# 行政院國家科學委員會專題研究計畫成果報告

計畫編號: NSC 96-2411-H-003-061-MY3
執行期限：98年8月1日至99年7月31日

主持人：黃居仁　中央研究院語言所　（第一、二期）
churen@sinica.edu.tw
謝舒凱　國立臺灣師範大學英語所語言學組　（第三期）
shukai@gmail.com
計畫參與人員：陳克健 中央研究院資訊科學研究所
柯淑津 東吳大學資訊科學系(所)
ksj@csim.scu.edu.tw
周亞民 景文技術學院 資訊管理系
milesymchou@yahoo.com.tw
施昶安 中央研究院語言學研究所
ms9408@cis.scu.edu.tw
黃勝偉 中央研究院語言學研究所
rainbon@gate.sinica.edu.tw
羅翊寧 國立台灣師範大學英語學系
ilikeswatch@hotmail.com
吳岳錚 國立臺灣師範大學英語學系
wyc.juju@gmail.com

## 1　中英文摘要

### 1.1　計畫中文摘要

**中文關鍵字：知識本體，知識表徵，多語化，全球詞網，詞彙語意關係**

當今國際社會與學術發展皆爲多語與多元文化環境，這是語言學研究的挑戰，也是突破與轉型的關鍵。由語言能表達知識、傳達訊息的功能看來，表達知識、傳承文化的最主要媒介便是語言。以此來說，語言學研究的新派典必須能夠解釋語言如何承載豐富的知識與文化訊息。這個新派典最重要的基礎便是一能夠同時表徵語言現象與知識內容的描述架構，此表徵架構，容許跨語言變異，也能夠描述不同領域的不同知識系統。有了這樣的架構，多語多元文化的比較與融合研究才可行。此架構有兩個重要基石，一爲上層知識本體，二爲全球詞網。本計畫整合此二基石，以建構多語多元文化時代的語言研究基礎架構。

## 1.2 計畫英文摘要

**Keywords: Ontology, Knowledge Representation, Multilingualism, Global Wordnet, Lexical Semantic Relations**

Modern international society and research are under the environment of multilingualism and multiculturalisom. This environment is the greatest challenge and yet the turning point for linguistics studies. Based on the primary function of language to represent knowledge and to convey information, language is naturally the most important media for knowledge representation and cultural inheritance. Therefore, a new paradigm of linguistics studies must explain how language represent rich knowledge system as well as cultural identity. Central to this paradigm will be a shared framework to represent both linguistic facts and knowledge content. This shared framework accomodates linguistic variations and also variations among different knowledge systems. With this infrastructure, a synergy of research on multilingualism and multiculturalism is made possible. This project builds the infrastructure on two cornerstones – upper ontology and global wordnet.

# 2 計畫緣由及目的

## 2.1 理論背景與重要性

多語與多元文化（Multilingualism and Multiculturism）不但是當今國際社會與學術發展面臨的最大挑戰，也是語言學研究突破與轉型的關鍵。由結構主義到衍生語法，研究的派典有一個共同點，就是由單一語法（grammar）出發。這個單一語法，可能是對特定語言的的描述（如結構主義），也可能是主張可以描述多種語言的共同架構（如衍生理論中的通用語法universal grammar）。要

言之，是以同一套規律來嘗試解釋語言。 以一套規律來解釋，自有其操作上的簡潔性，也符合由數學出發的自然科學派典。 但是其解釋能力僅限於符號與符號系統。 換言之，並無法去有效表達預測這個符號系統所承載的大量複雜訊息。 由語言是表達知識，傳達訊息的功能來看，更在近三十年來語料庫語言學堅實的研究基礎上，我們得到的結論是： 語言是表達知識與傳承文化的最主要資源。 在這個結論的基礎上，我們主張語言學研究的新派典必須能夠解釋語言如何承載表達豐富的知識與文化訊息。 我們所主張的語言學研究的新派典的成功，最重要的基礎架構就是能夠同時表徵語言現象與知識內容的描述架構。 這個共同的表徵架構，容許跨語言的變異，也能描述不同領域的不同知識系統。 有了這個共同的表徵架構，多語多元文化的比較與融合研究才有可行的基礎架構。 我們主張這個架構的兩個重要基石是上層知識本體 (upper ontology)與全球詞網（global wordnet）。 本計劃整合此兩基石，以建構多語多文化時代的語言研究基礎架構。

## 2.2 應用背景與重要性

在知識經濟，全球化，與多語化的趨勢中，掌握並運用更多知識來源，成爲競爭力的關鍵。「人類語言科技」（Human Language Technology）的研究應運而生。 這個研究奠基在計算語言學與語言工程近40年的研究成績上，但更注重知識的處理以及知識處理的結果如何爲人所用。 在網路的未來發展上，則是以語意網 (Semantic Web)與網格運算(Grid Computing)最值得注目。 語意網嘗試把網路轉變成可以自動處理語意知識（而非僅是數位資料）的媒介。 語意網的最主要創新之一，是重視知識體系表達及知識本體（ontology）的研究。 而網格運算則打破了目前資料與運算集中處理的模式，使資料可以存在全世界不同的點上，也可以在不同的點同時進行運算。 不但打破了資源集中衍生的許多問題，更把運算的規模與複雜度提高到以往難以想像的程度。 語言學的研究在這個重要的關鍵點上，將發揮樞紐的功能。 人是最主要的知識的製造者與消費者。而人最熟悉的知識表達方式，當然是以語言。 人類多種語言的演化與歧異，造成了跨語言知識處理的複雜度，也有全球分布的特性； 正適合語意網與網格運算的處理。 從文化的觀點，數位典藏內容要跨越原表達語言與文化的鴻溝、典藏知識要能超越原有領域的限制，做跨領域的整合與加值。 這些都是面對全球化與多語化網路社會必須解決的問題。 知識經濟的重大瓶頸，其關鍵在是否能跨越語言與文字表達的限制，直接針對數位內容作知識交換與分享。「跨語言知識表徵基礎架構」將是跨越這個瓶頸的關鍵。

