中文詞彙網路：跨語言知識處理基礎架構的設計理念與實踐

黃居仁[1] 謝舒凱[2] 洪嘉馡[3] 陳韻竹[1] 蘇依莉[1] 陳永祥[4] 黃勝偉[1]

[1]中央研究院語言學研究所，台灣

[2]台灣師範大學英語學系，台灣

[3]台灣大學語言學研究所，台灣

[4]台灣大學資訊工程學研究所，台灣

摘 要：中文詞彙網路(Chinese WordNet, 簡稱 CWN)的設計理念，是在完整強健的知識系統下兼顧詞義與詞義關係的精確表達。中文詞義的區分與詞義間關係的精確表徵必須建立在語言學理論，特別是詞彙語義學，的基礎上。而詞義內容與詞義關係的發掘與驗證，則必須得自實際語料中。我們採用的方法是與分析與語料結合。結合的方式則除了驗證與舉例外;更包括了在大量語料庫上，平行進行詞義標記,以反向回饋驗證。完整強健知識系統的建立，則是兼顧了知識本體(ontology)的完備規範(formal integrity)，以及人類語言系統內部的完整知識。我們採用了上層共用知識本體(SUMO)來提供知識的規範系統表徵。

关键词：中文詞彙網路；全球詞彙網路網格；知識本體；多語處理；跨語言整合

Chinese Wordnet: Design, Implementation, and Application of an Infrastructure for Cross-lingual Knowledge Processing

Chu-Ren Huang[1] **Shu-Kai Hsieh**[2] **Jia-Fei Hong**[3]
Yun-Zhu Chen[1] **I-Li Su**[1] **Yong-Xiang Chen**[4] **Sheng-Wei Huang**[1]
[1]**Institute of Linguistics, Academia Sinica， Taiwan**
[2]**Department of English, National Taiwan Normal University, Taiwan**
[3]**Graduate Institute of Linguistics, National Taiwan University， Taiwan**
[4]**Department of Computer Science and Information Engineering, National Taiwan University， Taiwan**

Abstract: The design concept of Chinese WordNet (CWN) is to build a complete and robust knowledge system which also embodies a precise expression of semantic relations. Such precise expression for the Chinese sense division and the semantic relations needs to be based on the linguistic theories, Lexical semantics especially. The source and the instances of the lexical content and their semantic relations in CWN are all from the practical language materials. The method adopted in the system is to analyze the language materials and then combine the analyzed result with other corpus by using sense tagging to reexamine the accuracy of the analysis. A complete and robust knowledge system needs to equip with not only the complete knowledge of human languages but also the formal integrity of ontology. Therefore, we adopted the Suggested Upper Merged Ontology (SUMO) as the system structure to present the knowledge.

Key Words: Chinese Wordnet; Global WordNet Grid; Ontology; Multi-language Processing; Cross-lingual integration

1 前言

為達到提供完整中文詞義區分資料以及呈現豐富語料與思慮完備的語言分析結果，透過網際網路平台開發一中文詞彙網路，是資訊基礎建設完善環境中最適當的選擇。我們的研究團隊—中文詞網小組(Chinese WordNet Group)，結合分析詳盡的中文詞彙詞義資料與網路科技的技術，初步開發了中文詞彙網路(Chinese Wordnet)，以利於提供中文詞彙詞義的相關訊息，便於從事中文詞彙詞義的研究所需。

在語言內部知識的完整表達上，則是建立在完整的詞義關係系統上，特別是利用「類義詞」(paranym) 整合對比語意關係為主的詞彙網路與界定語意場的不同分類系統(taxonomy)，更以完整標記的跨語言詞義關係作為多語知識系統對應的基礎。以上的設計理念，使得中文詞彙網路不但提供了中文詞彙語意深入研究的基本參考資料，更進一步能支持跨語言的知識合與應用，如全球詞網網格 (Global WordNet Grid) 的建構與生態環保領域的跨語言知識整合 (Kyoto 計畫, Vossen et al. 2008)。

本研究以中央研究院語言學研究所中文詞彙網路研究小組從 2003 年以來大量的詞彙詞義分析研究成果為基礎，收錄至 2008 年初的第五版資料，已有約 7,000 個詞形，20,000 個詞義深入分析之 synset 資料建構於中文詞彙網路上，以人性化整合查詢介面透過網際網路呈現，除了提供相關研究人員以及有興趣的使用者查詢檢索外，更希望藉此系統作為全球詞網—多語、跨語言的基礎知識架構對應的連結。

2 中文詞彙意義的知識檢索研究發展

2.1 詞義與義面區分的基礎與應用

在理論語言學上表現出來的，是詞彙語意學理論與解釋能力的蓬勃發展。而在計算語言學方面則是語意導向的研究題目逐漸成為主流。甚至在認知科學的研究中，意義的心理與腦神經處理問題也開始受到重視。在這些潮流中，詞網 (Wordnet) 是最重要的共同基礎架構，而詞義 (sense) 的區分則是最關鍵的基本研究議題。

詞網是以詞義與語意關係為經緯建立的人類語言知識表達基本架構。建構完成的詞彙語意網，一方面可以作為語言學研究的素材，另一方面在資訊處理上又可以作為自然語言處理以及諸多實際應用的基石。詞網裡有兩項重要的元素，一是以詞義為據的詞彙分組(即所謂的同義詞集(synset))，另一個就是連繫詞集的語意關係。為達到提供完整中文詞義區分資料，以及呈現豐富語料與思慮完備的語言分析結果，透過我們所開發的平台—中文詞彙網路，是資訊基礎建設完善環境中最適當的選擇。

以同義詞集為節點，透過語意關係相互連繫，就形成了表徵詞彙意義及其關係的語意網絡。其中，同義詞集的建立可說是最基礎的工作。建立同義詞義，便是把在語境中能表達相同詞義的詞彙歸為一組同意詞集，而多義詞則分處多組詞集，以表示其不同的詞義。據此可知，詞彙的詞義區辨及其同義詞的判斷與滙集，便成了最根本的工作。然而，雖有實際的需求，詞義區辨的原則在學術上卻是尚無定論的議題。爲了相關工作的進行，本文希望能討論並建立一組詞義區辨的操作原則，一方面能滿足一致性與合理性的要求，另一方面又能作為大量中文詞彙詞義區辨工作上有用的準則。有了一致性的詞義判準，語言知識才能有效處理，也才能把語言知識連結到知識本體 (ontology) 或轉換成概念表達。

