

## Contents

<b>1</b>	<b>詞彙語義管理系統</b>	<b>2</b>
1.1	SSMS設計準則	4
1.1.1	唯一序號	4
1.1.2	詞彙知識庫文集同時存取	6
1.1.3	連結Sinica BOW	6
1.1.4	交互參照裝置	6
1.2	SSMS系統建置	7
1.3	SSMS功能	8
<b>2</b>	<b>瀏覽系統與程式存取介面</b>	<b>13</b>
2.1	詞彙資料線上查詢	13
2.2	詞彙知識視覺化	14
2.3	中文詞彙地圖	14
2.4	PyCWN	14
<b>3</b>	<b>跨語言詞彙網路</b>	<b>14</b>
3.1	詞彙標示架構 (Lexical Markup Framework)	14
3.1.1	目標	14
3.1.2	CWN詞彙標示框架 (CWN-LMF)	15
3.2	全球詞彙網路網格與詞彙知識交換	23
3.2.1	Bootstrapping approach	23
3.2.2	跨語言上層知識本體表徵礎架構: 漢語義大利語	26
<b>4</b>	<b>其他應用</b>	<b>26</b>
4.1	釋義基本詞與知識本體	27
4.2	領域標記	27
4.3	詞義資料與詞義預測	27
4.3.1	詞義預測	27
4.4	詞彙網路模擬	29

# 詞彙庫系統設計與應用

## 1 詞彙語義管理系統

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。詞網的濫觴是普林斯頓大學建立的英語詞網（WordNet）。接著有歐語詞網（EuroWordNet）的建立後，不但受到普遍重視，更有其他許多其他語言詞網的的建立，然後詞網真正成為多語言與跨語言知識表達的最佳架構。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。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集，便成了最根本的工作。

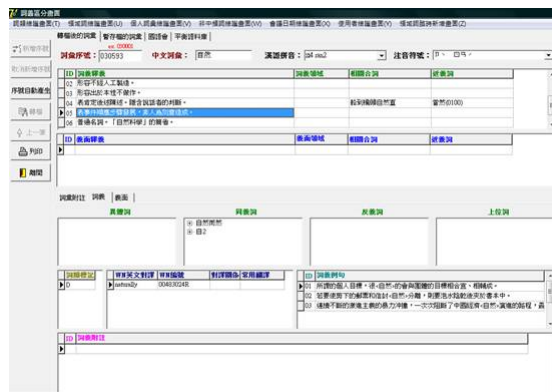


Figure 1: Sinica Sense Management System 介面

中研院詞彙語義管理系統（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，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以及中

文詞網資訊整併的來源。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。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，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。

詞彙資料庫綱要結構如下列各資料表所示：

資料表1：中文詞彙和詞類標記對照說明



Figure 2: 中文詞彙和詞類標記對照說明

資料表2：同義詞、反義詞、異體詞和上位詞

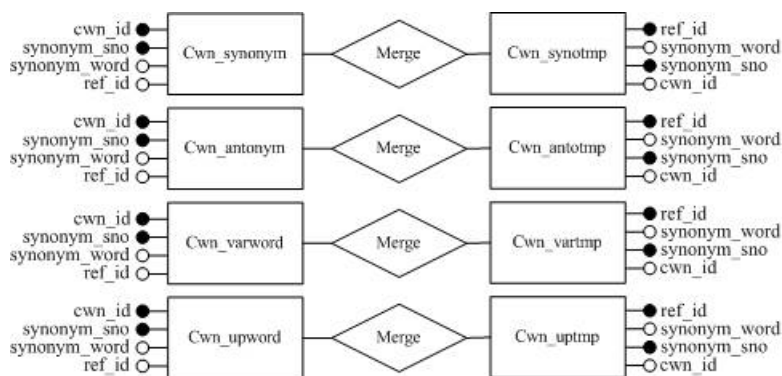


Figure 3: 同義詞、反義詞、異體詞和上位詞

資料表3：詞義和義面

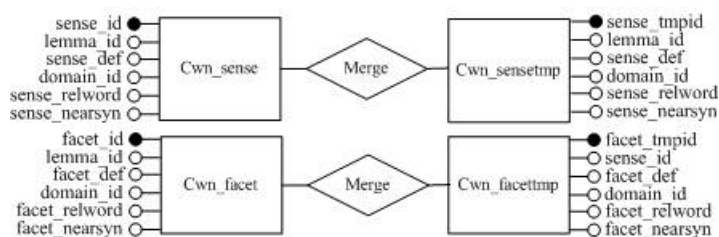


Figure 4: 詞義和義面

資料表4：例句、附註和詞類標記

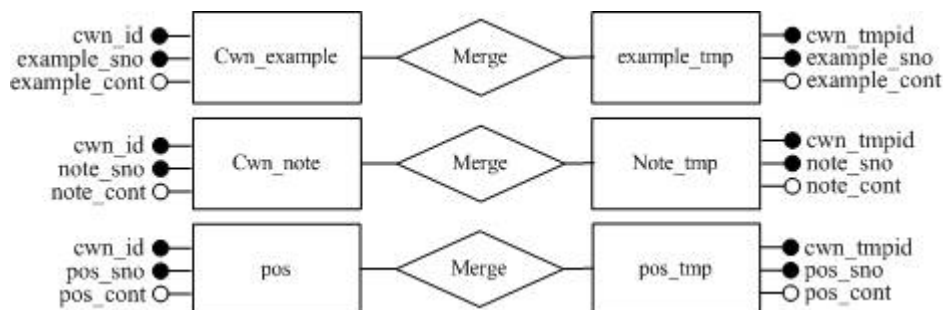


Figure 5: 例句、附註和詞類標記

資料表5：詞義與WordNet相關對應

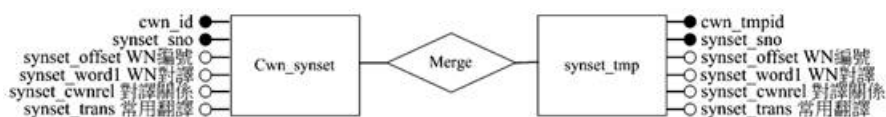


Figure 6: 詞義與WordNet相關對應

## 1.1 SSMS設計準則

### 1.1.1 唯一序號

SSMS中每一個詞義或詞義義面由唯一序號所識別，這跟Princeton詞網[Fellbaum 1998]對於每一個同義詞集賦予唯一號碼是相同的，然而，英文詞網雖然具有唯一序號但是卻不支持邏輯結構資訊，因此它不容易被追。其它作法方面，基於知識本體之下賦予語意詞義節點一個唯一序號，這是不方便於未事先指定所有可能概念和語識關係的情況下使用，另外，如果上層節點給予了編碼之後，因為下層節點是繼承自來源上層節點的關係，那下層節點將會發生隨機賦予序號的問題。

SSMS中的詞義序號可分成三個區段：1.時間序號，表示詞彙在什麼時候被新增。2.詞彙形式。3.詞彙詞義分類編碼(包含詞義義面階層)。以「bao4 zhi3(報紙)」為例，「報紙」有二個詞義以及第一個詞義有二個義面，詞彙詞目「報紙」如圖7所示。

中文詞目(詞義或義面)編碼方式如下：

詞目(lemma)：編碼佔用6字元

如：報紙 030018；前2碼為新增此詞目之年份，後面則為流水號。

詞義(sense)：編碼佔用2字元

如：報紙 03001801；前面6碼為此詞目之字元數，後2碼即代表詞義編號。

根據例子，表示此詞義是「報紙」的第一個涵義。

義面(meaning facet)：編碼佔用2字元

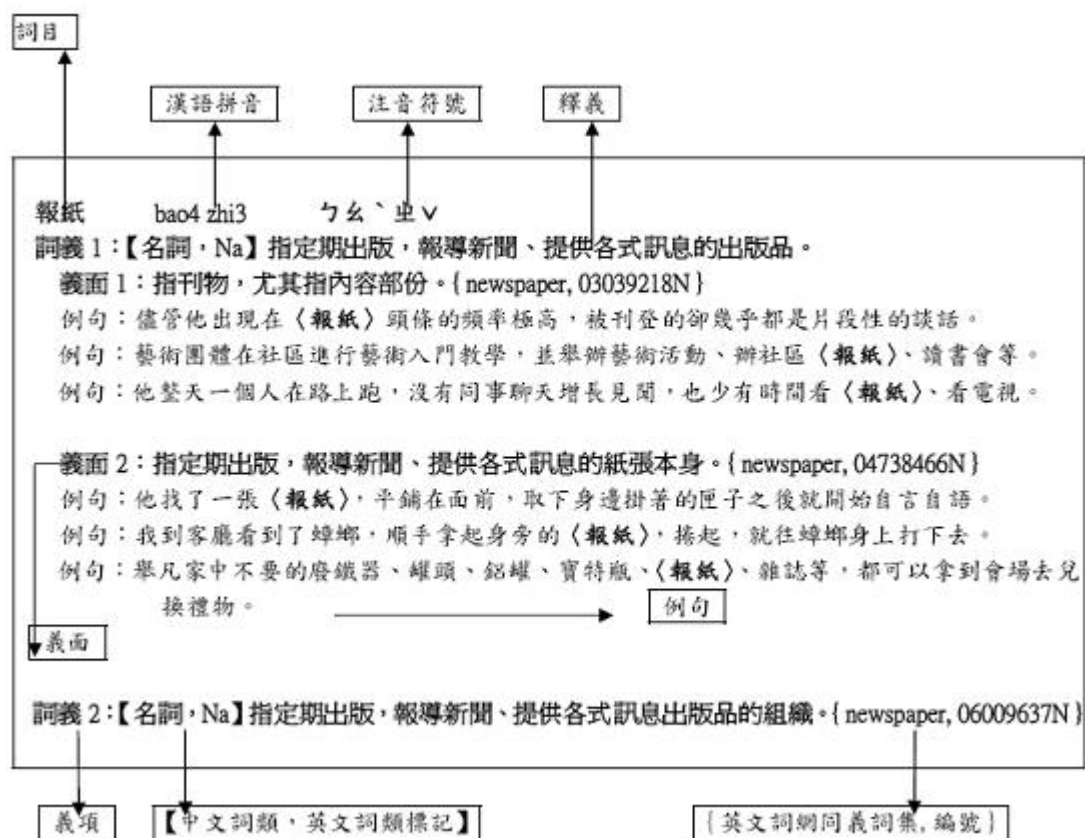


Figure 7: 詞義區分結果-以報紙為例

如：報紙0300180102；前面8碼為此詞目+詞義之字元數，後2碼即代表義面編號。根據例子，表示此義面是「報紙」詞義1在不同語境下所意義區分出的第二個義面。藉由唯一序號管理詞義(義面)有四個優點：