## 2.3 計畫內容–跨語言知識表徵基礎架構

此基礎架構可細分爲三個部分：

一、跨語言自然語言處理技術平台
二、跨語言語言知識表徵基礎架構
三、跨語言上層知識本體表徵基礎架構

### 2.3.1 跨語言自然語言處理技術平台

NLTK（Natural Language Toolkit）已成爲國際跨語言自然語言處理技術平台。但此平台上仍缺中文處理的基本資料與工具。 本計畫與創始與維護NLTK的國際同行合作（如：澳洲墨爾本大學，德國斯圖加特大學，與英國愛丁堡大學）， 取得技術上之支援，並用於NLTK 平台上整合發展中文處理技術。

### 2.3.2 跨語言語言知識表徵基礎架構

先期工作中，我們已利用由IEEE Standard Upper Ontology Working Group 所建立的SUMO (Suggested Upper Merged Ontology)，完成以概念出發之中、英詞彙資料檢索系統。 本計畫再繼續參與ISO相關國際標準制訂的大綱領下，同時積極參與亞洲，歐洲，美國等區域性整合計畫， 站穩相關技術領先群的地位。此平台與標準可分兩大項：

**一、詞彙知識表徵基礎架構** 大至整個語言，小到個別典藏的術語庫或領域詞彙庫，都是詞彙知識表達的一種形式。 這些詞彙知識若不能互通，典藏知識就不能交換。 我們參與 ISLE (International Standard of Language Engineering)，於2004年制訂MILES多語詞彙表達規範的基礎上（該規範現爲ISO TC37 SC4 WG4制訂相關標準草案的主要依據），與國際同行共同發展， 建立以語意網 OWL 語言爲表達語言，以Protege 爲程式語言的跨語言詞彙知識表徵基礎架構。

**二、跨語言詞網平台** 詞網（WordNet）已被公認爲語言知識本體（linguistic ontology）， 也就是說能由個別語言抽取概念關係，以及由文字表達對應到規範知識本體表達的最基本架構； 同時詞網也被認爲是最有跨語言互通性的語言及概念知識表達架構。 我們在「語言座標」計畫成績Sinica BOW及Chinewe WordNet（CWN）的基礎上，構建以網格（grid）架構的詞彙網路，共用及互享的基礎協作機制，並利用詞彙網路的基礎架構，建立跨語言知識交換的平台。 並參與Global WordNet Grid 國際合作計畫，對「多語的國際詞彙網路網格」之「中文節點」進行整體規劃。

策略上，我們彙整「語言座標」計畫已完成之詞形（約5000）、詞義數量（約10,000 — 12,000）及中文詞義關係基本資料庫， 對「國際詞彙網路網格

協作計畫」之「中文節點」進行整體進度規劃，並建構界面軟體工具。 以英語詞網網格節點 (4689 個共享基礎概念) 及其他語種詞網爲參考依據， 整理出中文詞網之共享基礎概念（Common Base Concepts），以英文詞網同義詞集及 SUMO 定義表示。 同時亦整理出未涵蓋在英語或歐語詞網中之中文特殊概念；最後成果以 XML 語言爲標誌標準，上傳至國際詞彙網路網格。 在此基礎上，比較國際上各種詞網之建構方法（例如：利用平行語料庫資源），進行國際合作與交流。 此研究項目亦可與其他研究項目作密切互動， 包括與漢字知識本體之關係研究，與全文語意標記及機器語意預測研究等等。

#### 2.3.3 跨語言上層知識本體表徵基礎架構

下一代網路知識內容表達，必須建立在知識本體(ontology)上， 而知識本體的互通性，則必須依賴上層共有知識本體。 這是爲什麼IEEE成立了上層知識本體工作小組（SUO），並提出SUMO（Suggested Upper Merged Ontology）。 語言座標計畫的重要核心成績，便是建立了中文本的SUMO及SUMO與中英文詞網的對應； 爲日後國家數位典藏適應語意網的永續經營，提供了必須的基礎架構。 我們在這個基礎上，與國際合作，繼續發展語言相關的知識本體技術與資料。

# 3 計畫成果自評

## 3.1 跨語言自然語言處理技術平台—PyCWN

Natural Language Toolkit（NLTK）已發展了WordNetCorpusReader， 可連至英文詞網及其他類似架構詞網，並抽取所需之詞義或語意關係等資訊，或計算語意相似度等。 然而NLTK中的WordNetCorpusReader因中英文詞網的架構設計不盡相同而不適用於中文詞網（以下列點解釋之），因此發展一特定適用於中文詞網的模組有其必要性。

**義面的區分**　　某個詞在語境中有明確單一詞義的前提下，若在不同語境中，有爲說話者接受的更細緻的區分，而此區分在概念上從屬於詞的單一詞義，中文詞網將此區分的細部語義歸爲「義面」（meaning facet）。義面的存在使得中文詞網的同義詞集架構較英文詞網更爲獨特，而不適用 NLTK 英文詞網的CorpusReader。

