知識天地

語言，文字，與知識架構-由漢字出發的知識本體研究

黃居仁（語言學研究所研究員）
周亞民（景文技術學院資訊管理系助理教授）
謝舒凱（語言學研究所博士後研究）

一．概述

語言是傳達與表達知識的約定俗成系統。語言研究的迷人之處，也是最大的挑戰，在於語言不但是知識的載體，也是知識的內容。語言工程的研究，強調語言的載體功能，著重如何把知識從載體分離。認知語言學的研究，強調語言是人類共享的知識表達系統，著重在如何由語言的系統中，理出人類認知行爲的特性。這兩個尖端研究，雖然都以語言爲標的，卻鮮少交集。我們的研究，強調語言是知識系統，可以兼顧語言科技與語言理論的研究需求。知識系統的研究觀點，不但使我們能宏觀研究完整的語言體系，進一步對人類的認知行爲，提供系統性的解釋；更因爲系統的完備性，也可以有語言科技上的有效應用。

語言作爲知識系統的研究，有兩個重要的研究背景。第一個是方興未艾的知識本體（ontology）研究。知識本體的研究，雖然有長遠的哲學研究傳統；但近來發展的最直接動力來自於「語意網」（Semantic Web）的相關研究。Tim Berners-Lee 於 2001 年在《科學美國人》中宣告語意網的願景是電腦能自動「閱讀」並理解網路資料的內容。而這個閱讀語意的能力，主要來自每個網頁上，都要以知識本體宣告網頁中知識的架構與內容。因此，如何建立共用的知識本體架構，能夠正確有效表達不同領域知識，並能跨越不同語言的知識鴻溝，成爲在相關學術領域中受到重視的研究議題。

另一個重要的背景是語言與文字跨越時空的約定俗成特性。使用同一個語言的人之所以能夠彼此瞭解，光靠使用同一套符號系統是不夠的；他們必須同時使用同樣的知識表達系統。這與即使全世界的網頁都有知識本體，還是不能保證網頁的意義能被所有的電腦閱讀，必須還要共同的知識架構；有異曲同工之妙。換句話說，任何語言都有內涵的知識本體，所有講這個語言的人，都不自覺的在使用這個隱含的約定俗成系統。從這個觀點出發，所謂「語言知識本體」（linguistic ontology）的研究，開始受到重視。從語言到文字，是更進一步的約定俗成。中文的漢字書寫系統跨越三千年，又爲漢字文化圈的不同語言所共同採用。因此他的隱含知識表達系統，可以說是全世界使用人口最多、最穩定，表達知識也最豐富的知識本體。我們進行漢字的知識本體研究，不但有語言與認知理論上及知識工程應用上的意義；也可以提供未來的中文人文學研究，一個知識爲本、系統性和可規範化的研究基礎架構。

我們的研究有兩個相輔相成的方向。一個是由漢字書寫系統本身的知識表達系統出發，把這個系統轉換成知識本體的規範表達形式，並提出解釋；也同時提供了漢字歷時與跨地理區域變遷的描述模式。這個研究，我們稱之爲「漢字知識本體」（Hantoglogy）。另一個方向，著重於概念化的表達，從概念集的觀點出發，把漢字系統看成知識本體與詞彙間的介面。這個研究，著重於漢字的表義系統在語言學理論中的定位。相較於文獻中已成爲標準的詞網（wordnet），這個研究我們稱爲「漢字網」（HanziNet）。

二．漢字知識本體

漢字知識本體（Hantology）表達的知識包括字形結構、意符、聲符、古音、中古音、現代音、字義、異體字關係和詞彙衍生，以及不同時期的形、音、義的變化和關係。漢字書寫系統的特性是採用意符，漢字書寫形式和字義是意符的衍生，因此，意符表達的知識系統是漢語書寫系統的核心。我們以說文的五百四十部首作爲漢字的