序號不僅提供唯一識別編號並且能夠讓專案管理者更易於追蹤工作流程。

將詞彙納入序號中，幫助使用者快速地辨識出相關的詞義。

詞彙代表著其詞義的一部份，為詞義序號提供了邏輯結構。

以四個位元代表該詞彙所屬的詞義和義面，前面2位元保留給詞義；後2都則給義面。這些字元允許用小量空間從資料庫中去辨識出確切詞義。例如，當指定一同義詞關係「word 0200」，意味著參照到指定詞彙的第二個詞義，不需要重覆顯示完整的詞義序號並且能夠唯一識別和支持追性。

	報紙 “bao4 zhi3 (newspaper)”
Lemma processing year	03-
Lemma form ID	-0018-
The first sense	-01-
The first meaning facet	-01

Figure 8: 詞義序號處理標示

### 1.1.2 詞彙知識庫文集同時存取

SSMS能夠在選擇查詢詞目的同時時間，取得詞彙知識庫文集中關於此條目的所有標記例子，讓使用者更進一步觀察詞義在例句中的使用和分布，而且，也可以自動選擇語料庫的句子段落範圍。當標記結果完成時，會將標記與原始句子文章合併在一起，並將查詢詞目平行地並列在上下列，這個特色能夠對詞目的每個詞義在真實文集中做追，也能幫助語言學家觀察提供資料下的每個詞義分類。例如，查詢一個詞目「you2 mu4 (游牧)」時，便會從平衡料庫中檢索出9個例子，如圖 9 所示。



Figure 9: 平衡語料庫詞彙詞類標記-以「游牧」為例

### 1.1.3 連結Sinica BOW

SSMS具有連結到中央研究院中英雙語知識本體詞網的候選英語同義詞集對應資訊，即查詢中文詞彙中結果畫面的英語同義詞集編號，這能夠幫助雙語間的追性和一致性。

### 1.1.4 交互參照裝置

SSMS能自動提示所有可能的交互參照。當欲分析一詞目時，所有存在著該詞目的所有記錄會被顯示出來，包括語意關係(如同義詞和下位詞)、詞義釋義、

附註，除了提供了豐富的語意關係資訊之外，也能對詞義關係清楚定義和偵測不一致性。釋義中任何表達格式的異常也能夠被偵測出來，也能夠改善詞義釋義的追性和一致性，這個處理確實地幫助我們僅需要專注在詞彙的詞義釋義上。例如，查詢詞目「you2 mu4 (游牧)」時，系統會自動提示相關參照「man3 (滿)」和「meng2 gu3 (蒙古)」，這二者都是參照到游牧民族生活區域。

## 1.2 SSMS系統建置

SSMS系統建置如圖 10

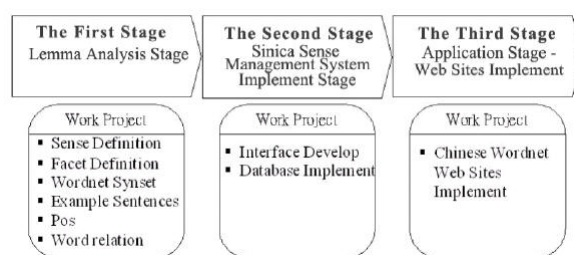


Figure 10: SSMS流程圖

分成三個階段。第一階段：詞彙分析階段，基於中文詞彙與意義操作原則[Huang2003]，對每個詞目區辨詞義與詞義義面，同時，平衡語料庫與英文詞網將被參照到詞類、例句以及英文翻譯，然後由系統所提供的字典資源或字詞對應來決定詞義關係。第二階段：包含了二個步驟，首先，我們必須設計SSMS資料庫網要結構，用來儲存詞彙分析結果，然後，為了資料庫管理的目的，我們開發了一個介面幫助詞網小組存取和修正資料庫；我們應用DELPHI程式開發工具設計介面，通過這個介面，能夠將資料庫轉換輸出成詞彙文件。最後階段：應用階段，我們計畫的主要目標是建立中文詞網讓使用者方便查詢，而網頁開發語言是HTML和ASP，使用者通過網際網路就能夠隨時隨地檢索詞彙資料庫的內容。要注意的是，當三個階段達到初始目標之後，第一階段的工作將持續擴展。

我們以圖11表達出SSMS整理框架，如圖中所顯示，中文詞網小組使用SSMS存取資料庫和電子文件做為詞彙報告。再者，使用者可以通過網際網路連結至網路伺服器，瀏覽HTML/ASP查詢頁面。

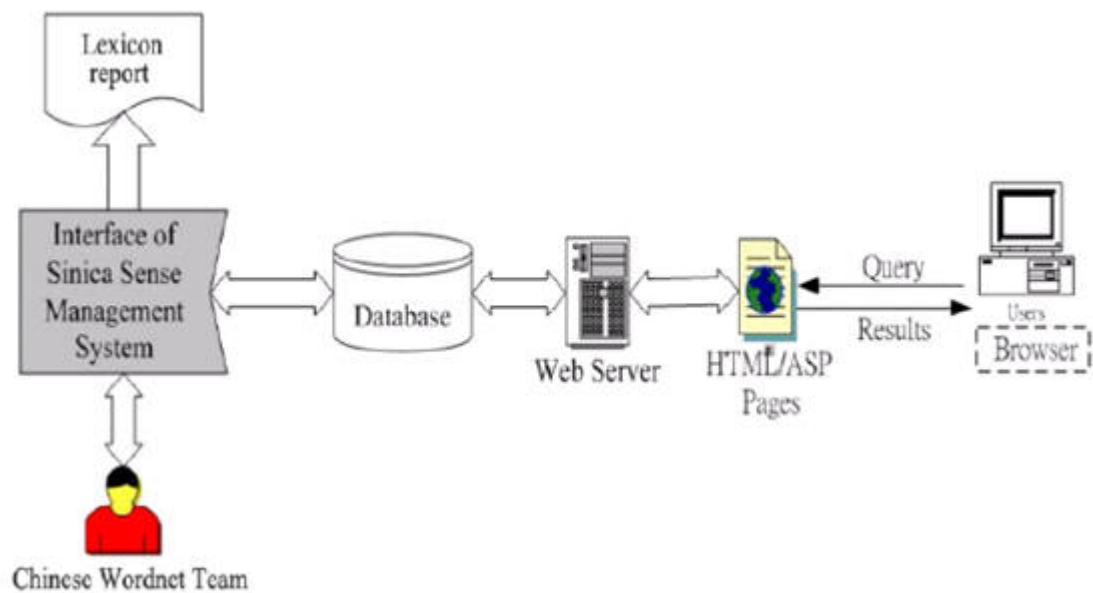


Figure 11: SSMS整體結構

### 1.3 SSMS功能

SSMS介面的開發語言為Delphi7.0，其程式執行架構如圖 12 所示。

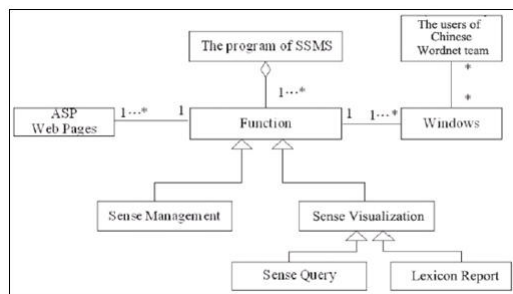


Figure 12: SSMS功能描述類別圖



SSMS包含了許多功能並且分別位在主視窗的各個子視窗以及交互連結的ASP網頁，圖 13 為SSMS主要系統畫面，提供了中文詞彙意義區辨小組分析與維護的親和式介面，當中詞義管理和詞義呈現是SSMS主要二個主要功能。

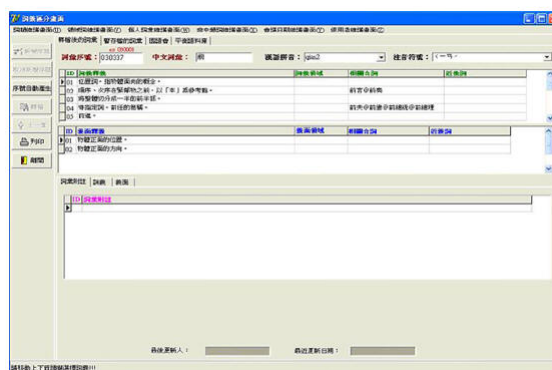


Figure 13: SSMS主要介面

在詞義管理上，中文詞網詞條可以被新增、修改並刪除，包含了詞彙詞目、詞義、義面、詞類、例句、英文同義詞集和詞彙語意關係等。

在詞義呈現上則可以分成二個部份：詞義查詢和詞彙報告輸出，畫面與輸出格式分別如圖 14、圖 15 所示，當語義查詢時，使用者可以輸入詞彙序號或者詞彙條目來得知相關資訊，單一字詞的多種資料顯示於工作視窗，包括了同義詞、反義詞、下位詞以及異體詞，透過這些清楚的展示使得詞彙語義關係能夠被瞭解和比較；詞彙報告輸出則使用Crystal Report9開發工具軟體產出電子文件。