**類義詞的標記**　　中文詞網定義類義詞爲兩詞彙之間屬於同一語意分類的語意關係(黃居仁等人, 2007)，因此類義詞此語意關係是標記在詞義上，而非同義詞集上。

**異體詞的標記**　中文詞網區分同義詞與異體詞。目標詞有一與之發音相同、意義相同但詞形（written form）不同的相對應的字詞，則稱此對應字詞爲異體詞。在我們發展的模組中，異體詞不另闢語意關係類別，而是整合至目標詞之同義詞集內。

**同形異義詞**　同形異義詞指的是有著相同詞形但意義迥異的字詞。中文詞網將之視爲不同詞條（lemma）。譬如，「連」有三個詞條，連1、連2、連3。我們發展的模組中無創建Lemma類別（class），但類別的資訊仍保留於詞義代碼中（以中文詞網的詞義代碼爲準，不做更動）。

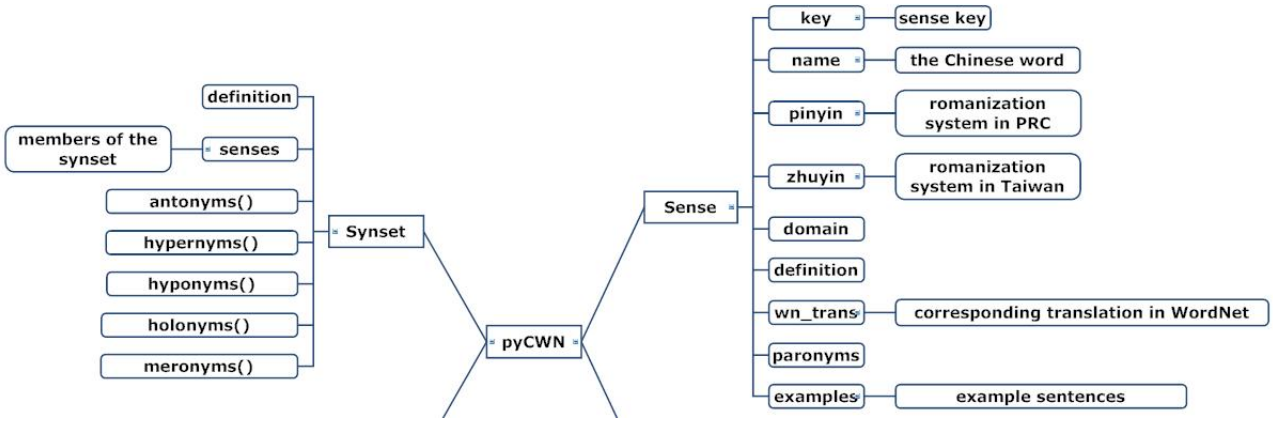


Figure 1: PyCWN之基礎架構

以現今情況而言，Python 已發展出 3.0 版本，使得中文處理更爲簡易流暢；然而 NLTK 許多套件仍無法與 python 3.0 相容，故我們發展模組以 python 2.x 版本爲主， 處理中文編碼，並使之連結到中文詞網，抽取特定所需資訊， 以便在python環境下進行更進一步的中文處理與應用。 此模組稱爲 PyCWN，模組結構詳見圖 1。 基於物件導向設計，此模組能更有效率地處理中文詞網的同義詞集及語意關係等語言資源。 以python語言寫成，此模組能在不同平台上使用，更可進一步整合入其他大規模的語言處理模組（如：Natural Language Processing Toolkits (NLTK)）。 它可依特定指令分別抽取中文詞網的不同訊息，包含標音(漢語拼音與國語注音)、詞類、釋義、例句、附註以及與目標詞有語意關係的詞彙。 中文詞網分析的語意關係含同義詞、反義詞、上位詞、下位詞、全體詞、部分詞、類義詞（paranomy）及事例詞（instance of）。而PyCWN中，可抽取同反義、上下位、部分全體、類義等與目標詞有語意關係的詞彙。 更重要的是，也可抽取目標詞對照至普林斯頓詞網（Princeton WordNet）的同義詞集及其編號（sense id，以WordNet 1.6爲主）。詳見圖 2。

連結到英文詞網，即是連結到歐洲詞網，此對跨語言知識整合助益甚大，更對未來詞網整合有莫大潛力。

```
>>> import cwn
>>> cwn.synsets('朝代')
['\xe6\x9c\x9d\xe4\xbb\xa3.n.0100']
>>> cwn.synsets('不加')
['\xe4\xb8\x8d\xe5\x8a\xa0.d.0100']
>>> print cwn.synsets('朝代')[0], cwn.synsets('不加')[0]
朝代.n.0100 不加.d.0100
>>> dynasty = cwn.Synset(cwn.synsets('朝代')[0])
>>> bu4jia1 = cwn.Synset(cwn.synsets('不加')[0])
>>> for member in dynasty.senses: # members of the synset
        print member


代1.n.0710
朝1.n.0510
>>> for member in bu4jia1.hypernyms():
        print member


不1.d.0110
勿.d.0200
無1.d.0310
>>> for member in dynasty.meronyms():
        print member


代1.n.0810
朝1.n.0610

>>> dy_sense = cwn.Sense(cwn.synsets('朝代')[0])
>>> print dy_sense.key, dy_sense.name, dy_sense.pinyin, dy_sense.zhuyin
04164201 朝代 chao2 dai4 ㄔㄠˊ ㄉㄞˋ
>>> dy_sense.wn_trans #WordNet translation
[u'dynasty.05976600N..']
>>> for eg in dy_sense.examples:
        print eg


任何一個文明或<朝代>，當時局有了變化，挑戰的力量即刻出現。
我國早在春秋戰國時代就有楚材晉用，而後每到國力發皇的<朝代>，常有延攬外國人擔任要職。
越是興盛強大的<朝代>，對人民性事的管束越是寬鬆；越是衰敗的<朝代>，對人民的控制越緊，性的禁錮也就越厲害。
```

