個詞類，

只有當動詞（＝往上運動）時與**昇**是異體字

（7.2.）同中有異：但是，當「往上運動」意義，引申為使動的「使往上升」，或經隱喻引申為「調高階級」的意義只能用**升**表達，沒有異體詞

- 因為老百姓已經全逃走了，根本沒人**升**火做飯。
- 她也希望兒子能**升**大學。

以上幾個例子中，同中有異（5），與異中有同（6）兩組的關係清楚，不另贅言。例（7）中，升與昇的對比，包括了同中有異，與異中有同兩個關係，我們用研究院語料庫[5]的資料，加以分析說明。在五百萬詞的平衡語料庫中，**升**當單字詞使用共 108 筆資料，而**昇**當單字詞使用共 9 筆資料。**升**的用例中，當名詞（公升）的，有 8 筆，這是與昇相異，不能取代的。但在 100 筆動詞用法中，當及物動詞的，有 84 筆。也就是說，只有剩下的 16 筆**升**與 9 筆**昇**可能是可代換的異體字。我們檢驗了這 25 個例句，也發現他們的確可以替代。[6]

中文裡用「異體字」與「異體詞」來表達與英文「同義詞」(homonym)類似的概念，這個用法本身就值得探討。「異體」所代表的概念，是由文字書寫系統出發的。表達的是用不同的變異字體形式，來表達相同的概念。而西方語言學理論中的同義詞，則強調不同詞形代表的是同一個義項。但是，不管側重為何，我們可以由以上的三組例子看出來，漢字是一個表義很強的書寫系統。即使是遇到異體字，這種系統上容許通用的字型，都會因為表達意義的不同，而產生對比。這些例子，給了我們更強的動機，來探究中文的知識表達系統。

4. 中文知識系統的表達

從＜說文解字＞到任何權威中文字典，中文部首帶有語義分類這個事實，應該是無庸置疑的。但是，這個書寫系統，代表的是什麼樣的知識體系，卻一直缺乏深入的研究分析。我們在上文中提到，任一個語言都代表了一個知識體系。這個說法的最基本論證，就是使用同一個語言的千千萬萬人，都可以互相溝通，交換知識；表示這個語言代表了一個基底的知識體系，是所有使用該語言的人都接受的，才有辦法把交談時得到的訊息，放在知識體系中正確的位置。我們又知道，語言的知識體系，是有相當高的涵蓋率的。基本上，我們可以說，任一語言，都

---

[5] 中央研究院現代漢語平衡語料庫，簡稱研究院語料庫(Sinica Corpus)， 為分詞完成，帶詞類標記之五百萬詞語料庫。將在今年內擴充為一千萬詞。網址為：http://www.sinica.edu.tw/SinicaCorpus/。

[6] **升/昇**這個例子裡的現象大致是這樣的；在往上運動的原始意時這兩個字可互換，為異體字。但詞義變化延伸時，只發生在較常用的「**升**」字上。這在語意演變與語言演化理論裡，是可以預測的。突變產生的基本要件，就是要事件發生的次數夠多。**升**的頻率是**昇**的頻率的 12 倍，當然較可能發生詞義改變。

可以表達所有的知識。[7]因此，研究語言背後的知識體系，對人類知識的組織，與表達知識的方法，應該可以有創見性的突破。這個研究方向以往的瓶頸，在於知識體系本來就是研究最基本的先驗架構，因此很難有更高階的理論基礎可供比較研究。但近來知識本體(ontology)的研究展開，提供了比較研究的基礎。我們底下的研究，是在上層共用知識本體 SUMO (Suggested Upper Merged Ontology, http://www.ontologyportal.org) 與詞網 WordNet 兩個基底架構上進行的。