Figure 14: SSMS詞義查詢畫面

中文詞義資料庫收錄的中文詞彙條目(entry)，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。

<p>拌 ban4 ㄅㄢˋ</p> <p>詞義 1:【及物動詞，VC】將任何材料混和、攪在一起。異體詞「伴」。(blend, 00274169V)</p> <p>義面 1: 將食材和食材，或食材和醬汁混和在一起。</p> <p>例句: 小女孩剛開始是給小黃吃鮮魚拌飯，小黃吃得津津有味。</p> <p>例句: 植物油及動物油的含量如何能攝取到最低的程度，通常拌沙拉以植物油處理。</p> <p>例句: 他把牛肉切薄片，並舀一點湯燙一下就上桌，而乾麵則只加一些醬油拌一拌，吃不到師傅的手藝。</p> <p>義面 2: 將一般液態和固態的材料混和在一起，使成為漿狀。</p> <p>例句: 他交給我一份以石灰水拌製海砂混凝土的研究計劃報告文件。</p> <p>例句: 牛原伯公口裡從來沒停止過嚼檳榔，雙手則忙著拌紅灰、包芒葉。</p> <p>附註:</p> <p>1.分義面的主要原因在於「拌」這個詞用在食物上的用法頻率上相當高，已經形成特殊用法，所以以義面方式處理，以標誌出這種情形。</p>
---

Figure 15: 詞彙報告輸出格式

## 4.2 中文詞彙網路(Chinese Wordnet)

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。



Figure 16: Chinese Wordnet 首頁

中文詞彙光碟版查詢介面，主要為了配合光碟版的輸出而產生並且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。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，透過視窗查詢介面可清楚呈現給使用者。中文詞彙網路(Chinese Wordnet)所能呈現的內容也均能和此介面與使用者互動產生結果完全相同。

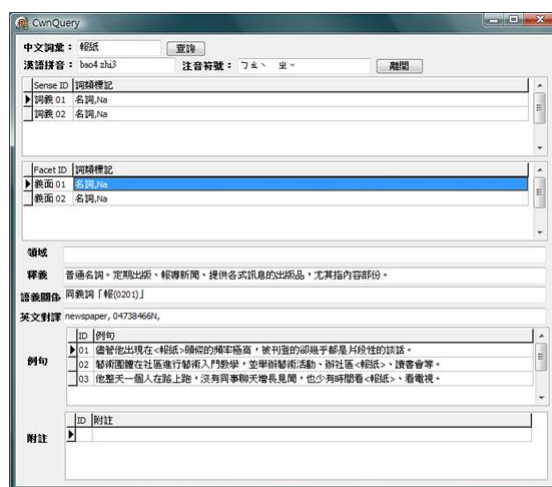


Figure 17: 中文詞彙光碟版查詢介面

詞彙資料庫綱要結構如下列各資料表所示：

資料表1：中文詞彙以及詞義和義面

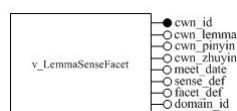


Figure 18: 中文詞彙以及詞義和義面

資料表2：例句、附註和詞類標記

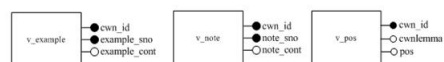


Figure 19: 例句、附註和詞類標記

欄位pos內容：依序為中文詞類，英文詞類標記。

資料表3：詞義語義關係

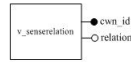


Figure 20: 詞義語義關係

資料表4：詞義與WordNet相關對應

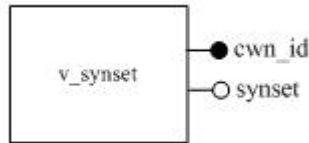


Figure 21: 詞義與WordNet相關對應

欄位synset內容：依序為WN英文對譯，WN編號，對譯關係，常用翻譯。

#### 4.3 中文同義詞集

雖然中研院詞彙語義管理系統(Sinica Sense Management System)儲存了中文詞網小組所收錄並分析的詞條、詞義等相關訊息，但是就同義詞集角度解釋之詞義關係仍然有所不足，原因在於此系統介面初始用意是針對詞條分析小組最簡單和直覺的使用方式所設計。因此，在詞義關係的連結上僅能顯示單一詞條間的詞義關係，而無法以同義詞集的概念去觀察詞集間的語意關係。最後，為了方便瞭解詞集間語意關係的狀態，我們依照圖22的步驟流程形成同義詞集。

經由上述步驟產生同義詞集之後，相關資訊如圖23所示。

其中相關欄位說明如下：

1、Synset 形式不同而意義相同或相近的詞集組合。兩兩詞彙間以逗號隔開而單一詞彙所屬sense-id則以「@」連接。 2、Wordcnt 同義詞集擁有的詞彙數量。 3、Offset 該同義詞集的唯一識別id，是以詞集領頭詞作為id來源。 4、Idex 資料表主鍵值。 5、Pntsymbol 包括目標同義詞集的指向數量、語義關係符號對應以及目標同義詞集offset。語義關係符號對應如表1所示，詞集語意關係如：反義詞、上位詞、下位詞等等計有八種關係。

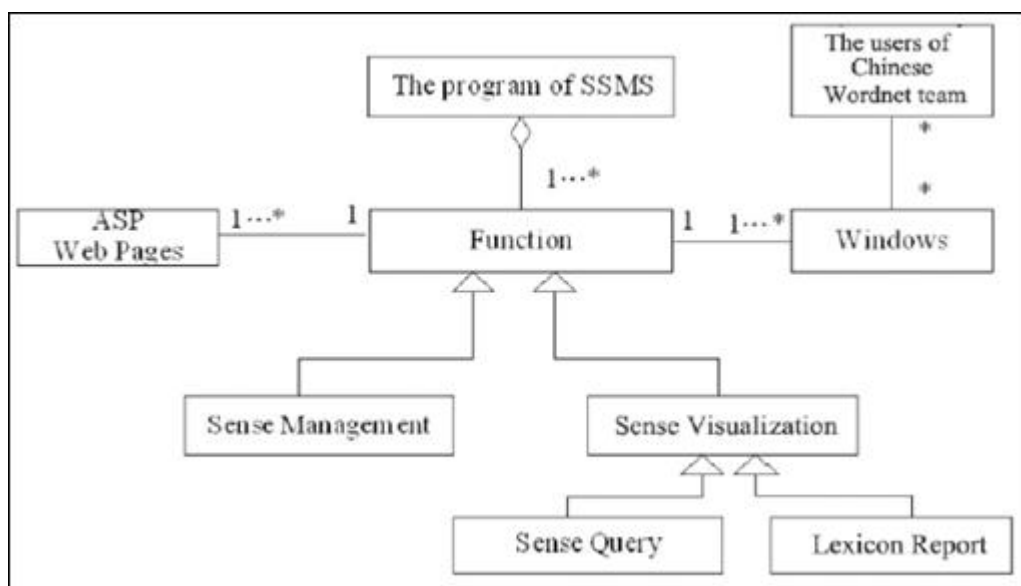


Figure 22: 同義詞集流程圖

Table 1: 語義關係符號對應

#	holonym
!	antonym
%	meronym
@	hypernym
~	hyponym
^	variant
&	nearsynonym
+	paronym

## 2 瀏覽系統與程式存取介面

### 2.1 詞彙資料線上查詢

(Chen et al)

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、

synset	wordcnt	idex	offset	pntrsymbol
03033704@前,05069001@前任,06551004@前面,	3	23462	03033704	1   03018204
04016205@代1,07042805@輩,08067009@紀2,	3	602	04016205	0
04016206@代1,07042806@輩,	2	603	04016206	0
0401630101@同,06025230@平1,06581201@相同,06586601@	4	604	0401630101	2   0668460101   0668460102
04016305@同,06586602@一樣,	2	605	04016305	0
04016501@轉寄,04016601@轉由,04016801@轉,06653103@借	4	606	04016501	0
04016701@幫,0411830101@助,	2	607	04016701	2 ~ 0512130101 ~ 0652060101
04016702@幫,04016901@替,	2	608	04016702	0
04016902@替,07104504@代2,	2	609	04016902	0
0401740101@用,0672240101@使用,	2	610	0401740101	1 ~ 05138201
0401740102@用,0672240102@使用,	2	611	0401740102	0
04017402@用,07053103@替1,	2	612	04017402	1 ~ 05138204
04017403@用,0506490102@花費,0516770102@費1,060314010	9	613	04017403	0
04017404@用,05227001@吃,06717016@進1,07029801@食1,	4	614	04017404	0

Figure 23: 中文同義詞集

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。

## 2.2 詞彙知識視覺化

(Hsu et al 2008)

我們遵循著廣被接受的設計模式，建立中文詞網的視覺化原型，可以從中觀察出詞義的分配以及特定同義詞集中有趣的和細微的資料。過去，資訊視覺化技術被應用在電腦科學或是生物科學，用它建構出大量資料間的關係，Ware (2000)提出了關於資訊視覺化效用的好處，如以下5項：