2.2 詞義判準原則

在本文中，「意義」與「詞義」二詞有特殊的界定。分析人員根據自身對某一詞彙在語境中傳達訊息的理解，區分出相同詞彙 (形) 的不同涵義，我們稱之為「意義」(meaning)；不同的人可能有不同的區分方式 (依據不同的標準或直覺)。進一步，根據適當的標準，判斷初步析分的合理性、進行意義的分合、細分等而獲致的最後結果，我們稱之為「詞義 (sense)」。在某些語境下，詞可能有受語境影響而改變的意義，人們可以區分出，但這些意義是暫時的，當下文語境改變時，又會出現不同的相對意義。這樣的意義區分，我們稱之爲「義面」(meaning facet)，是中文詞網中處理文獻中所謂「規則化多義」(regular polysemy) 的重要創新。如同一段文本中，「報紙」可以指涉閱讀的內容，或紙製品的實體。[1]以及規則化的動詞名物化，都是義面的重要例子。

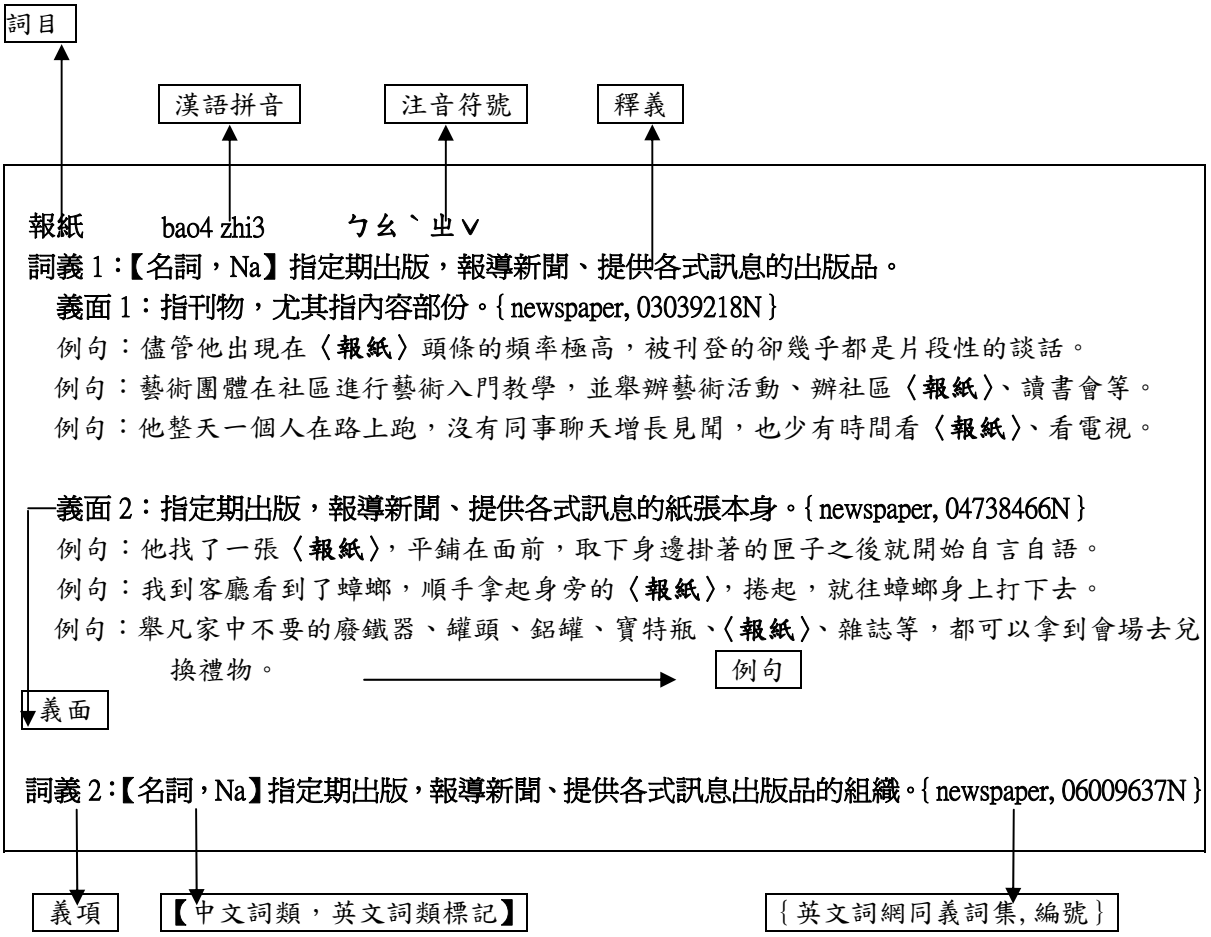
我們的詞義判準建立在五個基礎原則上：(一)一義一項、(二)一物一義、(三)一事一義、(四)義不隨境遷和(五)義面由觀點與語境定義。除了建立理論完整的詞義區辨原則外，並同時提供了可以實證的詞義區辨操作原則，並對每個原則提供實例。我們藉由這個分析原則進行詞義區分的工作，並建構工作介面，將中文詞網詞義區分資料庫的作業已全部線上化，中文詞網詞義區分的資料可直接進入資料庫，不用透過機讀格式的轉檔。

[1] 例句如：「今天的報紙登的是有趣的新聞，過期的報紙可論斤回收販賣；但也有人專愛在舊就報紙裡找漏網新聞。」報紙在整段中式連貫的主題，理論上必須有相同的詞義 (sense)，但又有不同的意義，故以「義面」區分之。

在本介面上，我們可以進行詞彙的查詢，詞義的新增、修改以及例句和 WordNet 同義詞集的查詢和輸入。本資料庫將可有效地管理詞彙與詞義，並更便利於技術報告的整理和編輯。此外，我們也藉由 Chinese WordShetch 提供較為明確的上下文語境的句子，以驗證我們分析的詞義是具可靠性的(Hong et al., 2007(a))。

2.3 詞義描述規範

黃居仁 等(2003)所提出之詞義區辨原則與操作原則，是中文詞義資料庫建檔與《詞義區辨小詞典》編纂的依據。《詞義區辨小詞典》所收錄的詞條(entry)，以現代漢語通用語詞為範圍，不列入現今已不用或罕用的詞彙。而收錄的中文詞彙條目，包含單字詞、雙字詞和多字詞。本詞典盡可能提供各詞目(lemma)完整而且正確的訊息，包含標音(漢語拼音與國語注音)、釋義、英文對譯、詞類、例句、附註，如圖一所示：



