國立交通大學

理學院(網路學習)學程

碩士論文

宋詞斷詞與本體論之建置

Building a Semantic Ontology with Song Ci Segmentation

研究生：許薰尹

指導教授：曾憲雄　教授

中華民國　九十五　年六月

宋詞斷詞與本體論之建置

Building a Semantic Ontology with Song Ci Segmentation

研 究 生：許薰尹　　Student：Shin-Yean Hsu

指導教授：曾憲雄　　Advisor：Shain-Shyong Tseng

國 立 交 通 大 學

理學院(網路學習)學程

碩 士 論 文

A Thesis

Submitted to Degree Program of E-Learning

College of Science

National Chiao Tung University

in partial Fulfillment of the Requirements

for the Degree of

Master

in

Degree Program of E-Learning

June 2006

Hsinchu, Taiwan, Republic of China

中華民國九十五年六月

宋詞斷詞與本體論之建置

研究生：許薰尹　　指導教授：曾憲雄 博士

國立交通大學理學院網際網路專班

中文摘要

宋詞又稱長短句，繼唐詩之後成為流傳千古的中國藝術結晶。由於宋詞採韻文書寫，對於現代人而言，不易學習。若能建構一個提供宋詞詞彙相關知識的本體，描述詞彙的語意，以及詞彙之間的關係，便可幫助現代人了解詞彙的含意。而欲建置本體的首要工作便是對詞句進行斷詞，以從中獲取所需的相關知識。

在本論文中，我們提出一個針對宋詞需根據詞牌倚聲填詞，按節奏停頓，以及宋詞特有的領字等特色進行斷詞，並透過詞彙語意的描述，來建置宋詞詞彙本體以輔助學習。論文包含兩大部份：**宋詞斷詞器**與**本體論建置**：

宋詞斷詞器利用規則式（Rule-Based）斷詞方式，截取詞句中的詞彙。包含六大斷詞模組：專有名詞、領字、典故、構詞模組、節奏斷詞模組、對仗模組。從斷詞實驗結果得知，召回率、精確度和效度最高可達 90%。

本體論建置則是將斷詞後所得到的詞彙，進行語意概念的分類，以及詞彙的前後連接詞彙、詞類、詞頻、同義詞、近義詞、反義詞、對仗詞與平仄等語意的描述。我們設計了語意編輯工具編輯詞彙的相關資訊，並且自動產生表達本體知識的 OWL 文件，大量降低本體建置的負擔。最後，我們設計「絕妙好詞」網站，讓使用者可以很容易地透過網際網路，檢索詞彙語意資訊，進行線上學習。

關鍵字：斷詞、中文斷詞、本體論、宋詞

Building a Semantic Ontology with Song Ci Segmentation

Student: Shin-Yean Hsu Advisor: Dr. Shian-Shyong Tseng

College of Science
National Chiao Tung University

Abstract

The Song Ci, known as Long Short Sentence, is the art of the ancient Chinese after Tang Poetry. Since Song Ci was written by verse (韻文), it's hard for modern people to learn. If we could construct an ontology to describe the semantic of words in Song Ci and the relationships among them, learning and the understanding of Song Ci will became easier. Before building the ontology, we will segment words contained in the sentence of Song Ci, and acquire all related information for this purpose.

In this thesis, we propose a method according to Ci Pai (詞牌), rhythm of poetry, and the Empty word (領字) of Song Ci to segment words. After that, we construct an Song Ci ontology based on the semantic of words. This thesis contains two parts: **Song Ci Parser** and **Ontology Building Module**.

Song Ci Parser, a rule-based parser, includes six modules for Song Ci segmentation: **Proper Noun Module**(專有名詞模組), **Empty Word Module**(領字模組), **Literary Quotation Module** (典故模組) , **Word building Module** (構詞模組) , **Rhythm Module** (節奏斷詞模組), and **Pair Module**. The experimental results show that the finest recall, precision, and effectiveness rate are 90%.

Ontology Building Module will use the words preprocessed by **Song Ci Parser** to build an concept hierarchy of words in Song Ci. Finally we design a Semantic Editor to

describe the semantics of word, E.g. Ci Pai (詞牌) , author name, frequency of words, word type, previous word, next word, antonym, near synonym, Synonym, etc. Finally, we build the “絕妙好詞” web site for people to learn the semantic of words from internet.