## 2.3 中文詞彙地圖

[Mashup] Hsieh et al. forthcoming

## 2.4 PyCWN

跨語言自然處理技術平台 (Julia and Hsieh 2010)

# 3 跨語言詞彙網路

## 3.1 詞彙標示架構 (Lexical Markup Framework)

詞彙標示框架 (Lexical Markup Framework，簡稱 LMF) 是國際標準組織 (ISO/TC37) 進行中的一項工作，目的在為自然語言處理與機讀字典的詞彙庫描述建立一個標準化框架。計畫範疇涵蓋對牽涉到多語溝通及文化差異的語言資源，對建立與交換這些資源的準則與方法做標準化處理。

### 3.1.1 目標

詞彙標示框架的目標有三。其一，為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。

### 3.1.2 CWN詞彙標示框架 (CWN-LMF)

LMF是基於LMF基本框架下，按照 WordNet詞彙知識表達模型所修訂出的詞彙資源編碼格式，而且，在遵循LMF框架和WordNet模型的同時，也試著將此效能最大化。以WordNet3.0的同義詞集「footprint\_1」為例子，其WordnetLMF如圖25所示。

```
<Synset id="eng-30-06645039-n" baseConcept="1">
  <Definition gloss="mark of a foot or shoe on a surface">
    <Statement example="the police made casts of the
    footprints in the soft earth outside the window"/>
  </Definition>
  <SynsetRelations>
    <SynsetRelation target="eng-30-06798750-n"
    relType="has_hyperonym">
    </SynsetRelation>
    <SynsetRelation target="eng-30-06645266-n"
    relType="has_hyponym">
    </SynsetRelation>
  </SynsetRelations>
  <MonolingualExternalRefs>
    <MonolingualExternalRef externalSystem="Wordnet1.6"
    externalReference="eng-16-01234567-n">
    <MonolingualExternalRef externalSystem="SUMO"
    externalReference="superficialPart" relType="at">
  </MonolingualExternalRefs>
</Synset>
```

Figure 24: LMF格式

中文詞網使用WordnetLMF表達詞彙語意，WordnetLMF的根元素是Lexical Resource，其下還有三個子元素，分別是GlobalInformation element、Lexicon elements (一個或多個) 和SenseAxes element (零個或一個)，LexicalResource即是一個容納許多詞彙資訊的容器，而語言間同義詞集的對應關係則包含在SenseAxes中，詳細說明如下：

### 1. GlobalInformation

用來描述詞彙資源的一般資訊，屬性「label」是個自由文字欄位(free textfield)。例如，<GlobalInformation label="Compile Chinese Wordnet entries using WordnetLMF">。

### 2. Lexicon

以多個 LexiconEntry 實例(instance)及其相關 Synset 元素組成單一語言資源。Lexicon 元素中包含許多屬性，依序有 languageCoding, language, owner, version, label 等 5 個，languageCoding 以 "ISO6393" 做為固定值；language 是 3 位元標準語言編碼(如，"zho"用來代表所指定的詞彙資源語言)，是必要的欄位；owner 是詞彙資源授權擁有者；version 為此資源的版本序號；label 則是用來記錄其它額外的資訊，是可選擇的。例如，<Lexicon languageCoding="ISO 6393" label="Chinese WordNet 1.6" language = "zho", owner = "Academia Sinica", version="1.6">。

#### • Lexicon Entry

Lexicon Entry有lemma和sense二個子元素，並且有可選擇的屬性id，做為唯一識別。

##### – Lemma

Lemma是按照詞形(word form)慣例，標示出的詞條。包括屬性有 partOfSpeech、writtenForm二個，partOfSpeech 根據 WordNet 方式，指定 POS 給每一個同義詞集，共有四個標記「"n", "v", "a", "r", "s"」使用在CWNLMF中，分別代表著名詞、動詞、形容詞、副詞以及其它POS標記。

##### – Sense

Sense元素記錄了詞條的意義和WordNet中屬於各自同義詞集的資訊。包括屬性有 id、synset 二個，id 根據中文詞網，指定 sense 的順序（例如，"環境\_1"表示詞目「環境」的第一個詞義）；synset表示此sense所在同義詞集的id。以詞目「環境」為例表示如下：

```
<LexicalEntry>
  <Lemma writtenForm="環境" partOfSpeech="n"></Lemma>
  <Sense id="環境_1" synset="zho1606640901n"></Sense>
</LexicalEntry>
```

#### • Synset

Synset元素有Definition以及可選擇的SynsetRelations和Monolingual ExternalRefs三個子元素，包括屬性有id, baseConcept兩個，id是唯一識別碼（例如，詞目「環境」的第一個詞義編碼為"zho-1606640901n"）；baseConcept屬性是一組數值名目(1,2,3)，基於NEDO 計畫 (Tokunaga et al. 2006)，歸類在第一類基本詞(basic words)者，編碼為1；歸類在第二類基本詞者，編碼為2；其它非基本詞者，則編碼為3。

##### – Definition

Definition元素允許同義詞集將釋義內容放在gloss屬性中；例句描述放在example屬性中。



- SynsetRelations  
SynsetRelations是一階層元素，集合SynsetRelation所有元素，SynsetRelation 記錄了同義詞集間詞義關係。包括屬性有 target、relType 二個，target 為關聯同義詞集的 id 值；relType 是關係種類，計有9種語意關係，分別是“has\_synonym”，“has\_nearsynonym” “has\_hyponym” “has\_hyponym” “has\_holonym” “has\_meronym” “has\_paronym” “has\_antonym” 和“has\_variant”，其中，語意關係類義詞（paronym）指的是二個詞彙句屬於同一語意分類(Huang et al.2008)下所形成的關係(例如，春/夏/秋/冬即是同屬於「seasons in a year」語意分類)。
- MonolingualExternalRefs  
MonolingualExternalRefs是一階層元素，集合MonolingualExternalRef所有元素，MonolingualExternalRef 用來表示詞義(同義詞集)和其它詞彙資源的連結，像是知識本體 (ontology)。包括屬性有externalSystem, externalReference, relType三個屬性，externalSystem描述其它系統資料名稱(例如，“domain”(Magnini and Cavaglia,2000), “SUMO” (Niles and Pease,2001) 以及 “Wordnet 3.0” 記錄sensekey值))；externalReference表示特定的識別或節點，是必要欄位；relType，如果“externalSystem”是“SUMO”的話，那麼“relType”即是SUMO知識本體節點的關係種類，可能的值有“at”，“plus”和“equal”。以詞目「環境」為例表示如下：

### 3. SenseAxes

SenseAxes是一階層元素，集合了不同語言詞網與WordNet 3.0間存在等價關係的對應。其中包括屬性，有id,relType二個，id是唯一識別碼；relType指定不同詞彙資料間，同義詞集的對應關係種類。例如，我們使用“eq\_synonym”表示中文詞網和 WordNet 3.0 間的同義詞關係，另外，當中文同義詞集 zho1606640901n 與英文同義詞集 eng3008567235n 存在同義詞關係對應時，則表示如下：

```
<SenseAxes>
  <SenseAxis id="sa_zho16eng30_5709" relType="eq_synonym">
    <Target ID="zho1606640901n"/>
    <Target ID="eng3008567235n"/>
  </SenseAxis>
</SenseAxes>
```

為配合中文詞彙日後授權事宜而採用與其它語言統一框架，將原有資料庫經過轉換過程並且輸出三個目的檔以作為xml格式(授權版本)資料取得之依據。下面為各個檔案中，逐一 欄位包含內容意義：

目的檔1：中文詞義清單。中文詞彙，詞目+詞類標記，版本+編號+詞類標記。

目的檔2：中文詞義內容資訊。版本+編號+詞類標記，BCS基本概念編號，釋義，例句，詞義所有語義關係，外部參照SUMO對譯，外部參照SUMO種

類。基本詞概念若該詞彙屬於基本詞，則該值為1；若為次基本詞，則該值為2；否則該值為3。

目的檔3：中文詞義與WN對應。版本+詞義編號+詞類標記，WN版本+WN編號+WN詞類標記。

另外，在LMF統一框架下所使用詞義編碼字元限制至多為8碼，因此也針對原有資料庫所使用至多為10碼重新調整以縮減成符合字元數。此調整動作是以詞義為基礎再對義面進行字元的縮減並根據流水號逐一編碼下去而達到要求。

以上圖詞目「報紙」為例，共有2個詞義並且在詞義01之下有二個義面。經過重新調整後，詞義01義面01將會取代詞義01並且詞義01義面02則會變成新編碼而成為03001803

ex. 090001	
詞彙序號: 030018	中文詞彙: 報紙 漢語拼音
<b>ID 詞義釋義</b>	
01	普通名詞。定期出版、報導新聞、提供各式訊息的出版品。
02	普通名詞。定期出版、報導新聞、提供各式訊息出版品的組織。
<b>ID 義面釋義</b>	
01	普通名詞。定期出版、報導新聞、提供各式訊息的出版品，尤其指內容部份。
02	普通名詞。定期出版、報導新聞、提供各式訊息的載體，通常為紙張。