圖一、中文詞彙條目內容範例

3 中文詞彙知識檢索系統設計

一般全文檢索系統，只能以所檢索的標的文件所含有的文字資訊進行檢索，無法就其字詞義或周邊相關資訊進行檢索，這種檢索功能顯然不能滿足語言研究的需求。由過去相關研究中可以整理出語言學研究所需要考量的各項詞彙資訊，本研究以中文詞彙為研究對象，經過嚴謹的分析研究後，對每一個中文詞彙呈現出詞目、詞義、領域、釋義、語義關係、英文對譯、例句、附註等內容。經過嚴謹分析的詞彙資訊，除可有系統性地保存詞彙知識外，更可滿足多元的語言學相關研究使用。

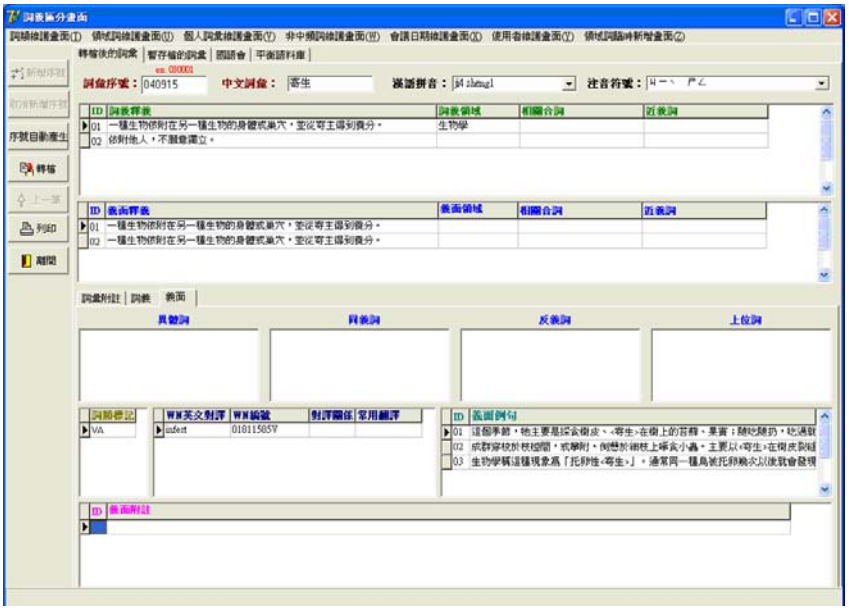
中文詞彙網路小組的研究成果，從 2003 年初起，至 2008 年 5 月止，目前累積的成果，共有超過 8500 個詞形，超過 20000 個詞義。為便於引用累積的資料庫，從 2005 年起，改以版次來稱呼中文詞彙網路資料庫的內容。原先累計到 2003 年底的資料，稱為第一版；累計至 2004 年底的研究成果，則稱為第二版；而至今，累計至 2007 年底之研究成果，目前稱之為第五版。

中文詞彙知識檢索系統之開發則將上述累積之工作成果依照結構化系統分析與設計方法在網際網路上建構一工作平台提供相關研究人員查詢使用，除了研究成果共享之目的外，更希望藉此作為中文詞彙知識網路研究之基礎架構。

3.1 SSMS 系統

為了可以讓機器讀取並儲存大量的詞彙詞義區分的資料，我們以詞彙知識為基礎，來整合詞彙詞義的訊息，開發了中研院詞彙詞義管理系統(Sinica Sense Management System)，簡稱 SSMS (Huang et al.,2005)。在 SSMS 裡，包含了中文詞網小組所收錄並分析的詞條、詞義等相關訊息。換言之，SSMS 包含的詞條訊息有：詞類、例句、對應 WordNet 的英文同義詞集(synset)、詞彙語意關係如：同義詞、反義詞、上位詞、下位詞等等。

從 2004 年 2 月起，本組中文詞網詞義區分資料庫的作業已全部線上化，中文詞網詞義區分的資料可直接進入資料庫，不用透過機讀格式的轉檔。在本介面上，我們可以進行詞彙的查詢，詞義的新增、修改以及例句和 WordNet 同義詞集的查詢和輸入。本資料庫將可有效地管理詞彙與詞義，並更便利於技術報告的整理和編輯。「中文詞網詞義區分資料庫介面」如下圖所示。



圖二、Sinica Sense Management System 介面

3.2 CWN 介面

中文詞彙知識檢索系統在設計階段參酌使用者角度與系統功能發展角度，共同建立起系統架構與操作流程，詳細描述系統範圍內相關之資料結構以及操作步驟，特別是設計一套整合式即時查詢的方式(陳 等，2006)，提供系統使用者一個整合查詢介面快速查詢以及瀏覽有興趣的各個詞義資訊。系統提供的查詢範圍，有：中文詞彙、釋義內文、英文對譯、中文詞彙模糊查詢、注音、漢語拼音等，使用者可依不同訊息或不同需求來選擇查詢的方式。主要的出發點是能對詞彙與語義相關連的內容，做廣泛而有效的檢索，也是藉著檢索的比對，來確保釋義語言及語義區分的一致性及強健性。在查詢結果之呈現上，以詞彙編號為主鍵由資料庫中提取出詞目、詞義、領域、釋義、語義關係、英文對譯、例句及附註等項目依序排列，透過瀏覽器可清楚呈現給使用者。