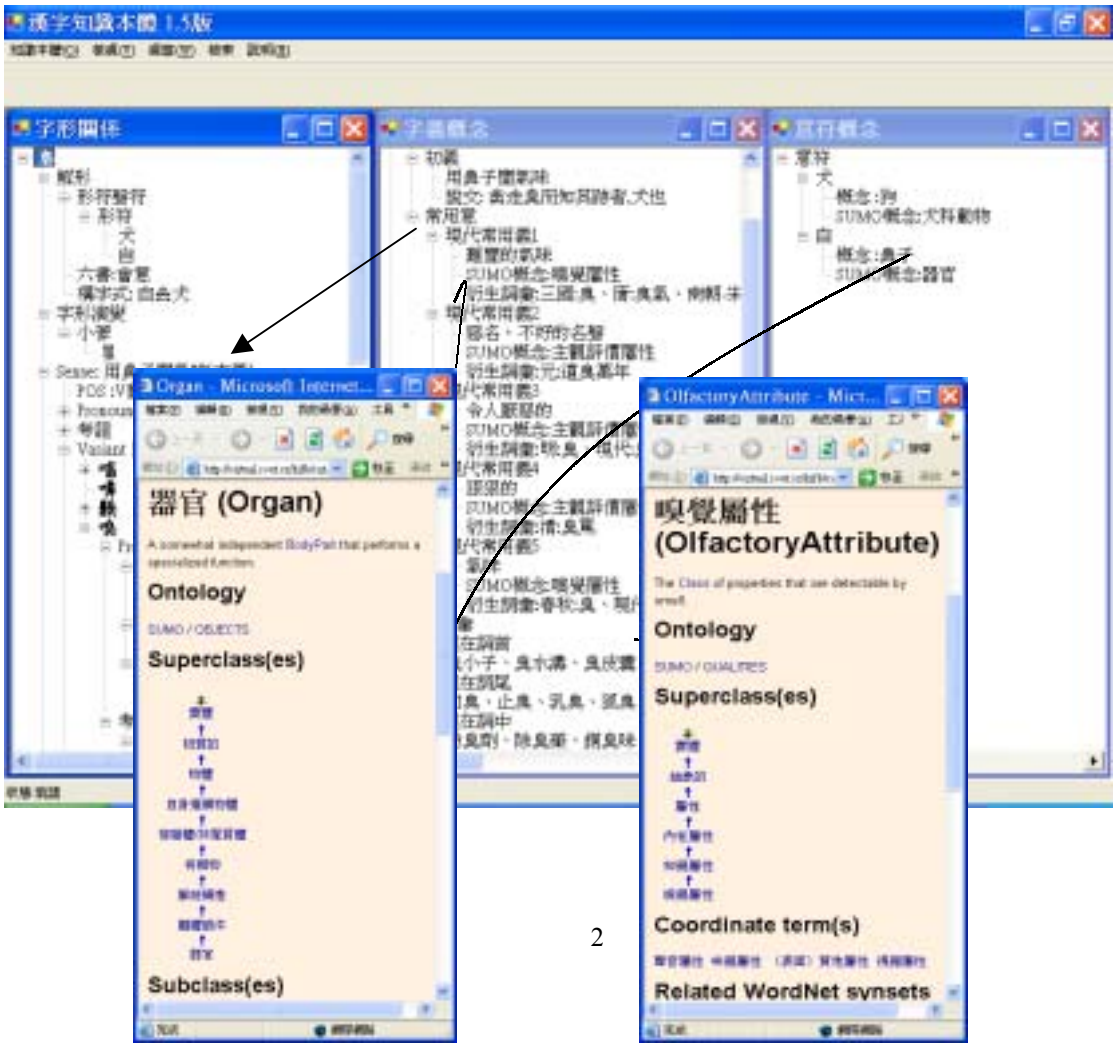
基本意符，分析部首作爲意符時所表達的概念，運用 IEEE Suggested Upper Merged Ontology（SUMO）表達和呈現意符的知識系統，還可以讓計算機直接利用漢字的意符進行推理。

漢字字義區分爲本義、引申和假借義，以表達字義的擴展和變化。本義以說文釋義爲依據，並描述音義之間的關係和變化，以及不同字義的衍生詞彙。漢字知識本體透過表達字義和衍生詞表達此重要的特性。由於漢字與漢字之間並非獨立，而是彼此有關係，表達這些關係主要考慮的是能夠反應漢字書寫系統的特性，尤其是相同的詞或詞素可以用不同書寫形式而產生異體字關係的特性。爲了能夠表達複雜的異體字關係，我們建立了由字音、字義、聲韻、構詞和時間所構成的異體字語境（context），以描述不同漢字在什麼語境可以交替使用。

漢字知識本體能夠有系統的表達意符到單字詞和詞素，與單字詞到雙字詞的發展過程，以及書寫形式、字義、聲韻、詞彙衍生和異體字的關係和變化。同時爲了讓漢字知識本體的知識更容易於被分享和利用，漢字知識本體的描述使用語意網的知識本體表達語言 OWL-DL。以提供計算機在自然語言處理所需的書寫、構詞、語法知識。

漢字知識本體這個基礎架構建立後相關研究中成果之一，是發現漢字知識表達系統直接表徵了人類的認知模式。過去的研究常把漢字的意符（部首），看成不完美的分類系統。但是，我們的系統性研究發現，每個一幅所衍生的漢字家族，事實上是一個小的知識體系。而這個體系，是建立在人類認知的顯著相關性（saliency）上的。以意符艸爲例，原先以爲他代表的是植物這一類。可是事實上，意符艸構成的漢字包括了不同的植物（蘭，菊），植物的部分（葉，芽），植物的功能（藥，薪），植物的樣貌（芬，蒼）。以意符金爲例，除了金屬名稱（銀，銅）外，也表現了金屬的功能（鐘，鎖，鏡），與金屬製造的過程與方法（鍛，鑄）。這與希臘哲學提出知識的經驗架構（qualia，近來爲 Pustejovsky 的衍生詞彙理論所引用）正好相符。也就是說，人類知識的衍生，主要的是由分類，組成，功用，與產生來源，這些經驗知識來的。我們證明了漢字意符表達的知識系統，正是建立經驗知識的基礎上，也驗證了知識的共同認知基礎。