Figure 25: 中文詞彙詞義區分以報紙為例

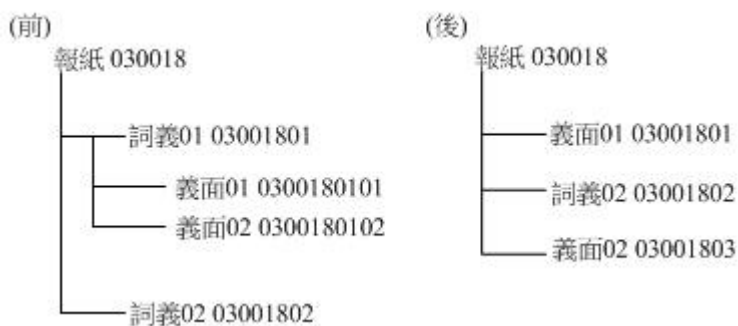


Figure 26: 詞彙編號編碼前後差異

LMF格式主要分為三個部分，分別是LexicalEntry、Synset、SenseAxes。僅以“自然”這個詞彙舉例說明(共125列,001:表示第一列開始)，其餘詞彙可類推。

- LexicalEntry(第006列到第029列)

”自然”這個lemma共有6個sense,分別是自然\_1、自然\_2、自然\_3、自然\_4、自然\_5和自然\_6。自然\_1(id 屬性)的書寫形式為自然(writtenForm 屬性)，詞性為名詞 n(partOfSpeech 屬性)。synset 編碼為 zho1603059301n，zho 是 ISO 6393 規範中文的代號，16 為 1.6 版，03059301 為這個 sense 在 CWN 中的編碼，最後 n 是這個 sense 的詞性。其餘 sense 同以上說明可類推。

- Synset (第030列到第097列)  
synset 為 zho1603059301n (自然\_\_1) 的釋義如 gloss 屬性，範例如example屬性。baseConcept 表示是否為基本詞，若該詞彙屬於基本詞列表則值為 1；屬於次基本詞列表則值為 2，否則其值皆為 3。SynsetRelations用來表示這個 synset 的語意關係，例如自然\_\_1有個同義詞(relType 屬性為 has\_synonym) 是大自然\_\_1(target 屬性為 zho1606653601n)，MonolingualExternalRefs 用來表示對外部系統的參照，例如自然\_\_1對應到 SUMO (externalSystem 屬性) 相當於 (ComplementFn) IntentionalProcess (externalReference屬性)，且屬於的類型是 plus (relType屬性)。其餘synset同以上說明可類推。
- SenseAxes (第099列到第124列)  
SenseAxes用來表示Chinese Wordnet 1.6版和Princeton Wordnet 3.0版的同義關係(relType屬性為eq\_synonym)。例如自然\_1(zho1603059301n) 相當於 nature (eng3011408559n)。其餘同以上可類推。

<範例說明：詞彙“自然”>

```
001: <?xml version = " 1.0" encoding = " UTF8" ?>
002: <!DOCTYPE LexicalResource SYSTEM " kyoto_wn.dtd" >
003: <LexicalResource>
004: <GlobalInformation label = " encoding of Chinese Wordnet entries using KyotoLMF by AS"
005: <Lexicon languageCoding = " ISO 6393" label = " Chinese Wordnet 1.6" language = " zho
006: <LexicalEntry>
007: <Lemma writtenForm = " 自然" partOfSpeech = " n" ></Lemma>
008: <Sense id = " 自然_1" synset = " zho1603059301-n" ></Sense>
009: </LexicalEntry>
010: <LexicalEntry>
011: <Lemma writtenForm = " 自然" partOfSpeech = " v" ></Lemma>
012: <Sense id = " 自然_2" synset = " zho1603059302v" ></Sense>
013: </LexicalEntry>
014: <LexicalEntry>
015: <Lemma writtenForm = " 自然" partOfSpeech = " v" ></Lemma>
016: <Sense id = " 自然_3" synset = " zho1603059303-v" ></Sense>
017: </LexicalEntry>
018: <LexicalEntry>
019: <Lemma writtenForm = " 自然" partOfSpeech = " r" ></Lemma>
020: <Sense id = " 自然_4" synset = " zho1603059304r" ></Sense>
021: </LexicalEntry>
022: <LexicalEntry>
023: <Lemma writtenForm = " 自然" partOfSpeech = " r" ></Lemma>
024: <Sense id = " 自然_5" synset = " zho16-03059305r" ></Sense>
025: </LexicalEntry>
026: <LexicalEntry>
027: <Lemma writtenForm = " 自然" partOfSpeech = " n" ></Lemma>
028: <Sense id = " 自然_6" synset = " zho1603059306n" ></Sense>
029: </LexicalEntry>
030: <Synset id = " zho1603059301n" baseConcept = " 3" >
031: <Definition gloss = " 普通名詞。天然生成的環境與事物。" >
032: <Statement example = " 風景畫家走出工作室，開始描繪戶外的自然。" />
033: </Definition>
034: <SynsetRelations>
035: <SynsetRelation target = " zho1606653601n" relType = " has_synonym" >
036: </SynsetRelation>
037: </SynsetRelations>
038: <MonolingualExternalRefs>
039: <MonolingualExternalRef externalSystem = " SUMO" externalReference = " (ComplementFn)
040: </MonolingualExternalRefs>
041: </Synset>
042: <Synset id = " zho-1603059302v" baseConcept = " 3" >
043: <Definition gloss = " 形容不經人工製造。" >
```

044: <Statement example = " 至於清潔用品，則以黃豆粉等自然物品取代，小朋友洗手、洗毛筆，都用水桶盛水使用。" />

045: </Definition>

046: <SynsetRelations>

047: </SynsetRelations>

048: <MonolingualExternalRefs>

049: <MonolingualExternalRef externalSystem = " SUMO" externalReference = " OrganicObject"

050: </MonolingualExternalRefs>

051: </Synset>

052: <Synset id = " zho1603059303v" baseConcept = " 3" >

053: <Definition gloss = " 形容出於本性不做作。" >

054: <Statement example = " 小朋友活潑自然又有禮貌，老師都很認真教書。" />

055: </Definition>

056: <SynsetRelations>

057: </SynsetRelations>

058: <MonolingualExternalRefs>

059: <MonolingualExternalRef externalSystem = " SUMO" externalReference = " (ComplementFn)

060: </MonolingualExternalRefs>

061: </Synset>

062: <Synset id = " zho1603059304r" baseConcept = " 3" >

063: <Definition gloss = " 表肯定後述陳述。隱含說話者的判斷。" >

064: <Statement example = " 他是文學博士，自然知識淵博。" />

065: </Definition>

066: <SynsetRelations>

067: <SynsetRelation target = " zho16-05196401r" relType = " has\_nearsynonym" >

068: </SynsetRelation>

069: </SynsetRelations>

070: <MonolingualExternalRefs>

071: <MonolingualExternalRef externalSystem = " SUMO" externalReference = " SubjectiveAsses

072: </MonolingualExternalRefs>

073: </Synset>

074: <Synset id = " zho1603059305r" baseConcept = " 3" >

075: <Definition gloss = " 表事件順應步驟發展，非人為刻意造成。" >

076: <Statement example = " 所謂的個人目標，很自然的會與團體的目標相合宜、相輔成。" />

077: </Definition>

078: <SynsetRelations>

079: <SynsetRelation target = " zho1605000204r" relType = " has\_synonym" >

080: </SynsetRelation>

081: <SynsetRelation target = " zho1604050801r" relType = " has\_synonym" >

082: </SynsetRelation>

083: </SynsetRelations>

084: <MonolingualExternalRefs>

085: <MonolingualExternalRef externalSystem = " SUMO" externalReference = " SubjectiveAsses

086: </MonolingualExternalRefs>

```

087: </Synset>
088: <Synset id = " zho1603059306n" baseConcept = " 3" >
089: <Definition gloss = " 普通名詞。「自然科學」的簡省。" >
090: <Statement example = " 自然是我最喜歡的一科。" />
091: </Definition>
092: <SynsetRelations>
093: </SynsetRelations>
094: <MonolingualExternalRefs>
095: <MonolingualExternalRef externalSystem = " SUMO" externalReference = " FieldOfStudy"
096: </MonolingualExternalRefs>
097: </Synset>
098: </Lexicon>
099: <SenseAxes>
100: <SenseAxis id = " sa_zho16eng30_17684" relType = " eq_synonym" >
101: <Target ID = " zho1603059301n" />
102: <Target ID = " eng3011408559n" />
103: </SenseAxis>
104: <SenseAxis id = " sa_zho16eng30_17685" relType = " eq_synonym" >
105: <Target ID = " zho1603059302v" />
106: <Target ID = " eng3001679907a" />
107: </SenseAxis>
108: <SenseAxis id = " sa_zho16eng30_17686" relType = " eq_synonym" >
109: <Target ID = " zho1603059303v" />
110: <Target ID = " eng3001799035a" />
111: </SenseAxis>
112: <SenseAxis id = " sa_zho16-eng30_17687" relType = " eq_synonym" >
113: <Target ID = " zho16-03059304r" />
114: <Target ID = " eng3000038625r" />
115: </SenseAxis>
116: <SenseAxis id = " sa_zho16eng30_17688" relType = " eq_synonym" >
117: <Target ID = " zho1603059305r" />
118: <Target ID = " eng3000488773-r" />
119: </SenseAxis>
120: <SenseAxis id = " sa_zho16eng30_17689" relType = " eq_synonym" >
121: <Target ID = " zho1603059306n" />
122: <Target ID = " eng3006000400n" />
123: </SenseAxis>
124: </SenseAxes>
125: </LexicalResource>

```

## 3.2 全球詞彙網路網格與詞彙知識交換

[Julia, NSC report]