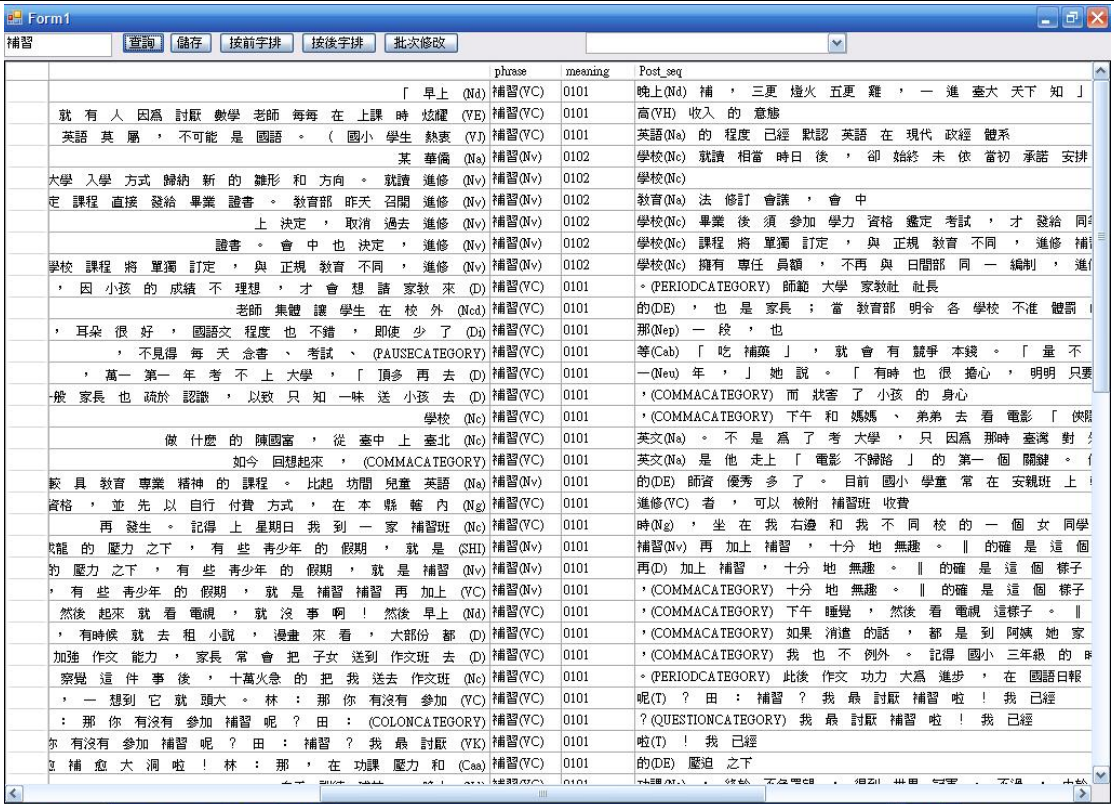
我們於 2006 年，將本組中文詞網詞義區分資料庫的成果網路化，以便提供給使用者直接查詢。目前，我們命名為「中文詞彙網路(Chinese Wordnet)」，如下圖所示。

圖三、Chinese Wordnet 首頁

3.3 Sense Tagging 系統

3.3.1 人工標記

為了徹底呈現語言的真實性，我們對於每一個詞條(lemma)的詞義(sense)和義面(meaning facet)做了詳盡的區分，同時，也藉由這些詞義和義面，開發出一套詞義標記系統(柯 等，2007；Ker et al.，2008)。中文詞彙網路的詞義，基準，承襲了普林斯頓詞網的繼成傳統，以同義詞集為節點，透過語意關係相互連繫。建立同義詞義，便是把在語境中能表達相同詞義的詞彙歸為一組詞集，而多義詞則分處多組詞集，以表示其不同的詞義。為了證實我們所分析的詞義可以完整地表現在實際語言上，我們開發了設計出一個超過十一萬詞的大規模中文詞義全文標示語料集系統，以我們已經分析過的詞義作為基礎，以中研院平衡語料庫為標示對象，從中摘錄 56 篇完整文章，利用 N-Gram 與搭配資訊等語言知識，並結合機器學習技巧以及機率模式的方式作為處理自動詞義標示的前置作業工作，最後為達高精確度之效果，再將自動產生之標示結果經由人工校訂而成。

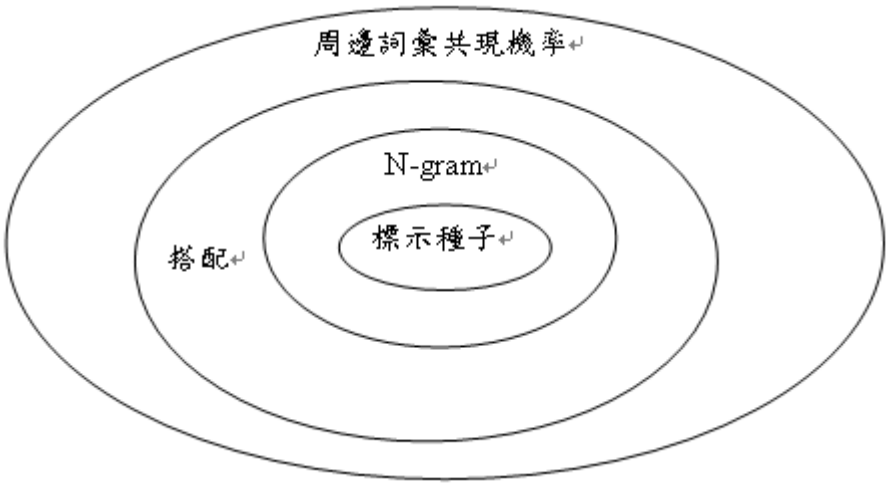


圖四、Sense Tagging 查詢結果

3.3.1 自動預測

大量精確的詞義標示資料，可提供多項計算語言相關研究的豐富素材。但是，中文語料庫詞義標記主要的瓶頸為缺乏足供自動標記參考的資料，而人工標示需要昂貴成本，造成語料庫標示語意工作的難產。近年來的許多研究，顯示出對大規模詞義標示語料集的大量需求，這些資源在建構上是否完備，往往會影響整個研究進行方向以及研究結果的正確性。為了克服此問題，本文提出一套半自動詞義標示方法，作為標示詞義的前置作業，再經由專門人士校訂。語料庫製作以中研院平衡語料庫為對象，從中摘錄文章，並對摘錄出之文章中之詞做詞義標示的動作，設計製作出一個大規模的中文詞義標示語料集以供自然語言處理研究使用（柯 等，2007；Ker et al., 2008)。

根據柯 等(2007)的研究，自動標示詞義的方法，採用誘導式方法（bootstrap）逐步放寬標示條件，來擴增標示語料，其系統組織圖如圖五所示。



圖五、標示詞義系統組織圖

自動標示詞義的第一階段採用 N-gram 模式，將標示出詞義的資料加入訓練集中，以作為第二階段的訓練語料。利用 N-gram 處理詞義標示是基於下面的假設：存在包圍目標詞彙前後 N 個詞彙完全相同的兩個子句，我們推論它們應擁有一樣的詞義。在此使用 N-gram 有兩項主要目的，第一是擴大訓練集，因語料庫中常可見到相似之子句。第二個目的是過濾訓練資料集的雜訊，以此檢驗人工標示資料之不一致性。第二個階段我們使用搭配資訊來增加標示集數量，搭配資訊是一種很強的語言關係，能決定目標詞彙之詞義。利用其所搭配的共現詞彙(collocation)與詞義特徵(semantic feature)等進行詞義的預測，第一部份詞義標示以詞義下再細分至義面為準，結果如表五所示，整體的正確率為 57.47%。至於，第二部分我們將詞義標示處理至詞義為止，不再細分義面，整體的正確率大約可提升至 64.51%。