Keywords: Segmentation, ontology, SONG-DYNASTY, Ci

誌謝

這份論文能夠完稿以及順利付梓，不是全憑我一己之力就能完成，而是藉由許多人的幫助才能產出的心血結晶。兩年來，在探索斷詞、本體相關知識的過程中，與老師、學長經過無數次的會議與討論，除了從中獲得許多寶貴知識外，也了解許多老師與學長們作學問、做人做事的哲學。

首先，感謝指導教授曾憲雄老師，為迷惘的我指引方向，讓我能夠撥雲見日，找到論文的方向，也才能如期地完成研究。同時要感謝元智大學羅鳳珠教授不吝於提攜後輩，為我提供文學知識的解答。當然也要感謝口試委員楊錦潭教授、莊祚敏教授、曾秋蓉副教授對於論文的寶貴建議，以讓論文更加完善。

其次要感謝楊哲青學長，犧牲無數個人時間，幫助我進行實驗，以及論文的寫作與校稿。同時也謝謝蘇俊銘、翁瑞鋒學長給予的建議與指導。謝謝迺仁、碧如、亮中、宗平、秀怡、書源、鳳琴在兩年的研究生活中不停的為我打氣，以及負責幫忙口試相關的事宜。謝謝所有實驗室的其他伙伴們，在研究生涯中對我的幫助。最後謝謝一直陪伴在身邊的家人與朋友們。

許薰尹

2006 年 6 月

目錄

表目錄

圖目錄

演算法目錄

第一章 緒論

幾千年來，中國的文人騷客，將情感寄託於詩詞歌賦之中，以文字表達豐富的情懷，寫下漢賦、唐詩、宋詞、元曲等等作品傳頌於後世，可堪稱是中國古典文學藝術之極致。這些文學作品常能表述我心，臨長江水岸，會不禁想起「亂石崩雲，驚濤裂岸，捲起千堆雪。」（蘇軾《念奴嬌 赤壁懷古》）一詞的意境；在失意時，以「三十功名塵與土，八千里路雲和月。莫等閒、白了少年頭，空悲切。」（岳飛《滿江紅》）舒發胸懷；在失戀時，無奈感嘆「多情自古傷離別，更那堪，冷落清秋節，今宵酒醒何處，楊柳岸，曉風殘月。」（柳永《雨霖鈴》）。

1.1. 研究動機

有人曾經這麼說過：「知識傳統是必須的，唯有知識化，才能保存已知，進而探求未知。」傳承、推廣中國文學，讓大眾了解古典文學之美，才能讓這些前人的智慧與心血能夠永遠地流傳。不過在欣賞或研讀這些作品的過程中，發現最大的困難點在於現在的人對於古典文學的知識化程度不夠，縱使有心者想要學習也無著力之點。大多數的原因是這些文學使用的多是韻文，與目前我們日常流通使用的白話文、語體文有著些微的差異。甚至是因時代的變遷，以致原韻文的涵意在現代有了新解，有些詞彙在現在已有完全不同的解釋，無法了解宋詞中使用詞彙的語意，這都讓接觸古典文學的人倍感挫折。

近年來，資訊科技技術的蓬勃發展，帶動數位典藏的風潮，這些代表中華文化的藝術與內涵的文學與典籍，藉由網際網路的觸手，漸漸伸展觸及國際舞臺。若能借助資訊科技之力，充份利用網際網路無空間、時間的障礙，做為古典文學推廣學習的推手，定能使知識的傳承更為容易。

韻文中，宋詞因具有音樂性的特色，在現代也常被配樂演唱。這種在流行音樂的旋律中搭配古典文學的氣息，於新潮中帶點懷舊的韻味，值得令人玩味。從事資訊科技工作多年的我，對於古詩詞也相當地感興趣。一個是走在時代尖端的科技，一個懷舊的古典文學，如何結合兩者的特色，讓古典文學能夠透過資訊科技，更容易進行網路學習，而不是只在網際網路上提供詞句的檢索。融合兩者，讓古典文學透過資訊科技更容易透過網路傳遞，以便更容易學習就成為本篇論文