Figure 2: PyCWN抽取中文詞網訊息範例

## 3.2 跨語言語言知識表徵基礎架構

### 3.2.1 詞彙知識表徵基礎架構—詞彙標示框架

詞彙標示框架是國際標準組織（International Organization for Standardization ISO/TC37 ）界定建立的一個標準化框架，目的在於應用於自然語言處理（Natural Language Processing）及機讀辭典（Machine-Readable Dictionary, MRD）的詞彙庫建立。 範疇涵蓋了對牽涉到多語溝通及文化多元的語言資源，對建立與交換語言資源的準則與方法做標準化處理。 詞彙標示框架的目標有三。 其一，爲詞彙資源的創造與使用提供共用模型。 其二，管理詞彙資源間的

資料交換。 其三，促進個別電子資源的整合以形成大規模的全球性電子資源。詞彙標示框架的種類包括單語、雙語或多語的詞彙資源。這三種分類亦適用於小型或大型詞彙庫、簡單或複雜詞彙庫，乃至於書面或口語詞彙表述。說明的範疇包含構詞學、語法學、計算語意學及電腦輔助翻譯。涵蓋的語言包括所有自然語言，並不侷限於歐洲地區。此計畫在自然語言處理的運用上不受限制。詞彙標示框架能呈現多數辭典，包括WordNet、EDR及PAROLE。 我們與國際並行發展，完成了

1 在ISO TC37 SC4 的架構下，與ISO TC37 SC4 參與會員國的正式代表共同制訂相關標準，並提供亞洲語言及台灣本土使用語言的觀點。

2 與義大利國家科學委員會計算語言學研究所（ILC-CNR, Italy）依照ISO標準平行發展工具平台。

3 在標準與共同平台與工具的基礎上，與東京工業大學、東京尖端科技研究所、泰國計算語言學研究所、中國富士通研究中心等研究機構，進行亞洲重要語言詞彙知識表達的實作，以調整並確認該標準及共享工具的可行性，及建立亞洲跨語言詞彙知識表達的核心架構。

### 3.2.2 跨語言詞網平台

我們以英語詞網網格節點（4689 個共享基礎概念）及其他語種詞網爲參考依據， 整理出了中文詞網之共享基礎概念（Common Base Concepts），以詞網同義詞集及 SUMO 定義表示之。 中文詞網更以此作爲釋義語言的基本標準，實現簡明精確之旨。中文詞網目前所分析之詞條（lemma）數已達 11407 條，詞義（sense）數高達 29169筆，義面（meaning facet）6485 筆。 中文詞網分析的語意關係含同義詞、反義詞、上位詞、下位詞、全體詞、部分詞、類義詞（paranomy）及事例詞（instance of）。 此對資訊檢索、資訊萃取及文本處理（例如：全文語意標記、機器語意預測）等自然語言處理應用有相當大的幫助。 再者，中文詞網以人工方式標記了英文詞網的對應詞，使中文詞網得以連結到英文詞網。 連結到英文詞網，即是連結到歐洲詞網（EuroNet）， 使中文詞網得以整合入WordNet grid，對未來全球語言知識整合的發展極有助益。另外，除了中研院的中文詞彙網路介面外，也發展了中文詞網視覺化網頁，如圖 3。