4.1. 知識表達的兩個基底架構：知識本體與詞網

4.1.1. SUMO 上層知識本體

知識本體(ontology)，在資訊科學與網路科技上的定義，就是用來描述一個系統內部知識體系的架構。這個架構通常由一組基本詞彙、定義、及該組詞彙所建立的關係共同組成。以往知識本體研究最大的瓶頸就是每個系統的知識架構都不同，因此即使建立了知識本體，也會因為彼此不相容而無法交換知識。由IEEE標準上層知識本體工作小組所建置的SUMO（Suggested Upper Merged Ontology， 建議上層共用知識本體, http://www.ontologypoprtal.org），其建立的目標就是要提供一個知識本體間建構與知識組織的共同標準。這工作小組的目的是發展標準的上層知識本體，以促進資料互通性、提高資訊搜尋和檢索的精確度、並容許自然語言介面與自動推理。知識本體（ontology）內的訊息遠遠超過字典或者術語表，能使電腦處理更多內容細節和其結構。上層的知識本體被限制在後設（meta）的概念、即一般、抽象或者哲學的概念，因此足夠涵蓋一般的廣闊範圍的領域區域。特殊領域具體的概念不被包括在上層知識本體中，但是這樣的知識本體確可提供特殊領域(例如： 醫藥、財政… 等等）的知識本體結構的建立與交換。SUMO 的設計是希望藉由共用最高層次的知識本體，來保證百家爭鳴的不同知識系統內容可以分享知識。SUMO只有3912個概念節點。採用SUMO為其上層共用本體的知識系統，理論上不會對這些上層做任何修正。而個別知識本體，內部表現的細緻知識，雖然很可能會與其他知識本體有所出入。但是這些知識，往上到較高的概念層次，就是SUMO表達的層次，一定有一致的分類。因此，即使是不同的知識體系，也可以藉由共用的上層本體作知識交換與結合。

除了維持上層本體的不變與鼓勵其他特殊領域知識本體以SUMO為基礎衍生出個別領域的知識本體，SUMO目前另一個關鍵設計是與英語詞彙網路的連結。以人類跨領域語言使用觀點，真正能超越領域障礙表達所有知識的本體架構，必須能夠連接常用的詞彙與語言概念到本體架構上。目前SUMO和英語詞彙網路 WordNet1.6/2.0 版本的連結，使得任何不分領域的知識，都可以藉詞彙的連結，建立正確的知識本體位置。這對將來語意網上，支援所有網頁都必須有的網頁知識本體，也有積極提供基礎資料的意義。

---

[7]當然有效率的問題。如用一個詞就能表達，一句話表達，還是必須用一大段文字表達。但理論上不能說某個概念，是某個語言不能表達的。

4.1.2. 詞彙網路 WordNet

詞網是詞彙網路 WordNet 的簡稱。詞語的意義會隨時間或用法而變，但詞與詞間的詞義關係則相當穩定。因此，以詞義關係為基礎的詞彙網路(WordNet, http://www.cogsci.princeton.edu/~wn/)，是表達語言內容最豐富，最穩定的資料庫之一。詞彙網路不但提供了詳盡的詞彙語意分析，並提供了由詞義出發的定義，及包括了 20 種以上可能詞義關係的標記。這些關係常常有認知上或邏輯上的必要性。因此，詞網是提供完整的自然架構與認知的基礎。換句話說，詞彙網路事實上是一個涵蓋了傳統的同義詞典所有內容的詞彙知識庫。詞彙網路與同義詞典有兩個很重要的差異：首先，詞彙網路收錄語言中所有的詞彙，而同義詞典只收有同（反）義詞的詞彙。這是如何做到的呢？正是因為第二：詞彙網路收錄了所有的詞彙語意關係，如部分-整體，上下位等。也就是說，它建立在每個詞都會與其他詞有詞彙語意關係的基礎上。相反的，同義詞典只限於同/反義關係。詞彙網路的建構與研究已經有相當的時間，自 1985 年普林斯頓大學心理與認知實驗室就在 George Miller 教授的領導下開始規劃進行。換句話說，詞彙網路原來的用意之一，其實是希望利用人類詞彙間的關係，來探討人類認知中的概念結構。