### 3.2.1 Bootstrapping approach

中文詞網提供了精確的詞義劃分和詞彙語意關係，也對部分詞義使用Domain LexicoTaxonomy (DLT, Huang et al. , 2004b) 標記上領域標籤，然而，英文詞網則有Princeton WordNet 2.0同義詞集以及相對應的領域，也通過Dewey Decimal Classification (DDC)對每個同義詞集半自動式的標記上至少一個領域標籤。

由於，雙語詞彙語意關係推論曾經將bootstrapping方法實驗在SinicaBow和Princeton WordNet (Huang et al. 2002; 2003b)，因此，我們也用它來建構中文詞網領域(Chinese WordNet Domains)，並且藉由目前英文詞網領域做為媒介，自動地將中文詞網詞義標記上領域標籤。此方法分成三個方面進行，分別是普林斯頓英文詞網比對(alignment)、詞彙語意關係以及領域分類對應，詳細說明如下所示：

#### 1. Alignmentmediated領域預測

當中文詞義能夠完全地對應到英文詞網時，可以使用Alignmentmediated方式bootstrap出中文詞網領域，在圖27中， $CW_1$  表示可以對應到英文詞義 $EW_1$ 的中文詞義；DDC (Dewey Decimal Classification) 表示英文詞網所選擇的領域標記方法。如果 $EW_1$ 經由DDC標記了至少一個語意領域，並且 $EW_1$ 對應到 $CW_1$ ，那麼 $CW_1$ 則可以被預測出具有和 $EW_1$ 相同的領域標籤。例如，詞目「愛迪生」的第一個詞義“04071401n”，對應到英文詞網2.0同義詞集“10235982n”，同時，同義詞集“10235982n”關聯到領域「person」，因此，基於Alignmentmediated領域預測，我們將詞義「愛迪生」“04071401n”領域標籤標記為「person」。

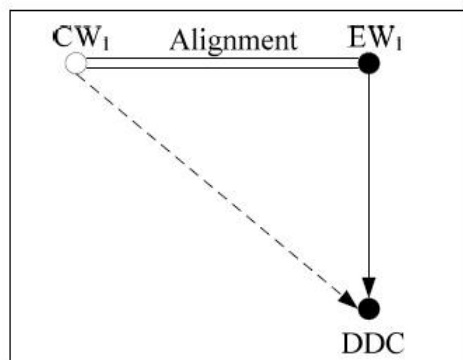


Figure 27: Alignmentmediated領域預測

#### 2. LSRmediated領域預測

中文詞網中，對每個詞義人工標記了許多的語意關係(同義詞、上位詞、下位詞以及反義詞等等)，如果一個詞義已經有alignmentmediated預測的領域標籤，我們就可以對這些定義好的詞彙語意關係(LSR)的詞義連結，進行下一步領域標記。在圖28中，CW<sub>1</sub>表示alignmentmediated中，經過DDC預測出領域的中文詞義；CW<sub>2</sub>和CW<sub>1</sub>存在LSR關係。通過LSR，我們可以預測CW<sub>2</sub>與CW<sub>1</sub>有相同的領域標籤。我們使用四種LSR來推論領域標籤，分別是同義詞、近義詞、類義詞以及異體詞，其中，類義詞是依照二個詞項(lexical item)之間屬於相同語意做為分類(Huang et al. 2008a)，例如，春/夏/秋/冬在“seasons in a year”主要概念下，屬於類義詞關係。再者，異體詞關係中文字的變形，例如“為什麼”和“為甚麼”二者都是“why”的意思，但是在第二個字元使用不同的書寫形式。另外，以為例，第二個詞義“05085502 n”(水瓶座 2)由alignmentmediated標記為“astrology”，並且和詞目「摩羯座」第二個詞義“05181002n”(摩羯座 2)二者之間有類義詞關係，所以基於LSRmediated領域預測，決定“astrology”為“05181002n”的領域標記。

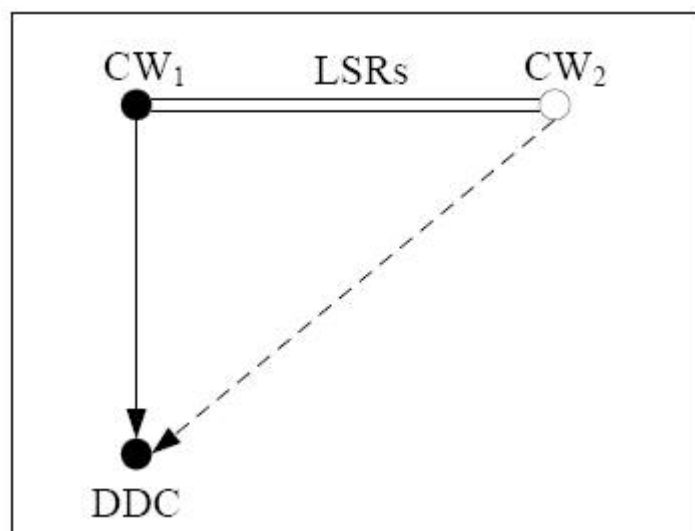


Figure 28: LSRmediated領域預測

### 3. Mappingmediated 領域預測

迄今，中文詞網有些詞義已經通過DomainLexicoTaxonomy (DLT) (Huang et al. 2004b)標記出領域節點，如果DLT領域節點和DDC語意領域之間的對應能夠明確界定的話，就可以利用英文詞網領域標記出中文詞義的語意領域。在圖29中，CW<sub>1</sub>表示有DLT領域標記的中文詞義，如果DLT和DDC二者之間的對應能夠被建立的話，那麼CW<sub>1</sub>便能夠通過DLTmapping預測出DDC領域標籤。例如，詞目「佛」第一個詞義“06736101n”的 DLT領域節點是“佛教”，經過DLT和DDC的對應之



後，“佛教”對應到“religion”，所以“06736101 n”可以被英文 詞網領域標記為“religion”。

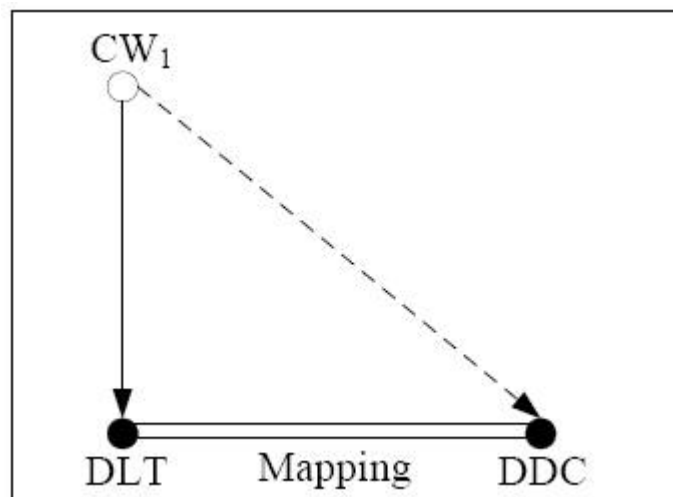


Figure 29: Mapping-mediated domain prediction

我們使用中文詞網 1.6 和英文詞網領域 3.2 做為 bootstrapping，以便取得中文詞網領域。在中文詞網中，已經分析的詞目和詞義分別是 8628 與 25938 個數量，而詞義間包含語意關係中，有18789同義詞關係、1801近義詞關係、3029類義詞關係以及923異體詞關係存在。另外，中文詞網的2541個詞義已經標記上DLT 領域標籤，而英文詞網領域3.2(45個基本領域、168個領域)也包含了與英文詞網2.0同義詞集的領域對應，共有115424組同義詞集標記上DDC領域標籤，其中，40 995組同義詞集領域標記為“Factotum”，表示不屬於任何一個語意領域。

然後，我們對每一個詞義進行bootstrap，得到每一個領域下究竟有哪些詞義涵蓋在裡面，而為了瞭解這個方法結果的效能，我們使用了在多標籤分類課題上知名的三個測 measure做為檢測工具，依序定義如下：

方程式中， $\| \text{predicated\_labels} \|$  表示詞義領域預測的標記數量； $\| \text{correct\_labels} \|$  表示詞義領域標記正確的數量； $\| \text{correct\_labels} \cap \text{predicated\_labels} \|$  表示詞義領域預測的正確標記數量。例如，如果已知正確的詞義領域標籤是A, B, C, D和bootstrapping預測的詞義領域標籤是B, C, E，那麼 $\| \text{predicated\_labels} \|$  就是3 (i.e. B, C, E)； $\| \text{correct\_labels} \|$  就是4 (i.e. A, B, C, D)； $\| \text{correct\_labels} \cap \text{predicated\_labels} \|$  就是2 (i.e. B, C)，precision和recall則為0.67和0.5。precision值為0.67表示，當預測領域標籤為3時，至少有二個標記正確；recall值為0.5則表示使用bootstrapping能夠預測出一半正確的領域標籤。

最後，我們也對bootstrapping方法做了效能評估，而為了能讓未預測到任何領域的詞義也列入評估，我們將它們人工標記為“Factotum”，評估的結果如表 mediated達到最佳結果(precision = 97.15%， recall =

$$\text{Precision} = \frac{\|correct\_labels \cap predicated\_labels\|}{\|predicated\_labels\|} \quad (1)$$

$$\text{Recall} = \frac{\|correct\_labels \cap predicated\_labels\|}{\|correct\_labels\|} \quad (2)$$

$$F - measure = \frac{2 * \text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}} \quad (3)$$

Figure 30: mediated領域預測

Methods	#Sense	Precision (%)	Recall (%)	F measure (%)
Alignment-mediated	6,669	83.86	82.72	82.46
LSR-mediated	174	70.4	69.83	69.54
Mapping-mediated	2,350	97.15	96.33	96.6

Figure 31: bootstrapping效能評估

96.33%，F measure mediated為基於領域預測的boots mediated方式來的好。由於人工建立語言資源會造成許多不一致和錯誤，所以我們也希望藉由自動地boot strap語意領域標籤，得到較為滿意的結果。

### 3.2.2 跨語言上層知識本體表徵礎架構：漢語義大利語

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，共同進行中文為基礎之跨語言、跨領域詞義關係推導機制及概念推理 (entailment)研究，作為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）為例，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。

## 4 其他應用

For a detailed description of specific possibilities for linguists, please refer to the LinguistLyX page on the LyX wiki key "linguistlyx" (feel free to enter your own hints there).