的研究動機。

本研究希望能夠利用資訊科技，建構一個提供線上輔助學習的本體論。描述宋詞詞彙的語意和概念分類，幫助學習者能夠了解詞彙中的語意、隱藏於字面下的典故，和作者寄情於詞句之中的情感。但要建置一個符合需求的本體論之前，還需要收集宋詞相關的資訊，加以分門別類、研究匯整，而對宋詞斷詞得到詞句中所採用的詞彙就是建置本體論之前的首要工作。因此，本研究將以宋詞斷詞為基礎，先萃取出詞句中的詞彙，再為這些詞彙加上語意資訊，以建置出宋詞詞彙本體，達成輔助線上學習的目標。

1.2. 研究目標

由於目前中文斷詞的研究都偏向處理語體文，對於韻文處理的研究較少，韻文和語體文之間的差異，使得不管是採用現存的自然語言、或規則式的中文斷詞系統對宋詞斷詞的精確度都偏低，常會將一些詞組切割成一個個的單字詞。

本篇論文以宋詞與生俱來的特色來建構一個斷詞系統：宋詞需根據詞牌倚聲填詞，按節奏停頓，而節奏點的位置常是詞彙的識別項目。再者，宋詞中使用特有的領字句（虛字）結構，除賦予吟唱時能增添抑揚頓挫的風味之外，也能輔助辨識出虛字詞。此外本論文嘗試辨識專有名詞、典故詞彙，以解決傳統斷詞方式將這些名詞切為雙字詞而喪失原意的問題，如「白鷺洲」是地名，但是被切成「白鷺/洲」。最後再利用詞中對仗的特色協助斷出三字詞。

本研究以規則式（Rule-Based）斷詞方法為基礎，以決定要採用的斷詞模組和斷詞順序。斷詞模組包含：專有名詞、領字、典故、構詞規則、節奏斷詞與對仗模組。最後，從斷詞的結果中萃取出詞彙，來建置詞彙語意的本體。本體論中包含了詞彙的詞類、詞頻、前後詞彙、同義字、反義字、近義字、對仗字等相關的資訊，以幫助學習者了解詞彙的語意。

論文包含兩大部份：**宋詞斷詞器**是用以將宋詞的詞句進行斷詞；**本體論建置**則是利用斷詞處理後得到的字組進行概念上的分類，以及詞句使用的背景和語意的描述。

本論文提出的貢獻如下：

(1) 宋詞詞彙本體知識的建立。

(2) 宋詞斷詞召回率與精確度可達 90%。

1.3. 論文架構

本篇論文第一章為緒論，介紹研究之動機、方法及論文之架構，第二章介紹宋詞的基本知識、相關的研究背景與語料庫的建置。第三章則描述宋詞斷詞器的設計，以及如何建置宋詞詞彙本體的細節。第四章為實驗成果，最後一章為本篇論文的結論與未來的展望。

第二章 研究背景與相關資源

本章節將介紹宋詞的專有名詞、相關術語等基本概念，並描述本體論與中文斷詞的相關研究。最後說明研究過程中參考到的詞庫、資料庫的建置與收集。

2.1. 宋詞簡介

詞起源自隋唐時代，經五代到兩宋，達到詞的全盛時期。它上承唐詩，下啟元曲，為中國文學史上重要的文學體裁之一，也是詩歌的主要形式之一。在詞初創時，具有音樂性，能夠配合樂曲演唱。一般是先有樂曲的存在，再根據樂曲的長短、節奏來按樂填入詞句，也因此寫詞被稱之為「依聲填詞」。直到南宋時期，漸漸的和音樂脫離，不再用來歌唱，成為一種具有特殊形式的文學。

詞是詩的一種，但是又和近體詩[1]不太相同。近體詩的格式只有幾種，對於詩的字數、句子數目、平仄和用韻方式都有一定的規定。但詞更為複雜，規定更為嚴格，有一定的詞調和格式，在字數、句子數目、平仄和用韻和格律[2]方面都受到樂律的規範。