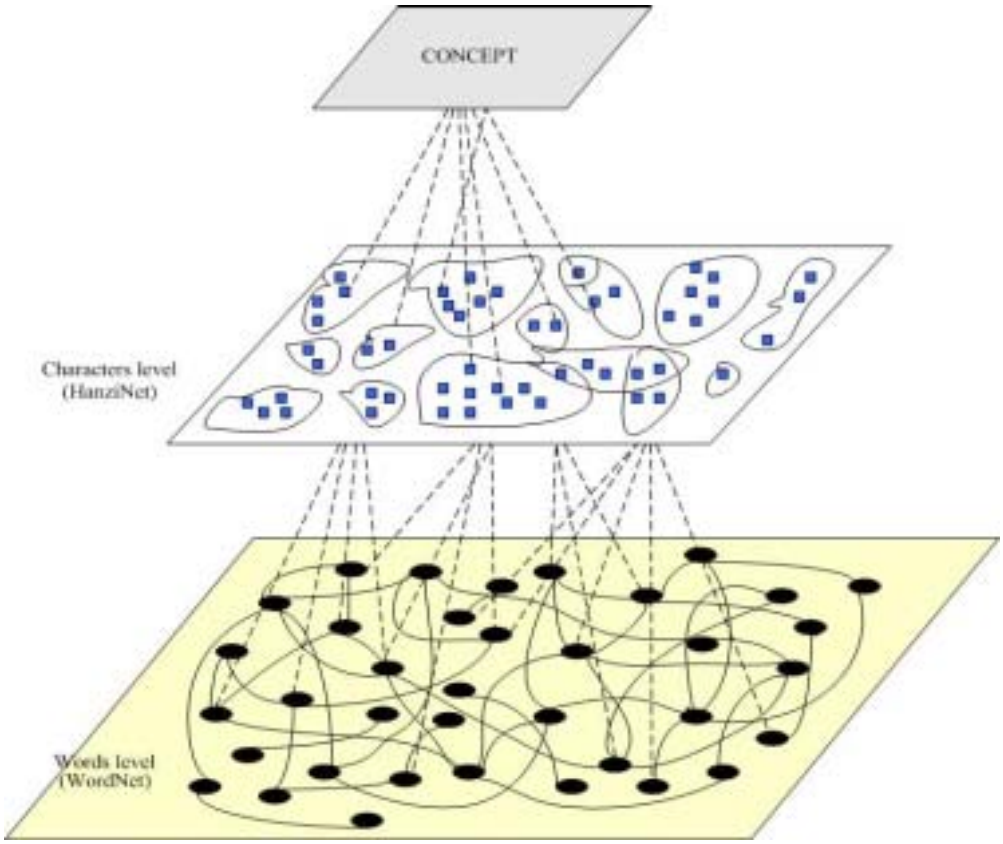
漢字知識本體的網路介面還在構建中，圖一提供了這個知識架構資料庫的一個例子：



圖一、漢字知識本體與 SUMO 的連結

三．漢字網

「漢字網」（HanziNet）從一些字源資料中（包括從意符訊息、訓詁學之本義推敲），彙整了一部「漢字知識本體字彙集」（Hanzi-grounded Lexicon）。其中內容包括了每個字的「概念義」（平均 2-3 個）、組字之統計訊息、與意符之劃分與表達。這是漢字網的第一個核心組成成分。漢字網的第二個核心組成成分是「知識本體」的「概念集」（conset）。相對於詞網之「同義詞集」（synset），漢字網將概念義相同的漢字取出，置入同一個「概念集」中。目前做出了存在三百零九個「概念集」之預設，以及標明其上下位關係之概念階層表。並且將上述「漢字知識本體字彙集」中之千餘漢字作爲實例 （instances），充實了（populating）這個知識本體架構。漢字與概念及語言知識本體的關係如圖二。



圖二： conceptual ontology,characterized ontology and lexicalized ontology

爲了與其他國際資源接軌，目前正在進行將「漢字網」所提出的 309 個「概念集」（conset）與「歐語詞網」（EuroWordnet）所提出的 164 個共享基本概念集（Base Concepts）作匹配，接下來則朝向與 4689 個 Common Base Concepts 作匹配工作，以期「字彙驅動」的概念知識與「詞彙驅動」之詞彙化概念知識能夠相互參照。漢字網下一步之重要目標，包括與「中文詞網」之整合介面 （interfacing）問題，以及作爲一種知識資源，如何協助進行中文詞網之語義關係預測。

四．結語

Multilingualism 是當今全球共同面臨的重大挑戰之一，最重要的難題在於語言與知識的不對稱性，以及不同語言間知識表達的鴻溝。我們希望語言知識本體的研究，在結合了語言學、人文科學，認知科學與資訊科技的基礎上，可以建立重要的橋樑。在解決科學問題的同時，也能協助人類面對挑戰。