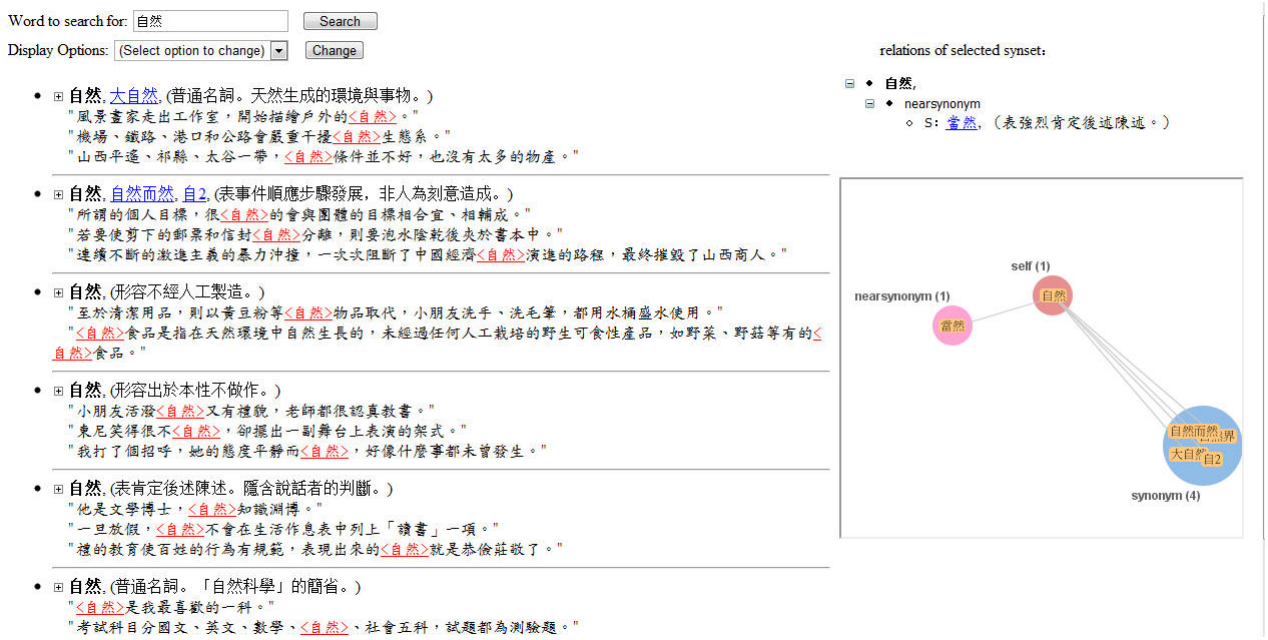


Figure 3: 中文詞網視覺化介面範例

## 3.3 跨語言上層知識本體表徵基礎架構

我們與義大利國家科學委員會計算語言學研究所（ILC-CNR, Italy）、義大利國家科學委員會應用知識本體實驗室（LAO-CNR, Italy）與IEEE SUMO Editor等機構，共同進行中文爲基礎之跨語言、跨領域詞義關係推導機制及概念推理(entailment)研究，作爲在知識本體的基礎上，整合並衍生典藏知識的實驗。我們以中央研究院中英雙語知識本體詞網（Academia Sinica Bilingual Ontological Wordnet, 簡稱BOW）及義大利詞網（ItalWordNet）爲例，觀察研究詞彙資源的半自動化整合及互通之所需條件與環境。做法上，我們首先連結各地當地的詞網，以網格的方式將各地詞網分別以XML 資料庫形式儲存於當地，此爲詞彙階段（lexicon level）；第二步爲應用階段（application level），使用wordnets grid發展出應用於豐富各詞網詞彙的多語詞網服務（Multilingual WordNet Service, MWS）；第三步合作階段（cooperation level）我們整合不同詞網，使各個詞網模組能整合於一大體環境下，形成全面性的工作流。見圖 4。

抽取語意關係方面，以部分詞爲例。我們先取義大利詞網中的一同義詞集「passaggio,strada,via」。此同義詞集對應到內部語言索引（interlingual index, ILI）是「road,route」，再對應到 BOW 同義詞集之「道路、道、路」（ 5-A）。中文之同義詞集「道路、道、路」與「彎」有一部分全體關係（ 5-B）。「彎」對應至ILI之「bend, crook, turn」再對應至義大利詞網之「curvatura, svolta, curva」（ 5-C）。因此顯示義大利詞網之「passaggio,strada,via」與「curvatura, svolta, curva」兩詞集可能有全體部分之語意關係。