3.4 中英雙語知識本體詞網

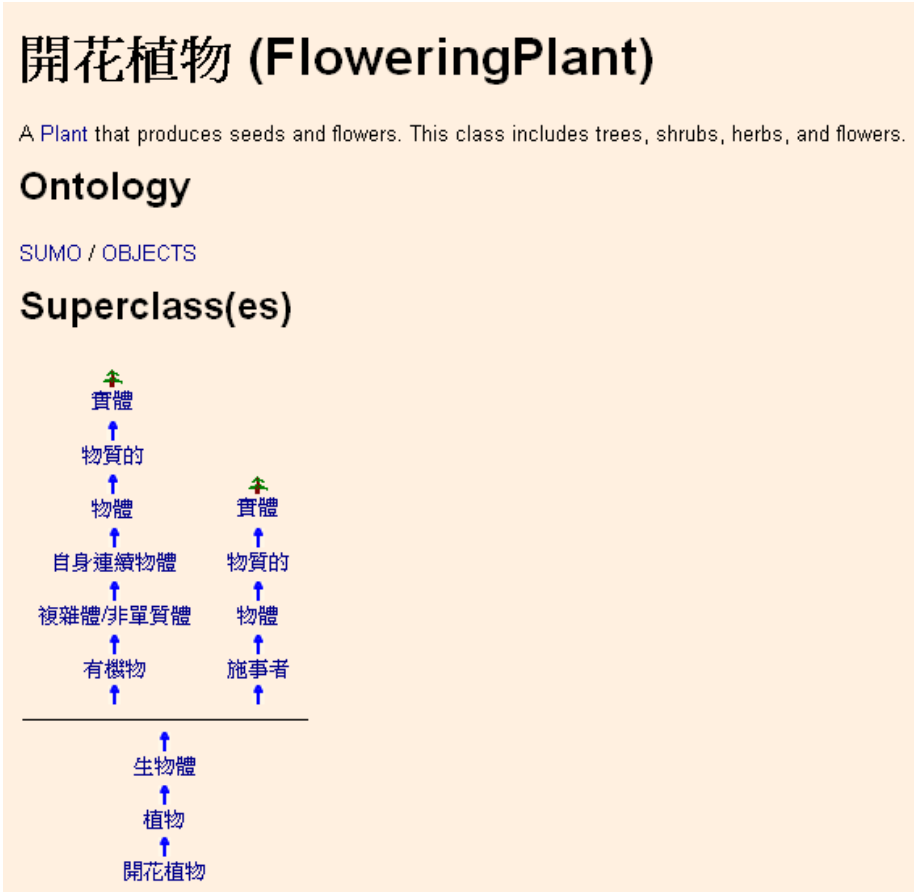
為了追求語言知識架構的豐富性，我們採用建議上層共用知識本體(Suggested Upper Merged Ontology，簡稱 SUMO)為基礎來進行語言知識的對照。在自然語言中，常會有一詞多義的現象，甚至於有模擬兩可的現象，諸如此類的詞彙，並不能以一般的詞義就能分析得精準，因此，在我們的系統裡，還引入了義面(facet)的概念，不但使我們的系統有更強的表達能力，更使我們能夠描述某些隨語境轉變成共現的語義區分，同時，我們利用類義詞則容許我們表達同類姊妹詞義間的同類聚合關係(Huang et al., 2008(a))。

中央研究院中英雙語知識本體詞網(The Academia Sinica Bilingual Ontological Wordnet, 簡稱 Sinica BOW）以 WordNet 為基礎，加入台灣地區所使用的中文經驗，搭配領域以及 SUMO（Suggested Upper Merged Ontology，建議上層共用知識本體），並以 WordNet 的 1.6 版和 1.7.1 版之名、動詞的單詞義和多詞義對應資料為基礎，作為媒介連結了領域詞彙庫和領域知識本體。系統藉由多元、友善的介面，將功能切割為詞網、知識本體以及索引三個主要單元，提供跨語言資訊轉換、詞義的區分與詞義關係的連結、語言資訊與概念架構（知識本體）的連結以及使用領域等訊息。

Sinica BOW主要使用的資源包含WordNet、ECTEC (English- Chinese Translation Equivalents Database)以及SUMO。ECTEC是以WordNet為基礎，經由現有英中或中英電子辭典的詞形對應，替每個同義詞集的詞義找出可能相對應的中譯詞組，再經由人工檢驗。尋找對譯的過程中，盡可能的以詞彙而非描述性短語表達，目的在於讓每個同義詞集都有最適當的一至三個左右的中文對譯。

SUMO則是由IEEE標準上層知識本體工作小組所建置，其目的在於促使自然語言處理、資訊檢索、自動推論以及資料互通性等工作的進行。知識本體類似於字典或詞彙表，但訊息更豐富，以便於電腦處理其內容。知識本體以格式化的方式表達概念(Concept)、關係(relation)以及公理(axioms)。上層知識本體是將一般性、後設性(meta)、摘要性以及哲學類的概念指出，所以特殊領域的概念可由其中的概念所涵蓋，但特殊領域概念的知識本體則期許由各領域自行制訂。(Niles and Pease, 2001)目前SUMO已經與WordNet1.6以及2.0版本結合，且以同義(synonymy)、上位(hypernym)、體例(instantiation)這三種類別顯示同義詞集和SUMO概念間的對應關係，例如：同義詞集cell（細胞）與細胞概念(cell)是同義。Hockey(曲棍球)屬於運動概念(sport)，兩者間的關係為上位，也就是說運動涵蓋hockey(曲棍球)。China(中國大陸)屬於國家(nation)這概念的體例。圖五與圖六，是我們利用Sinica BOW的系統查詢WordNet的詞彙，所得到的相關訊息與SUMO的訊息。

圖六、Sinica BOW系統查詢WordNet詞彙的結果



圖七、Sinica BOW系統查詢SUMO的結果

4 詞彙語意關係表達與預測

4.1 詞彙語意關係表達

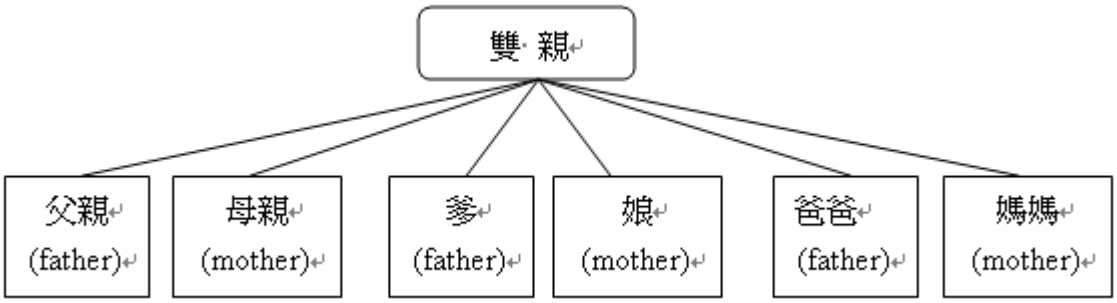
在英語及其它的歐洲語言裡，詞彙語意關係已有相當充分的研究(蔡 等，2002)。例如，歐語詞網