詞彙網路構成的元素其實就是表達特定語言所代表的知識體系所需的要素。首先是該語言內所有的詞彙。這裡形義並重，把任何一個詞形 lemma 與詞義 sense 的獨特配對為一個詞彙。詞義網路架構的第一個準則是以詞義為基準，把有相同詞義的所有詞彙放在一個同義詞集（SynSet）。這麼一來，同義詞集即是表達相同概念的所有詞的集合。也就是說，當一個語言（或次語言）經過這樣的分析歸納之後，這個語言裡所有的概念都整理出來了。可能有一個以上的方法來表達某些概念，但是當把所有的詞意的集合列出來之後，這個語言裡所有的概念都已經有了。詞彙的區辨是以意為主，以形為輔的。意義不同，不論是否共用一個詞形，一定要視為兩個詞。意義若不分，當有一個以上可互換的詞形（如異體字）時，則視為同一個詞彙。如英文的 bank 這個詞，講銀行是寫 bank，講河岸也是 bank，講船跟飛機傾斜的動詞也是 bank 等等之類的，或者說你到某個銀行開戶等，也叫 banking、to bank，動詞也是 bank。這麼多意義，它每一個意義跟每一個形的組合就是一個詞彙。在詞彙網路中，bank 這個詞形共有 17 個詞義。這個就是**依義不依語**：根據它的意義。每個意義是一個單位，而不因它外表的語音或文字形狀相同而歸做一個單位。如此，每個語言所能表達的所有概念，就是所有詞義 sense 的集合，正好在詞彙網路中表達出來。

其次，更重要的是一組基本的詞彙語意關係。這是因為意義可以經由關係來定義。一個詞的意義是什麼？怎麼樣去表達它呢？有些基本的概念幾乎是無法定義的。但是這些詞意跟詞意之間，一定有些關係存在。而這些關係是事先定義過，固定且有限的。我們認為詞彙驅動的概念 Synset 是知識的單位。也就是說我們

用語言來表達知識的時候，最小單位是一個概念。這個概念的具體呈現就是——詞彙。任何一個概念的實質直接描述，當然不容易。特別是所謂的基本概念單位，本來就找不到任何比他更基本的概念來描述。但是，詞彙一定會與其他詞彙產生關係，而詞義關係的網路就是概念的網路。換句話說，概念的關係可以藉著詞義使用的搭配與環境表達。所以，語意關係（Semantic Relation）不但說明了概念連結與知識衍生的基本關係，更提供了語言理解的知識背景架構。這正是我們採用知識本體 SUMO 做為上層架構，但是以詞網作為完整知識表徵系統的原因。

4.2. 漢字的知識本體；外顯分類與知識架構一例

從知識本體與詞彙網路的觀點，中文是非常特殊的書寫系統。因為中文的組成，大部分是由意符與聲符共同組成。而意符，即所謂的部首，表現出來的是一個知識分類的體系。早在許慎的<<說文解字>>就針對這 540 個意符分類。雖然許慎並無知識本體的概念，他對部首的知識分類，還是適用於現在。

最近有兩篇即將完成的博士論文，分別嘗試由說文的 540 個部首著手，建立漢字的知識本體。除了研究上讓我們更深入瞭解中文字的孳乳過程，及構詞與意義的互動外；其實在實際的語言工程方面，也可建立更方便的系統，讓中文處理時，也可使用必定存在的部首訊息。