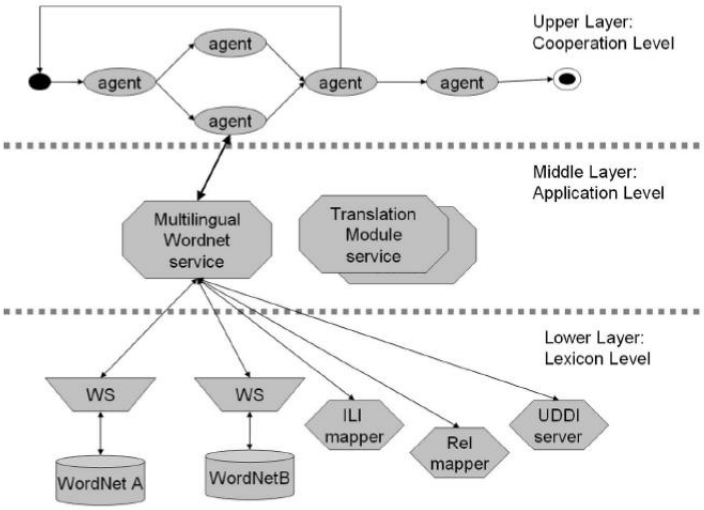


Figure 4: 三階段整合詞彙資源示意圖

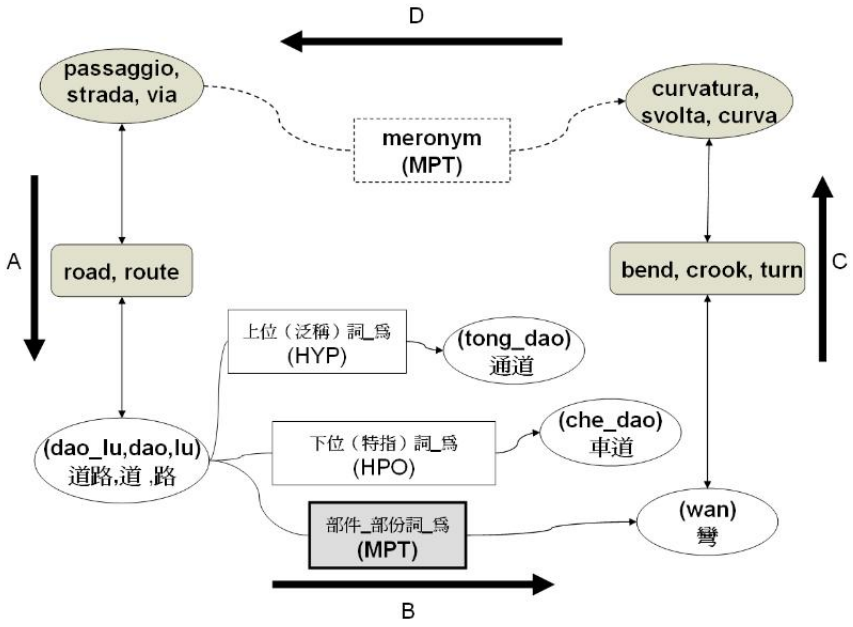


Figure 5: 利用ILI自動提議語意關係之示例圖

此法有幾項特點：一、可豐富化現有的詞彙資源。單一詞網的詞彙資源一般而言資訊不夠完整，藉由連結二詞網，使其互通，並自動抽取（boots）語意關係及其他資訊，可使詞網中資源更豐富。 二、可應用於創造新資源。藉由連結不同詞網的內部語言索引，各國語言詞彙可自動抽取出，創造新語言資源。三、可應用於檢驗現有詞彙資源。若目標詞網中同義詞集分類及語意關係之區

分能見於其他詞網，則此語言資訊則具有普遍性，也指明了詞網的適切性。 因此我們得以在跨語言框架中善用各語之詞彙資源，對未來助益甚大。

# 4 參考文獻


1. Calzolari, N. 2006. Technical and Strategic issues on Language Resources for a Research Infrastructure In S. Furui (eds.) *Proceedings of the International Symposium on Large-scale Knowledge Resources* (LKR2006). Tokyo Institute of Technology, 53-58.

2. Calzolari, N., Soria C. 2005. A New Paradigm for an Open Distributed Language Resource Infrastructure: the Case of Computational Lexicons. In Proceedings of the AAAI Spring Symposium "Knowledge Collection from Volunteer Contributors (KCVC05)", Stanford, CA: 110-114

3. Calzolari, N., Bertagna, F., Lenci, A., Monachini, M. (eds.). 2003. *Standards and Best Practice for Multilingual Computational Lexicons* MILE (the Multilingual ISLE Lexical Entry). ISLE CLWG Deliverable D2.2 & 3.2. Pisa.

4. Chang, Ru-Yng, Chu-Ren Huang, Feng-ju Lo, Sueming Chang. 2005. *From General Ontology to Specialized Ontology: A Study based on a Single Author Historical Corpus* Presented at the Fourth OntoLex Workshop. October 15. Jeju, Korea.

5. Chou, Ya-ming, Chu-Ren Huang. 2005. *Hantology: An Ontology based on Conventionalized onceptualization* Presented at the Fourth OntoLex Workshop. October 15. Jeju, Korea.

6. Church, Ken. W. and Hanks, Patrick. 1989. Word association norms, mutual information and lexicography. Proceedings of the 27th Annual Meeting of ACL. Pp. 76-83. Vancouver.

7. Daud, J., Padr, L. and Rigau, G. 2001. A Complete WN1.5 to WN1.6 Mapping. *In Proceedings of NAACL Workshop "WordNet and Other Lexical Resources: Applications, Extensions and Customizations"* Pittsburg, PA, United States, 2001.

8. Fellbaum, Christine. (ed.). 1998. WordNet: An Electronic Lexical Database. Cambridge, MA: MIT Press.

9. Francopoulo, G., G. Monte, N. Calzolari, M. Monachini, N. Bel, M. Pet, and C. Soria. (2006). Lexical markup framework (LMF). In Proceedings of LREC2006. Genova, Italy.

10. Guarino, N. and C. Welty. An overview of ontoclean. In S. Staab and R. Studer, editors, Handbook of ontologies, chapter An overview of ontoclean, pages 151—159. Springer Verlag, 2004.

11. Gulrajani, G. 2002. SHAWEL: Sharable and Interactive Web-Lexicons. Paper available at: http://emeld.org/workshop/2002/presentations/wittenburg/shawel-paper-final.doc
12. Hong, Jia-Fei and Chu-Ren Huang 2006. Using Chinese Gigaword Corpus and Chinese Word Sketch in linguistic research. To be presented at The 20th Pacific Asia Conference on Language, Information and Computation. November 1-3, Wuhan, China.
13. Hork, Ales and Pavel Smrz. 2004. VisDic – WordNet Browsing and Editing Tool. Proceedings of GWC 2004, pp. 136—141.
14. Hsieh, Shu-Kai, Shu-Ming Chang, Feng-Ju Lo, Ru-Ying Chang, Chun-Han Chang, Yi-Shuan Zhou, Chu-Ren Huang. 2006. GuangQunFangPu: e-Humanities Combining Textual and Botanic Information. To be presented at e-Humanities — an emerging discipline. Workshop in the 2nd IEEE International Conference on e-Science and Grid Computing. December 2006, Amsterdam, Holland.
15. Hsieh, Shu-Kai and Chu-Ren Huang. 2006. When Conset Meets Synset: A Preliminary Survey of an Ontological Lexical Resource based on Chinese Characters. Presented at the 2006 COLING/ACL Joint Conference. July 17-21. Sydney, Australia.
16. Huang, Chu-Ren, Nicoletta Calzolari, Aldo Gangemi, Alessandro Lenci, Alessandro Oltramari, and Laurent Prevot. Eds. Forthcoming (2007). Ontologies and the Lexicon. *Cambridge Studies in Natural Language Processing* Cambridge: Cambridge University Press.
17. Huang, Chu-Ren, Siaw-Fong Chung and Kathleen Ahrens. 2006. An Ontology-based Exploration of Knowledge Systems for Metaphor. In Rajiv Kishore, Ram Ramesh, and Raj Sharman Eds. Ontologies: A Handbook of Principles, Concepts and Applications in Information Systems. Berlin: Springer.
18. Huang, Chu Ren, Wei Yun Ma,Yi Ching Wu and Chih Ming Chiu. 2006. Knowledge-Rich Approach to Automatic Grammatical Information Acquisition: Enriching Chinese Sketch Engine with a Lexical Grammar. To be presented at The 20th Pacific Asia Conference on Language, Information and Computation. November 1-3, Wuhan, China.
19. Huang, Chu-Ren. 2006.大數與求眞：如何以十億字語料庫進行語言分析與研究Invited Speech, The 4th Annual Meeting of Society of Chinese Teachers in Taiwan, October 28. Kaohsiung, Taiwan
20. Huang, Chu-Ren. 2006. Towards Global Wordnet Grid: Infrastructure for language technology in the age of multilingualism 全球詞網網格倡議：多語社會中的語言科技基礎建設. Invited Speech, ROCLING2006. September 7-8. Hsinchu,