（EuroWordNet, Vossen 1998）就是一個以語意關係來勾勒詞彙詞義的資料庫。也就是說，詞彙意義的掌握是透與其它詞彙語意的關連來獲致的。為了確保資料庫建立的品質與一致性，歐語詞網計畫就每一個處理的語言其詞彙間的詞義關係是否成立提出相應的語言測試。實際經驗顯示，利用這些語言測試，人們可以更容易且更一致地辨識是否一對詞義之間確實具有某種詞義關係。而且，每一個使用資料庫的人也可以據以檢驗其中關係連結的正確性。換句話說，對一個可檢驗且獨立於語言的詞彙語意學理論而言，這些測試提供了一個基石。

詞彙網路是以「同義詞集」(synset)與詞彙語意關係為核心所架構出之詞彙知識系統。也就是說，詞彙網路架構表達的不僅是詞彙本身的概念性知識，它亦表達了詞彙之間的語意關係。然而，從普林斯頓英語詞網以及歐語詞網(EuroWordNet)的建構經驗來看，這是一項相當費時耗力的龐大語言工程。對於經費取得困難、使用頻度較低之語言而言，建立此項語言資源更為不易。從詞彙語意與知識表達的角度觀察，我們認為不同語言對於概念原素(conceptual atoms)，可能有著不同之表達方式，但是在詞彙語意關係的表達上，則應具有更大程度之「普同性」。因此「借力」(bootstrapping)於已發展成熟之英語、歐語詞網之語意關係，以加速新的詞網雛形成形，就成了一個自然而然的另類選擇。

在 CWM 裡面，我們參考了普林斯頓 WordNet 的語意關係連結，例如：同義詞、反義詞、上位詞、下位詞…等等，另外，我們也開發了「類義詞(paranymy)」(Huang et al., 2008(a), 2007)的語意關係連結，主要是以 WordNet 為框架，在姊妹詞彙(sister terms)中解釋豐富的概念關係 (rich conceptual relations)。根據黃 等(2007)在之前的研究，以詞彙語意關係的角度來定義這些姊妹詞彙，我們就稱之為「類義詞」。增加「類義詞」的運用，可以讓我們在詞網的描述上更具完整性也更豐富性的 ontological 知識。一整個完整的「類義詞」語意關係會依照同一個概念原則來將他們歸類為同一類，我們深深地相信，我們將「類義詞」當作詞彙的語意關係，對於我們在處理詞彙語意關係上，有更明確的描述與對應，也可以藉此提供更明確的訊息以及富有 ontological 的詞網。

根據黃 等(2007)的分析，我們將「類義詞」分為兩大類：一、相對類義詞(contrary paranymy)；二、重疊類義詞(Overlapping Paranymy)。「相對類義詞」，通常是有比較級和最高級的;可以用 very, almost 等詞修飾的詞彙，這個語意關係的詞彙也可以是中等程度的詞彙，所以在描述上有可能既不是這個也不是那個。例如：某個東西可能被認定是「溫的」，因為它既不是熱的也不是冷的。除此之外，通常「相對類義詞」又可被分「類成認知、感官類」(perceptional paradigms)或「約定俗成類」(conventional paradigms)。「認知、感官類」是基於人類的認知、感知與感官，例如：快/慢的上層節點是速度。至於速度是快或是慢，這個就完全依靠個人的感知，然而這樣的感知是不同於其他人的。「約定俗成類」就如同在中文裡，我們對於雙親這個概念所使用的稱呼詞彙，我們可以圖七來表示這樣的關係。



圖八、雙親稱謂示意圖

至於「重疊類義詞」，簡單來說，則是這個類型的兩個姊妹詞彙共同擁有一些相同的特徵，在 CWN 裡面，我們基於語言約定俗成的用法來做解釋與區分這類重疊類義詞，通常與語言的使用與經驗相符合。例如：「箱子」和「盒子」，所隱含的語意與概念是相同的，但是，當我們要裝置較大的物品，如：電視機或電腦的時候，我們會使用「箱子」；當我們只是要放置較小的物品，如：杯子或餅乾，我們就會使用「盒子」。又例如：「江」與「河」，流量較大、規模較大的，我們稱之為「江」；反之，則稱之為「河」。儘管如此，「江」與「河」，所呈現的語意概念，在某些方面與程度，是相同的。

4.2 詞彙語意關係預測

在上述背景之下，我們提出之一個模型(Bootstrapping from Multilingual Wordnets)。此模型是基於中

文詞彙網路小組一系列之相關研究所得出(Huang et al. 2003, 2005, Hsieh et al 2006)。主要的論點在於，我們假定在詞彙語意關係之標記上，可以借力於其他已成形之詞網的跨語詞義關係資源。我們提出了平行進行詞義標記所涉及之邏輯條件,並以反向回饋驗證。在(Huang et al 2003)中，並曾針對 210 個中文詞形(lemma)做過小規模之試驗與評價。在接續之前的研究基礎(Hsieh et al 2006)，進一步在規模上與多語擴充兩個面向上作延伸試驗 。亦即，在規模上，我們將目前在中研院中文詞網小組所定義完成之八千多筆中文同義詞集為主；在多語擴充上，我們將歐語詞網(Vossen 1998)亦納入實驗對象。其中包括了德語、法語、捷克語、荷語、西語、義語與愛沙尼亞語等七種歐洲語言。

5 語言知識整合與應用

著眼於作為一個跨語言知識處理架構，中文詞彙網路的發展過程中，亦與歐洲語言(法語、義大利語)、日語以及兩岸中文之詞彙對應進行了語言知識整合與應用之嘗試。

5.1 跨語言知識系統的對比與應用

在跨語言知識資源平台設計上，中文詞彙網路小組與義大利國家計算語言學研究所，針對跨語詞彙知識資源之協作機制，共同提出了稱之為 LexFlow 之分散式計算架構，並以義大利文與中文為例成功地展示了初步之實驗成果(Bertagna, et al 2007)；在跨語詞彙知識表達之形式化上，我們亦與法國土魯斯大學合作，利用晚近之圖形處理與相似度計算技術，建構了中法動詞語意對應網路(Gaume, et al 2008)，同時亦透過心理語言學之實驗得到心理處理歷程之初步驗證。此外，透過 Wordnet 的訊息作為中介，我們也比對了中、日文之漢字知識表徵之相同與差異(Huang et al., 2008(b))。