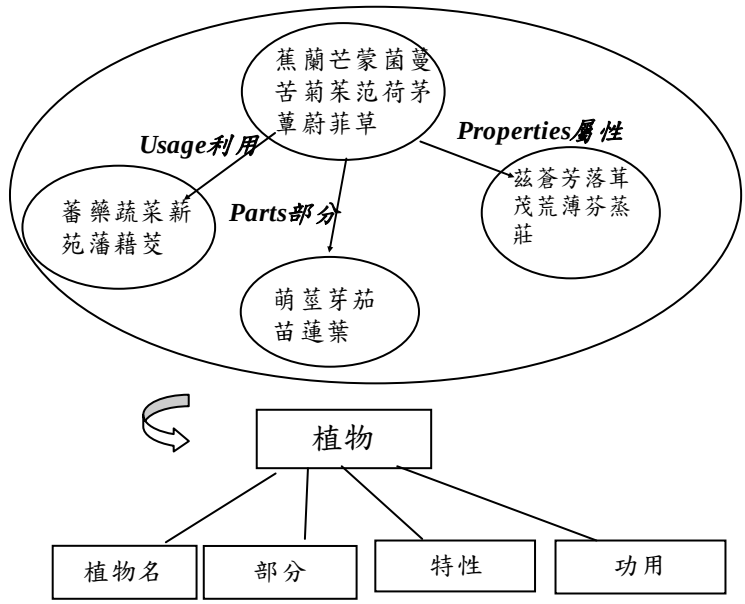
進行漢字知識本體的博士論文的兩位分別是台大的周亞民與德國圖賓根(Tubingen)大學的謝舒凱。限於篇幅，本文只介紹周亞民的部分研究成果。這兩篇論文完成後，對漢字代表的知識系統，一定有深入的突破。

周亞民的漢字知識本體，其出發點是要把中文歷代的語言與詞彙演化，以詞彙庫的方式表達。因此他除了蒐集大量各時代的中文使用，與其使用的意義，更作出時代的區分，以及新詞義從何時產生及使用。因此他也參考了大量文本。這個漢字知識詞網的完成，將提供研究中國歷代語義與文化知識架構的基礎。

對 540 說文的初步研究，顯示出他們的知識系統性。540 個概念比 SUMO 的接近一千個概念節點少。漢字知識本體與 SUMO 的直接對比中，也顯示出 2000 年前漢文化與現在知識系統的對比。漢字知識本體，與上層知識本體 SUMO 對比，也顯示了系統性的差異。主要在節點集中於較具體的知識範圍。更進一步的分析，特別是加入了歷代漢字知識本體的演進資料，應該可以提供我們知識系統與漢字表義方式演化的重要分析。

周亞民的漢字知識本體初步研究中，已經有了一個有趣的重要結果。部首作為意符的傳統看法，早已建立部首就是語意分類的分析。這個分析也廣為國際學者，包括漢學家和語言學家所接受。但是，從知識關係著手，周亞民發現，同一部首所帶領的一組字，其實與部首所帶的概念間，有比較複雜的關係。並非是單純的分類關係。底下以原來認為代表植物類的艸部為例。

漢字部首 艸(艹)的詞義分類



圖一。漢字艸部的詞義分類

上圖中顯示了部首的知識體系重並非只有單純的分類關係,也就是說，並非所有屬於艸部的漢字，都是一種（ISA)植物。在概念上，屬於艸部的字，除了屬於植物類的分類關係外，還有專指植物部位的一組字（如：苗，葉），描述植物特性的一組字（如：芳，落），及描述其功用的一類（如：藥，薪）。這裡顯示的是漢字部首構成一個極具衍生力的知識體系。這個結果，也與最新的語言學理論符合。普士德耀夫斯基(Pustejovsky 1995）的衍生詞彙理論（generative lexicon)[8]對詞彙的認知與運算衍生能力，提出了創新性的分析，解決了以往理論詞彙語意學所不能解釋的問題。而他的理論核心之一，就是說任何一個詞彙概念，可以在適當環境下，衍生不同的詞義。但是，可衍生的詞義關係，必須在詞義的經驗結構(qualia structure)中。而他提出的經驗結構，分成了四個層面，物質(formal)，組成(constitutive)，功用(telic)，產生(agentive)。物質就是對該類的直接描述（如材質，外觀），組成就是列出該類可以被分析的部位（如動物有腳），功能就是該類用於人類生活中的典型功能（如筆用於書寫），而產生則是該類如何產生或被製造（如小說是被寫出來的）。草字部首，在分類概念外的三種知識關係，正好是與衍生理論中的物質，組成，與功用三個層面符合。而因植物非人造物，沒有產生這個層面的意義衍生，也是可以預測的。