Taiwan
21. Huang, Chu-Ren. 2006. 語言學理論與分析在計算語言學中的應用. Keynote Speech 全國學生計 算語言學會議 (SWCL). August 23. Shenyang, China
22. Huang, Chu-Ren, Wan-Ying Lin, Jia-fei Hong, and I-Li Su. 2006. The Nature of Cross-lingual Lexical Semantic Relations: A Preliminary Study Based on English-Chinese Translation Equivalents. Proceedings of the Third International WordNet Conference. Pp. 180-189. Jeju. January 22-25. 23. Huang, Chu-Ren, Alessandro Lenci, and Alessandro Oltramari. Eds. 2005. Proceedings of OntoLex 2005. A post-IJCNLP 2005 workshop
24. Huang, Chu-Ren, Shiang-bin Li, and Jia-fei Hong. 2005. The Robustness of Domain Lexico-Taxonomy: Expanding Domain Lexicon with Cilin. Presented at the Fourth SigHan Workshop on Chinese Language Processing. October 14-15. Jeju, Korea
25. Huang, Chu-Ren, Feng-ju Lo, Ru-Yng Chang, and Sueming Chang. 2004. Reconstructing the Ontology of the Tang Dynasty: A pilot study of the Shakespeareangarden approach. Proceedings of the OntoLex 2004 Workshop. Lisbon. May 30, 2004
26. Huang, Chu-Ren, Chang, Ru-Yng, Lee, Shiang-Bin. 2004. Sinica BOW (Bilingual Ontological Wordnet): Integration of Bilingual WordNet and SUMO". Proceedings of the 4th International Conference on LanguageResources and Evaluation (LREC2004). Lisbon. Portugal. 26-28 May, 2004
27. Huang, Chu-Ren, Feng-ju Lo, Ru-Yng Chang, Sueming Chang. 2004. Sinica BOW and 300 Tang Poems: An overview of a bilingual ontological wordnet and its application to a small ontology of Tang poetry. Invited talk. Workshop on Possibilities of a Knowledgebase of Tang Civilization: Towards a new comprehensive digital archive of Tang China. Institute for Research in Humanities, Kyoto University. February 20-21
28. Ide, N., A. Lenci, N. Calzolari. RDF instanciations of ISLE/MILE lexical entries (2003) Proceedings of the ACL'03 workshop on Linguistic annotation: getting the model right
29. Kilgarriff, Adam, Chu-Ren Huang, Pavel Rychl, Simon Smith, and David Tugwell. 2005. Chinese Word Sketches. ASIALEX 2005: Words in Asian Cultural Context. Singapore
30. Kilgarriff, Adam, Pavel Rychl, Pavel Smrz and David Tugwell. 2004. The Sketch Engine. Proceedings of EURALEX, Lorient, France
31. Kilgarriff, Adam and Tugwell, David. Sketching Words. 2002. In Marie-Hlne Corrard (ed.): Lexicography and Natural Language Processing. A Festschrift in

Honour of B.T.S. Atkins. Euralex
32. Kishore, Rajiv, Ram Ramesh, and Raj Sharman Eds. Forthcoming (2006). Ontologies in the Context of Information Systems. Berlin: Springer
33. Levin, Beth. 1993. English Verb Classes and Alternations: A Preliminary Investigation. University of Chicago Press
34. Lin, Dekang. 1998. Automatic retrieval; and clustering of similar words. Proceedings of COLING-ACL. Montreal. 768-774
35. Ma, Wei-yun, and Chu-Ren Huang. 2006. Uniform and Effective Tagging of a Heterogeneous Giga-word Corpus. Presented at the 5th International Conference on Language Resources and Evaluation
36. Niles, I., and Pease, A., (2003). Linking Lexicons and Ontologies: Mapping WordNet to the Suggested Upper Merged Ontology, Proceedings of the IEEE International Conference on Information and Knowledge Engineering. (IKE 2003), Las Vegas, Nevada, June 23-26, 2003
37. Nirenburg, Sergei and Victor Raskin. 2004. Onotological Semantics. Cambridge: MIT Press
38. Oltramari, Alessandro, Chu-Ren Huang, and Alessandro Lenci. Eds. 2006. Proceedings of OntoLex 2006. A post-LREC 2006 workshop
39. Peters, W., Vossen, P., Diez-Orzas, P. and G. Adriaens. 1998. Cross-linguistic Alignment of Wordnets with an Inter-Lingual-Index. In: N. Ide, D. Greenstein, P. Vossen (eds), *Special Issue on EuroWordNet, Computers and the Humanities* Volume 32(2-3):221-251
40. Romary L., Francopoulo G., Monachini M., Salmon-Alt S. (in press). Lexical Markup Framework (LMF): working to reach a consensual ISO standard on lexicons. Accepted for publication in *Proceedings of LREC2006* Genoa, Italy
41. Roventini A., Alonge A., Bertagna F., Calzolari N., Girardi C., Magnini B., Marinelli R., Speranza M., Zampolli A., ItalWordNet: Building a Large Semantic Database for the Automatic Treatment of Italian. In Zampolli A., Calzolari N., Cignoni L. (eds.), Computational Linguistics in Pisa, *Special Issue of Linguistica Computazionale* Vol. XVIII-XIX, Istituto Editoriale e Poligrafico Internazionale, Pisa-Roma, 2003
42. Prevot, Laurent, Chu-Ren Huang, I-Li Su, 2006. Using the Swadesh list for creating a simple common taxonomy. To be presented at The 20th Pacific Asia Conference on Language, Information and Computation. November 1-3, Wuhan, China
43. Pustejovsky, James. 1995. The Generative Lexicon, The MIT Press
44. Ruimy N., Monachini M., Gola E., Calzolari N., Del Fiorentino M.C., Ulivieri