為了解決全球多語化所帶來的問題，我們需要一個跨語言的知識資訊整合平台。我們需要一個知識資訊系統，其設計之核心主軸，在於產生內容可協作的(content interoperability)標準化製作、跨語言之分散性知識資源共享與交換機制，及其存取與檢索介面。在實作方法上，我們將以知識本體驅動的方式，利用上層知識本體與全球詞彙網路網格之串接作為知識資源核心,並輔以文本知識發掘與語意分析技術(請參見 Kyoto 計畫, Vossen et al. 2008)。實作出之知識資源，將以 wiki 平台之呈現，使其得以透過相關領域專家與使用者之貢獻回饋得到維護與永續性。最後，我們亦將會拓展此模式到不同語言與文化領域。此跨國合作計畫之最後目標，即在於設計與實作這樣的系統與資源，對於鉅量之分散全球環保知識，可以用一致的格式加以表達呈現，並進行深層之概念檢索與發掘。此項研究成果，特別是對於全球之中小型企業，包括非營利性組織，將有相當大之助益。

5.2 兩岸詞彙對應

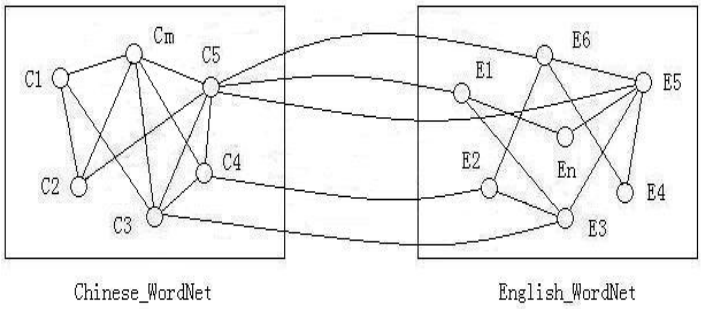
在我們的中文詞彙網路裡，我們結合了 Wordnet 的訊息，利用各種不同的語義關係，將每個原本屬於看似獨立的詞義連結起來；也以 Wordnet 為中界，比對同樣是中文而區分出來繁體中文系統(Chinese Wordnet, CWN)與簡體中文系統(Chinese Concept Dictionary, CCD)這兩個系統(Hong et al., 2007(b)；Hong et al., 2006)。

CCD，中文概念辭典(Chinese Concept Dictionary)，是一個中英雙語的詞網，北京大學計算語言學研究所開發，整個架構發展也是來自於 WordNet(于和俞, 2004; 于 等, 2003; 劉 等, 2003)。在 CCD 的發展手冊裡記載，研究團隊描述這些詞義的首要條件，是不可以破壞原本 WordNet 對於同義詞集定義概念與其語義關係的架構。另一方面，CCD 的研究團隊考量到可以存在許多在中文與英文的不同描述架構，所以，他們不止表現對中文詞彙內涵的表達，也發展了中文詞彙語義與概念的關係性，以利於強調中文的特質。

CCD 的研究團隊專注在整個 CCD 的架構，提出同一概念的同義詞集的定義，其所呈現的概念、定義和概念網的上下位語義關係，每一個同義詞集都有其基本關係，彼此之間亦有語義關係的存在。至於 CCD 的邏輯推演原則在語義網上的呈現，是運用到數學的形式而來的，是可以幫助研究者在中文語義分析上的使用。

自從 2000/09 開始，北京大學計算語言學研究所就已經開始著手以 WordNet 為基準，研究 CCD，並建立一個中英雙語的詞網，一個可以提供各種不同研究的詞網，如機器翻譯(MT)，訊息擷取(IE)…等等。

基於 WordNet 英文概念與 CCD 中文概念是屬於兩個不同知識背景，也因此 CCD 中，他們兩者間的相互關係與概念，是非常複雜、繁瑣的。CCD 包括了大量且繁雜的成對、成組的小網絡，大致上，差不多有 105 的概念節點和 106 的成組小網絡的概念關係，他們的關係，呈現如下圖：



圖九、WordNet 小網絡中複雜的關係結構

繁體中文系統的英中對譯(CWN)與簡體中文系統的英中對譯(CCD)，依不同詞類，區分成：名詞、動詞、形容詞和副詞四大類來進行對比，以 WordNet 為主，檢測在同一個 Synset 中，繁體中文系統的對譯詞彙和簡體中文系統的對譯詞彙，然後再進行比對。

在四大詞類中，我們可以清楚得知，在同一個 Synset 中，繁體中文系統，可能有多個相對應的對譯詞彙，同樣地，簡體中文系統也可能有個相對應的對譯詞彙。在這些對譯詞彙裡，又有可能是兩邊使用的對譯詞彙完全一樣，稱之「完全相同」；如果，兩邊使用的對譯詞彙，沒有一個相同的，稱之「完全不同」，也就是「真正不同」；或者，只有使用其中一個或一個以上對譯詞彙，這個狀況，稱之「部份相同」，而在「部份相同」的對譯詞彙，如果兩邊的對譯詞彙使用的詞首相同，稱之「詞首相同」，如果只是使用到相同的字，則稱之「部份字元相同」，如：

表一、CCD 和 CWN 對譯的各種分佈狀況

Synset	CCD 對譯詞彙	CWN 對譯詞彙	
bookshelf	書架、書櫃、書櫥	書架、書櫃、書櫥	完全相同
lay off	下崗	解雇	完全不同
immediately	立即	立刻	詞首相同
according	據報	根據	部分字元相同

6 結論

本文的研究，以中文詞彙網路建構為開端，以 WordNet 的資訊為橋樑，透過各種語義關係，串連起資訊豐富的多語語料，比對同樣是中文而區分出來繁體中文系統與簡體中文系統這兩個系統；也比對了中、日兩種不同語言在表達同一概念的相同與差異。當然，在因應全球多語化的驅動，我們也試圖開發出一個跨語言的整合知識訊息平台，並結合上層知識本體與全球詞彙網路網格之串接作為知識資源核心，進而發掘語義分析的技術。

如上述所提,我們的研究團隊—中央研究院語言學研究所中文詞彙網路研究小組從 2003 年以來大量的詞彙詞義分析研究成果為基礎，以人性化整合查詢介面透過網際網路呈現，除了提供相關研究人員以及有興趣的使用者查詢檢索外，希望藉此系統作為全球詞網—多語、跨語言的基礎知識架構對應的連結，更希望藉此系統作為進階中文詞彙知識研究之系統化實作參考與基礎，進而達到研究成果共享與學術交流之目的。

参 考 文 献

Francesca Bertagna, Monica Monachini, Claudia Soria, Nicoletta Calzolari, Chu-Ren Huang, Shu-Kai Hsieh, Andrea Marchetti, and Maurizio Tesconi. 2007. Fostering intercultural collaboration: a web service architecture for cross-fertilization of distributed wordnets.In: Ishida, T., Fussell, S.R., Vossen, P.T.J.M. eds.: *Intercultural Collaboration I. Lecture Notes in Computer Science*, Springer-Verlag .