## 4.1 釋義基本詞與知識本體

在進行詞義區辨時我們需要使用一些釋義文字來定義所分析的詞彙。第二章的釋義準則中已經說明釋義文字以使用基本詞為原則，目的在於希望能以最基本的概念與淺顯易懂的詞彙將觀念明確的表達。通常基本詞具備以下幾項特點：

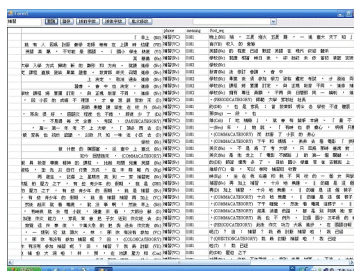
## 4.2 領域標記

[略] 同 3.2.1 Bootstrapping approach

## 4.3 詞義資料與詞義預測

### 4.3.1 詞義預測

詞義標記系統裡面，我們分為兩大類：一、人工標記；二、自動預測。在人工標記方面，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。初期中文詞網小組以人工的方式進行詞義標記，以「中央研究院現代漢語平衡語料庫」作為語料標的，以詞條為節點，在該詞條的每個詞義下標記出20個句子。用於標記的詞義代碼為四位數整數，前兩位為詞義序號，表明標示詞義出現在字典中之詞義順序，第三碼是詞形標碼，第四碼為義面編碼（如圖32所示）



詞條	詞義序號	詞形標碼	義面編碼
他	01	01	01
他	02	02	02
他	03	03	03
他	04	04	04
他	05	05	05
他	06	06	06
他	07	07	07
他	08	08	08
他	09	09	09
他	10	10	10
他	11	11	11
他	12	12	12
他	13	13	13
他	14	14	14
他	15	15	15
他	16	16	16
他	17	17	17
他	18	18	18
他	19	19	19
他	20	20	20

Figure 32: sense tagging 介面

至於在自動預測方面，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，而上述的人工標示需要耗費大量的時間和昂貴成本，造成語料庫標示語意工作的難產。為了克服此問題，我們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。

這套半自動詞義標示方法是先對語料作初步的詞義標示處理，以作為人工標示之前置作業。我們所設計之半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 33 所示。

首先蒐集一些已被標示過詞義的資料作為詞義標示工作的種子訓練資料。其來源主要有兩個部分，第一個部分為詞義區分詞典中之例句，第二個部分為辭

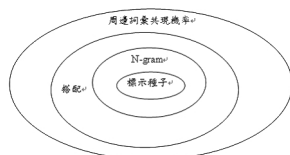


Figure 33: 標示詞義系統組織圖

典編撰小組，在搜尋整理詞義過程中所標示的語料庫部分內容。若來自這兩部分的例句數量不足時，我們會隨機從研究院語料庫中選出部分文句，交由人工標示詞義後加入成為種子標示句。將上述已標示資料合為訓練集，以本文選出來的56篇標示集文章，則作為測試集。

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。本文利用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。我們先以詞頻、搭配詞與目標詞彙距離變異量等條件作為選擇搭配詞彙之初步依據，最後再經過相互資訊MI計算來檢驗搭配詞與目標詞彙之間的相關程度。

經過N-gram 及搭配資訊兩個階段的處理工作，我們將訓練語料標示量做了實質擴增。接著，再經過機率模式計算，盡可能將大部分詞彙標上詞義資訊。最後為求標示語料之高精準度，我們將經由自動標示詞義處理過後的整個標示語料，再交由原字典編撰小組成員進行人工校正處理。

整個自動標示部分之實驗結果我們分為兩部分說明，第一部份詞義標示以詞義下再細分至義面為準，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，整體的正確率可提升至64.51%。

為了避免以人工標示詞義，而需要耗費大量的時間和昂貴成本，我們試圖利用機器比對的方式作為前置作業，因此，我們對於詞義預測，也採用了分群的分析方法，在以分群分析的詞義預測研究中，主要有兩個策略：一、詞形相似分群的分析；二、概念相似分群的分析，而概念相似的分析。在詞形相似分群階段，我們的目標工作是對在目標動詞的特徵擷取步驟中，所產出的特徵詞彙做分群。我們使用dice係數 (Dice, 1945)的計算方式來計算兩個詞彙的相似度，再以決定分群結果。例如：飯和米飯，共同的詞素是飯；案和案件，共同的詞素是案。一開始，我們假設每個詞彙都自成一個群組，接著計算兩兩群組間的相似度，並將擁有最大相似度的兩個群組合併成為新群組，直到所有群組結合完畢或者到達我們設定的群組數。在詞形相似分群階段的準確度約有61.08%。

在詞形分群演算法僅考慮特徵詞彙的詞形，雖然能將一些詞形相同、詞義相近的詞彙分到同一群，但同時也可能將詞形相同，但詞義並不相近的詞彙分到同一群，造成分群上的錯誤，如：山藥、藥。

因此，我們提出一種將特徵詞彙透過知網 (HowNet)轉為概念，再透過分析每個詞彙的義原組合來進行相似度的計算方式，做為分群的依據。在概念相似分群的分析中，又有兩個步驟：1)義原相似度；2)概念相似度。對於每個特徵

詞彙我們透過知網擷取出其概念的組成義原。由於多個詞彙可能會對應上同一組概念，這些詞彙某種程度上可以被視為同義，如「西瓜」、「柿子」、「蘋果」、「葡萄」...等的概念都是水果，都可以視為同義，應該要被分到同一群，也就是說，可以將概念做為特徵並計算概念相似度。對於計算概念相似度，我們使用了修改過的dice係數計算公式來做概念相似度的計算，每個概念都自成一個群組，計算群組間的兩兩相似度，並將擁有最大相似度的那兩個群組合併成為新群組，反覆歸併直到群組達到設定的群組數。在概念相似分群的分析中，其準確度則有85.90%。

為了檢測以分群分析的詞義預測的有效性，我們以Chinese Wordnet的詞義區分結果作為檢測對象，在詞形相似分群的分析階段，詞義預測的平均召回率 (recall)有82.25%；在概念相似分群的分析階段，詞義預測的平均召回率 (recall)則高達90.66%。

#### 4.4 詞彙網路模擬

「意義之模型與量度：中法動詞語意圖形模式之跨學科整合之研究」計畫(Model and Measurement of Meaning: A Cross-lingual and Multi-disciplinary Approach of French and Mandarin Verbs based on Distance in Paradigmatic Graphs，簡稱M3)以台法跨國合作之方式，對於詞彙語意之計算與心理模型進行跨學科整合之研究。在心理語言學之實驗設計上，首先錄製十七個影片，每個影片中會有人做一個動作，例如「撕報紙」、「削紅蘿蔔」等。在受試者觀看影片後，使其描述影片中的動作，採集所回答的動詞作為語料。十七個影片所表達的動作如圖34所示。

/TO DETERIORATE/	/TO TAKE OFF/	/TO SEPARATE/
1- bursting a balloon	6- peeling a carrot with a peeler	12- sawing up a piece of wood
2- crumpling-up a piece of paper	7- peeling an orange by hand.	13- making bread-crumbs with one's hands
3- breaking a glass with a hammer	8- pulling the bark off a log	14- cutting a slice of bread with a knife
4- crushing a tomato with the hand	9- undressing a baby doll	15- breaking off a piece of bread with one's hands
5- tearing up a newspaper	10- taking down a structure made of a Lego set	16- cutting up parsley with a knife.
	11- peeling a banana	17- tearing a shirt

Figure 34: 影片所表達的十七個動作

受試者包含七十九位小孩，分佈在四個年齡組：三歲有二十位、五歲組二十位、七歲組二十位及九歲組十九位。另有六十位成人，以其語料作為小孩語料的對照。採集到的語料分項進行心理語言學及計算語言學的分析，然後再進行台法雙方結果的交叉比對(如圖35)。

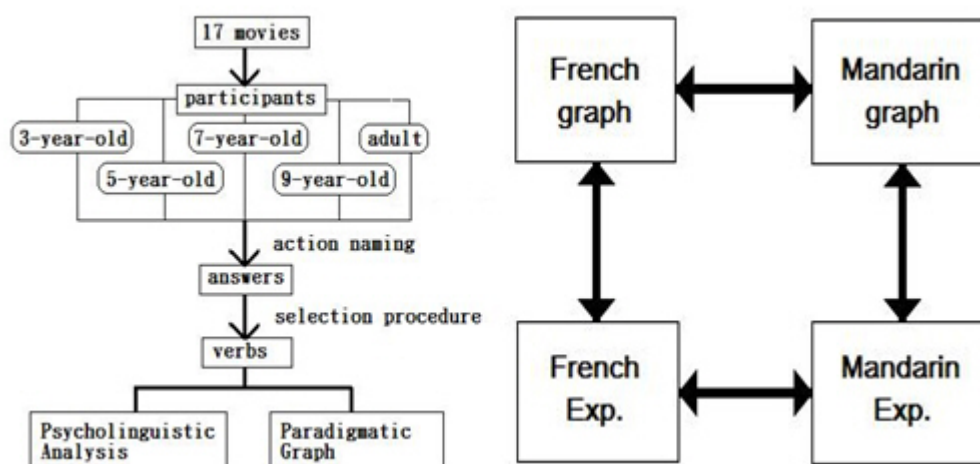


Figure 35: 研究流程圖示

在計算語言學的研究方面，我們主要提出了一種關於動詞語意之三維圖形表達模型，計算三維語意空間中詞義相似度之演算法。除了使用上述實驗語料之外，也利用中文詞網的詞彙建構三維圖表。關於此圖形表達模型，具體說明如下：