M., Rossi S. (2003). A Computational Semantic Lexicon of Italian: SIMPLE. In Zampolli A., Calzolari N., Cignoni L. (eds.), *Computational Linguistics in Pisa, Special Issue of Linguistica Computazionale* Vol. XVIII-XIX, Istituto Editoriale e Poligrafico Internazionale, Pisa-Roma, 2003

45. Schalley, Adrea C., and Dietmar Zaefferer. Eds. Forthcoming (2006). *Ontolinguistics. How Ontological Status Shapes the Linguistic Coding of Concepts* Berlin/New York: Mouton de Gruyter

46. Sinclair, John. M. (editor). 1987. Looking Up: an account of the COBUILD project in lexical computing. Collins

47. Soria, Claudia, Maurizio Tesconi, Andrea Marchetti, Francesca Bertagna, Monica Monachini, Chu-Ren Huang, and Nicoletta Calzolari. 2006. Towards Agent-based Cross-lingual Interoperability of Distributed Lexical Resources. Proceedings of the 2006 COLING/ACL post-conference workshop ‘Multilingual Language Resources and Interoperability.’ July 23. Sydney, Australia

48. Soria, C., Tesconi, M., Bertagna, F., Calzolari, N., Marchetti, A., and M. Monachini. 2006. Moving to Dynamic Computational Lexicons with LeXFlow. Accepted for publication in *Proceedings of LREC2006, Genova* Italy

49. Swadesh, Moriss. Lexicostatistical dating of prehistoric ethnic contacts: With special reference to north american indians and eskimos. In Proceedings of the American Philosophical Society, volume 96, pages 452—463, 1953

50. Tokunaga, Takenobu, Virach Sornlertlamvanich, Thatsanee Chareonporn, Nicoletta Calzolari, Monica Monachini, Claudia Soria, Chu-Ren Huang, YingJu Xia, Hao Yu, Laurent Prevot, and Kiyoaki Shirai. 2006. Infrastructure for standardization of Asian language resources. Presented at the 2006 COLING/ACL Joint Conference. July 17-21. Sydney, Australia

51. Tokunaga, Takenobu, and Chu-Ren Huang. 2007 (To Appear). Guest Editors. Special Issue on Asian Language Technology. Language Resrouces and Evaluation, Vol. 41.Nos. 2-3. (AHCI, SCI)

52. Vossen P. 1998. Introduction to EuroWordNet. In Ide N., Greenstein D., Vossen P. (eds), Special Issue on EuroWordNet. *Computers and the Humanities* Volume 32: (2-3) 1998. 73-89

53. Wittenburg, P., Kemps-Snijders, M. (2006). Some LIRICS Topics (available at lirics.loria.fr/doc_ pub/lirics-barca.ppt # 480,2,LMF Topics)

54. 周亞民,黃居仁. 2005. 漢字意符知識結構的建立.[Construction of a Knowledge Structure based Chinese Radicals.] Presented at the Sixth Chinese Lexical Semantics Workshop. Xiamen. April 21-24

55. 洪嘉馡,黃居仁, 巫宜靜. 2005. 異體字與異體詞詞彙語意初探.[Towards a

study on the Lexical Semantics of Character- and Word-Variants.]Presented at the Sixth Chinese Lexical Semantics Workshop. Xiamen. April 21-24
56. 張如瑩,黃居仁. 2004. 中央研究院中英雙語知識本體詞網(Sinica BOW)：結合詞網，知識本體， 與領域標記的詞彙知識庫。發表於第十六屆自然語言與語音處理研討會（ROCLING XVI） September 2-3. Greenbay, Taipei
57. 黃居仁. 2005. 漢字知識表達的幾個層面：字、義、與詞義關係 (Knowledge Representation with Hanzi: The relationship among characters, words, and senses). 漢字與全球化國際學術研討會 (International Conference on Chinese Characters and Globalization). January 28-30. Taipei
58. Francesca Bertagna, Shu-kai Hsieh, Andrea Marchetti. 2007 Exploring Interoperability of Language Resources: the Case of Cross-lingual Semi-automatic Enrichment of Wordnets