Gaume, Bruno, Laurent Prévot, Chu-Ren Haung, Shu-Kai Hsieh and Chao-Jan Chen. 2008. Building and Aligning Chinese and French Paradigmatic Graphs. CIL18. Seoul: Korea.

Hong, Jia-Fei, Chu-Ren Huang, and Kathleen Ahrens. 2007 (a). Event Selection and Coercion of Two Verbs of Ingestion. Proceedings of Chinese Lexical Semantics Workshop 2007, Hong Kong Polytechnic University, May 20-23.

Hong Jia-Fei, Chu-Ren Huang and Ming-Wei Xu. 2007 (b). 以中文十億詞語料庫為基礎之兩岸詞彙對比研究. 第十九屆自然語言與語音處理研討會（ROCLING '07）台北：台灣大學. 2007.9.6-7.

Hong, Jia-Fei, Chu-Ren Huang, and Yang Liu,. 2006. WordNet Based Comparison of Language Variation : A study based on CCD and CWN. Proceedings of the Third International WordNet Conference. Pp. 61-68. Jeju. Januaray 22-25.

Hsieh, Shu-Kai, Simon Petr And Chu-Ren Huang. 2006. 大規模詞彙語意關係自動標記之初步研究：以中文詞網 (Chinese Wordnet) 為例. 中華民國計算語言學國際會議, 新竹：交通大學。

Huang, Chu-Ren, I-Li Su, Pei-Yi Hsiao, Xiu-Ling Ke. 2008(a). Paranymy: Enriching Ontological Knowledge in WordNets. To be presented at the 4th Global WordNet Conference. Szeged, Hungary. January 22-25.

Huang, Chu-Ren, Chiyo Hotani, Tzu-Yi Kuo, I-Li Su and Shu-Kai Hsieh 2008 (b). Wordnet-anchored Comparison of Chinese-Japanese kanji Word. The fourth Global WordNet Conference. Hungary: Szeged. January 22-25.

Huang, Chu-Ren, I-Li Su, Pei-Yi Hsiao, and Xiu-Ling Ke. 2007. "Paranyms, Co-Hyponyms and Antonyms: Representing Semantic Fields with Lexical Semantic Relations." Chinese Lexical Semantics Workshop. May 20-23. Hong Kong: Hong Kong Polytechnic University.

Huang, Chu-Ren, Chun-Ling Chen, Cui-Xia Weng, Hsiang-Ping Lee, Yong-Xiang Chen and Keh-jiann Chen. 2005. The Sinica Sense Management System: Design and Implementation. Computational Linguistics and Chinese Language Processing. 10.4.417-430.

Huang, Chu-Ren, Chang, Ru-Yng, Lee, Shiang-Bin. 2004. Sinica BOW (Bilingual Ontological Wordnet): Integration of bilingual wordnet and SUMO". The 4th International Conference on Language Resources and Evaluation (LREC2004). Lisbon. Portugal. 26-28 May, 2004.

Huang, Chu-Ren. Elanna I. J. Tseng, Dylan B. S. Tsai, Brian Murphy. 2003. Cross-lingual Portability of Semantic relations: Bootstrapping Chinese WordNet with English WordNet Relations. Language and Linguistics. 4.3.509-532.

Ker, Sue-Jin, Chu-Ren Huang, Jia-Fei Hong, Shi-Yin Liu, Hui-Ling Jian, I-Li Su and Shu-Kai Hsieh. 2008. Design and Prototype of a Large-scale and Fully Sense-tagged Corpus. To be presented at the Third International Conference on Large-scale Knowledge Resources (LKR2008). Tokyo, Tokyo Institute of Technology. March 3-5.

Ker, Shu-Jin, Chu-Ren Huang, Jia-Fei Hong, Shi-Yin Liu,Hui-Ling Jian and I-Li Su. 2007. 中文詞義全文標記語料庫之設計與雛形製作. The 19th ROCLING Conference. Taipei：National Taiwan University. 2007.9.6-7.

Niles, I., and Pease, A. "Toward a Standard Upper Ontology". In Proceedings of the 2nd International Conference on Formal Ontology in Information Systems (FOIS-2001). Chris Welty and Barry Smith, eds, Ogunquit, Maine, October 17-19, 2001.

Niles, I., and Pease, A. "Linking Lexicons and Ontologies: Mapping WordNet to the Suggested Upper Merged Ontology". In proceedings of the IEEE International Conference on Information and Knowledge Engineering. (IKE 2003). Las Vegas, Nevada, June 23-26, 2003.

Vossen, P., E. Agirre, N. Calzolari, C. Fellbaum, Shu-Kai Hsieh, Chu-Ren Huang, H. Isahara, K. Kanzaki, A. Marchetti, M. Monachini, F. Neri, R. Raffaelli, G. Rigau, M. Tesconi, J. VanGent. 2008. KYOTO: A System for Mining, Structuring, and Distributing Knowledge Across Languages and Cultures, To be presented at the 4th Global WordNet Conference. Szeged, Hungary. January 22-25.

于江生, 俞士汶。2004。中文概念詞典的結構。中文信息學報(Journal of Chinese Information Processing), vol. 16 No. 4 (2004)12-21。

于江生, 劉揚, 俞士汶。2003。中文概念詞典規格說明。Journal of Chinese language and Computing, 13(2) 177-194。

陳永祥, 洪嘉馡, 黃麗婉, 黃居仁. 2006. 網際網路中文詞彙知識檢索系統之建置. 第七屆漢語詞彙語義學研討會(CLSW-7). 交通大學. 2006.5.22-24.

蔡柏生，黃居仁，曾淑娟，林貞儀，陳克健，莊元珣. 2002. 中文詞義關係的定義與判定原則。中文信息學報(Journal of Chinese Information Processing). 16.4.21-31.

劉揚，俞士汶，于江生。2003。CCD語義知識庫的構造研究。2003中國計算機大會

Websites:

Chinese WordNet http://cwn.ling.sinica.edu.tw
Sinica BOW http://BOW.sinica.edu.tw
SUMO http://www.ontologyportal.org
Princeton WordNet http://wordnet.princeton.edu/