以上這個結果，消極的意義，是替普士德耀夫斯基的衍生詞彙理論，找到了強而有力的跨語言與跨時代證據。積極的意義，則是證明了漢字的知識本體，其實不是單純的分類系統。在外顯的分類階層(taxonomy)下，漢字部首表達的知識本體，已經具有衍生詞彙一樣強的衍生與知識推導能力。，

[8] James Pustejovsky. 1995. The Generative Lexicon. Cambridge: MIT Press.

4.3. 中文的詞彙語意架構：關係與分佈的內涵知識一例

語言中知識的表達，有時候是藉著系統性的對比，而不是藉著規律性的定義表達的。中研院的同仁們最近對「聲」「音」這兩個近義詞的研究[9]，是一個蠻好的例子。「聲音」是基本常用詞，在一般說話者的語言知識上，大概不能也不會把這兩個字的意義區分。但是深入的詞彙語意研究後，我們發現了，實際使用上，「聲」用於聲音的產生，而「音」用於聲音的感知。也就是說，production 與 perception 這兩個認知科學上的重要概念，其實在中文的知識系統裡，早以詞彙化表達出來。這個分析支持我們認為語言是完整知識本體的想法。只是語言的概念結構，常常是詞義關係或詞彙對比表達，必須再進一步發掘。這樣的隱涵知識系統與關係，就是我們想藉中文詞網發掘的知識體系。

在研究院語料庫中，「聲」的共現詞彙總共有 425 種，「音」的共現詞彙總共有 168 種。可見他們的構詞、衍生力都強。我們先看多為成語的四字詞，如：(9)

• 先聲奪人	• 靡靡之音
• 異口同聲	• 餘音裊繞
• 聲淚俱下	• 弦外之音

從成語中可看出，含聲的成語描述的是發出聲音的事件，而帶音的成語則描述對所接受到聲音的感知與評價。除了四字詞結構的詞彙，二字詞與三字詞結構的「聲」(總共有 372 筆資料)與「音」(總共有 160 筆資料)，經過語料分析，可以很清楚得知，大多數的詞義組合，大致可呈現如下：

(10)「聲」的構詞共現—詞義組合

發聲來源、器具等(source) ＋ **聲** 發聲動作(action)

而其分佈與例子如下：

[9]洪嘉馡,**黃居仁. 2004.**「聲」與「音」的近義辨析：詞義與概念的關係. 漢語詞彙語意研究的現狀與發展趨勢國際學術研討會. 11 月 7-8 日，北京大學.

(11)

聲		
二字詞與三字詞的語義組合	發聲來源、器具等(source)	96
	發聲動作(action)	236
其他		40
Total		372

(12)a 屏氣凝神之餘，只聞巨輪沈重而有力的<**馬達聲**>，破浪前進。

b 她的古銅色肌膚透著健康俏麗，銀鈴般的<**談笑聲**>，吸引住眾人的目光。

(13)「音」的構詞共現—詞義組合

接收器（receiver） 接收動作（action）	＋ **音**

而其分佈與例子如下：

(14)

音		
二字詞與三字詞的語義組合	接收器(receiver)	38
	接收動作(action)	58
其他		64
Total		160

(15)a 防震材料如泡綿等，也許可以稍微減少<**走音**>的機會，減少調音的麻煩。

b 有些老人說話帶著濃重的<**鄉音**>，溝通困難。

由上述對於「聲」與「音」的構詞共現比對，以及實際語料中的例句來看，

「聲」的構詞共現詞彙數多於「音」的構詞共現詞彙數，其主要的原因是，「聲」是聲音的來源，是製造聲音的開端，是將聲音傳送出去的，只要是可以發出聲音的來源，就可以各種不同的器具、動作來產生而成的，受限的條件也就會比較少，相對地，就可以有各種組合而成的聲音來源。