詞彙在人類腦海中的連接方式與彼此間的距離是個有趣的議題，當我們提到或是聽到某一個詞彙，腦海可能馬上就會聯想到其他的詞彙，而這些詞彙彼此間也許存在可能是同義、反義或是其他隱藏的關係。如同人類世界中的關係存在著六度分隔理論(小世界現象)，若是把腦海中的詞彙庫作為一個語言世界，我們是否可以大膽假設其間也存在著相同的理論，即不同詞彙間最多僅需通過六個與其有關係的詞彙便能連接。由於CWN的詞彙已經對各個詞義分析出許多的語義關係，因此我們可使用這些資料來作實驗。

資料來源

中文詞網(Chinese Wordnet，以下簡稱CWN)：使用至2010年2月為止的CWN資料，共有10,533個lemmas，包含30,989個詞義(senses)，另外還有9種類型的語義關係 (relation types)，其數量分別如圖 36所示。

		數量
Lemma		10,533
Sense		30,898
Relations	Synonym 同義	28,815
	Antonym 反義	4,636
	Holonym 整體	28
	Meronym 部分	12
	Hyponym 上位	971
	Hyponym 下位	956
	Variant 異體	1,560
	Nearsynonym 近義	1,840
	Paronym 類義	2,351
	total	41,169

Figure 36: 中文詞網相關統計資訊(至2010年2月)

我們僅挑選與其他動詞具有同義關係(此處我們擴大包含異體關係及近義關係，以下同)的動詞作為實驗資料，也就是說我們將捨棄沒有任何同義關係的動詞，原因將在後段說明。最後我們共選出3,499個動詞。

同義詞詞林(以下簡稱Cilin)：雖然CWN擁有十分詳盡的詞義數量(深度)和語義關係，但在所擁有lemma的數量(廣度)上卻不多，因此我們另外增加同義詞詞林中的詞彙來擴充lemma的涵蓋率。

而從M3計畫原來的實驗語料中，我們僅考慮成人的答案，因其具有代表性。從中選擇4個使用頻率最高，並且沒有出現在CWN和Cilin的4個詞彙，分別是『敲碎、鋸、脫掉、撕成兩半』。

計算距離 完成資料的篩選後，我們接著要計算每個動詞之間的距離。我們首先定義詞彙之間若有詞義存在同義關係(此處我們擴大包含異體關係及近義關係)，則將這兩個詞彙視為鄰居，可以透過對方與其他同義的詞彙相連，一步步的往外擴展連連結。接著我們定義詞彙之間的權重值(weight)，代表的是彼此之間有多少個詞義具有同義關係，若兩詞彙間共有3個詞義具有同義關係，則權重值為3，若是無同義關係則為0。兩詞彙間的權重值越大，代表同義的詞義數量越多，因此詞彙間的可替換性越高，相對的距離就越短。由於僅考慮同義關係，因此沒有任何同義關係的詞彙將無法與別人連接，故在資料篩選



時即捨棄之。

CWN的動詞彼此間本身已有標記語義關係，經過統計後可得到16,815個權重值(非0)。

針對Cilin部份，兩個詞彙出現在同一行內則視為彼此間具有同義關係，共得到100,657個權重值。

M3語料的部份，由於新增的4個最高頻動詞並不屬於CWN與Cilin，因此我們定義若是其他詞彙(僅考慮CWN或Cilin中出現的)與新增之詞彙有出現在同一個影片的 結果裡(僅考慮成人的答案)，則視為彼此間具有同義關係。由於在CWN中，同一詞形的詞可能會分為不同lemma，如「看」有「看1」、「看2」、「看3」等，與其他資料來源不同，因此我們僅分辨詞形，將「看1,看2,看3,...」統一視為「看」，重新計算CWN動詞間的權重值。我們將三筆資料來源的權重值加總計算，一共有13,439個動詞以及111,985個權重值。

實驗結果 圖37與圖38分別是以「看」和「裝」為例，與其最相近的50個詞彙。詞彙間的可藉由其他鄰居(與其同義的詞彙)與其他看似不相關的詞彙連接，並且相近的詞彙將聚集在同一處，而即使同詞形但不同lemma的詞最後也將經由與其他動詞之間的關係在圖上區分開來。

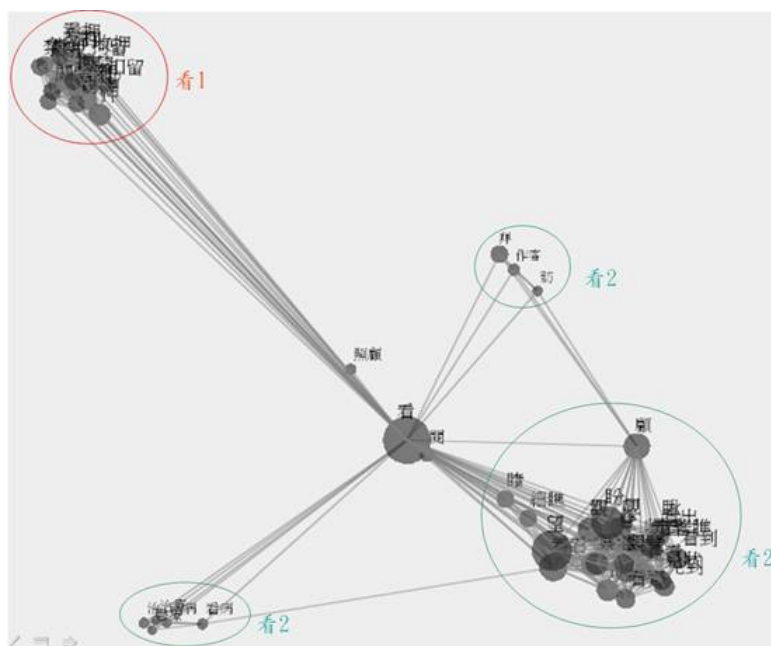


Figure 37: 中文詞彙詞義區分以看為例



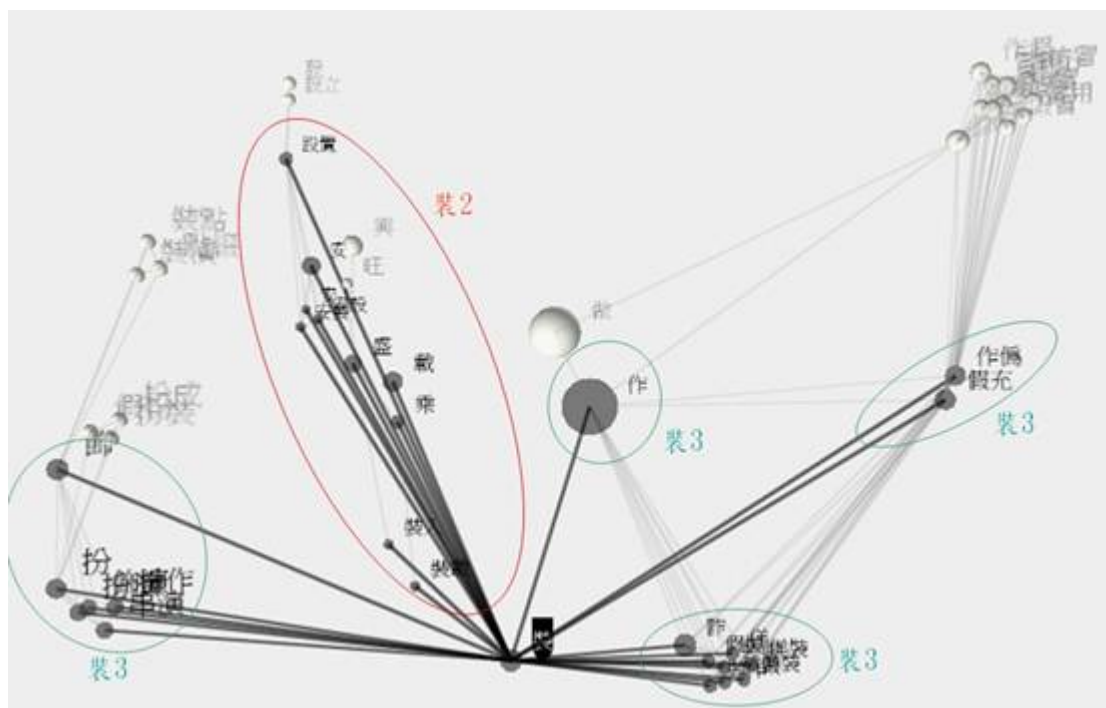


Figure 38: 中文詞彙詞義區分以裝為例

但連接的中介點數量過多(路線過長)，以及無法擴及所有詞彙，推測可能原因：1. 資料涵蓋率過低：中文字(詞)與詞義涵蓋範圍非常廣大，而能蒐集的資料僅佔其中一小部份。(節點不夠多＝鄰居的損失＝可能的連接路線消失) 2. 僅考慮同義關係：也許可增加反義等其他語義關係，甚至是其他非語義的關係。