反之，「音」是聲音的接收，是把聲音收回來的，不論接收者是生命體或無生命體，或者是接收的動作，都是接收者在收到訊號後，必須經過適當的解讀、認知概念的理解、機器的轉換等過程而得到的。接受認知分類的結果，當然是要歸納為有限的類，因此類的總數會受限。而且就接收者而言，接收聲音是單一過程，且處於接收聲音的終點，受限的條件較為複雜，當然就不會有太多聲音接受。

雖然「聲」與「音」都是可以用來表達與聲音相關的概念，但是從詞彙的核心詞義出發，區分兩詞彙的獨特性質，以辨析兩者在聲音傳送的歷程中所扮演的角色，再經過不同觀點的分析與討論，了解「聲」的詞義成分含有聲音的製造者，通常是隱含著某種聲音的製造方法，是傳送聲音的源頭；「音」的詞義成分則含有聲音的接收者，通常指射接收一段聲音的訊息，且必須經過接收者的認知、理解等過程而得到的概念結果。

因為「聲」與「音」在詞彙功能上的差別，影響了兩詞彙在構詞共現上的顯著對比，當然，詞義功能的差異也會影響到兩詞彙在構詞共現上的構詞結構。以(17)和(18)的共現分佈情形來分析，不論是聲音的產生或接收，雖然動態性動作多於靜態性人事物，但兩種情況仍都存在，也就是說，「聲」與「音」同時存在兩種複合詞的共現，四種不同組合的構詞結構，如下：

(17a)　「聲」與「音」在名名(N＋＊)複合詞中的對比

聲	**音**
• 歌	• 嗓
• 掌	• 鄉
• 人	• 喉
• 腳步	• 裝飾
• 風	• 尾
• 鐘	• 哨
• 水	• …
• …	

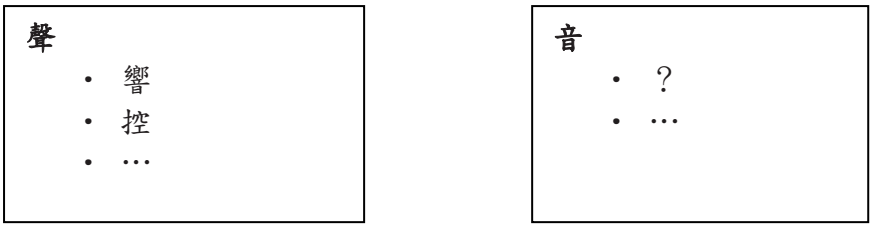
(17b)　「聲」與「音」在名名(＊＋N)複合詞中的對比

聲	音
• 帶	• 高
• 色	• 域
• 道	• 段
• 調	• 節
• 波	• 階
• 量	• 群
• …	• …

(18a)　「聲」與「音」在動名(V＋＊)複合詞中的對比

聲	音
• 叫	• 注
• 笑	• 錄
• 呼	• 拼
• 哭	• 播
• 鳴	• 調
• 放	• 正
• 發	• 走
• …	• …

(18b)　「聲」與「音」在名動(＊＋V)複合詞中的對比



以上四組例子，再次驗證了「聲」「音」的意義對比，在複合詞的環境中也成立。我們可以很有把握的說，這個發聲與聽音的概念對比，是中文詞彙知識系統中的一部份。

5. 結語

本文以中文內部表達語意與知識的系統為研究對象。由漢字表義的特性出發，釐清中文字與詞間複雜的關係。並以知識本體與詞網的觀點，來探討漢字內涵的知識系統。我們特別提出最近漢字知識本體的研究，及一些系統性的成果。由幾個例子中，我們希望說明的是漢字與中文都具有強健的知識表達系統，系統本身即有豐富的知識。我們目前的初步研究，只是一小步。後續研究，希望能對漢字文化的知識傳承與衍生，甚至人類的認知系統，提出解釋性的新發現。