2.1.1. 詞的別名

詞來自於民間文學，在初期發展時是專為搭配樂曲而演唱的，因此最初的名稱被叫做「曲子詞」或「曲詞」，也有人認為是從樂府[3]演進而來，因而也稱為「樂府」；有些人認為詞是由近體詩演變而來的，因此詞也稱為「詩餘」指的是「詩之餘享」的意思。

詞又有「長短句」的別稱，詞句中的字數從三到十一個字不等，形式比較活潑，但全篇的字數是一定的，每一句的平仄也有一定。

2.1.2. 詞的一般用語

- **闋**

[1] 格律詩，是唐代出現的新詩體，分成律詩和絕句兩種。
[2] 格律，指韻文的體裁與格式的法則。如篇幅的限制、句法、字聲、用韻的規定。
[3] 樂府，在古代指專門掌管音樂的官署，主要任務是製定樂譜、培訓樂工、蒐集歌詞。樂府詩，是指此官署為配製樂曲而由文人製作或從民間採集的詩歌。魏晉時期開始把「樂府詩」簡稱為「樂府」。

樂曲終了稱為闋，一首詞稱為「一闋」，表示曲子到此已經結束了。

- **單調、雙調、三疊、四疊**

  詞有單調、雙調、三疊、四疊的區別，詞可以分多段，類似今日在文章內分的段落。不分段的稱為單調，分為兩段的稱雙調，分三段稱為三疊，分四段稱為四疊。

- **片**

  詞的每一個分段稱做「片」。雙調的上段又叫「上片」，下段叫做「下片」。也有稱「上闋」、「下闋」。

- **詞調**

  詞一開始是配樂曲歌唱，因此每首詞都會有樂譜，詞牌使用的曲調稱為宮調[4]，具備一定的音律與節奏。

- **詞牌**

  詞的格式的名稱。每一種詞調都擁有一個名稱，如《念奴嬌》、《菩薩蠻》、《西江月》，這個名稱就是指「詞牌」。若採用一個詞牌填詞，就必需根據此詞牌定義的段數、句子數目、字數、平仄、用韻方式來填寫，若不照規定填詞，就不能夠使用這個詞牌的名稱。

- **詞譜**

  詞譜的著作是從明代開始，將每一個詞牌編成歌譜。根據清朝萬樹《詞律》一書中描述，詞譜共有一千八百多個，而清代的《欽定詞譜》中則錄有二千三百零六個。

- **小令、中調、長調**

  從宋代《草堂詩餘》一書開始，習慣依詞的字數多少將詞調分為小令、中調和長調三類。清朝毛先舒則在《填詞名解》書中描述：「五十八字以內為小令，自五十九字始至九十字止為中調，九十一字以外者俱為長調。」不過，這只是一種大致的分類，並非所有人都同意這個說法。

2.2. 本體論工程

本體論[1][2][3]源起於哲學領域，探討存在問題。在近幾年來本體論被應用到人工智慧領域，用來表達知識，很快地成為熱門的知識工程技巧，讓知識能夠不受限於開發工具，或應用平台而能夠透過本體共用或重複利用，不需從無到有重

[4]宮調，類似今日的C調、G調。

新建置，以節省大量的人力、物力上的浪費。

2.2.1. 本體論工程研究方法

近幾年本體論工程[1]的研究相當地蓬勃，比較常見的方法論包含：

(1) Mike Uschold dde & King [3]的「骨架」法

由愛丁堡大學人工智慧研究所及它的合作伙伴共同製定，建立在企業本體的基礎之上，提供相關企業之間相關術語和定義的集合。

(2) Gruninger & Fox的「TOVE評價法」[2]

由多倫多大學提出的多倫多虛擬企業本體（TOronto Virtual Enterprise Ontology），使用一階邏輯進行描述。TOVE 本體包括了企業設計本體、工程本體、計劃本體和服務本體。

由於 TOVE 評價法採用一階邏輯進行描述，很容易結合其它的人工智慧系統或本體表示語言一起使用。同時它也是現階段較為嚴僅的本體建置方法，故本研究以 TOVE 評價法為基礎，做為建立宋詞詞彙本體的參考方法。

2.2.2. TOVE 本體論工程

TOVE本體論工程將本體論的建置分為六個階段，參考圖 1：TOVE本體論流程：

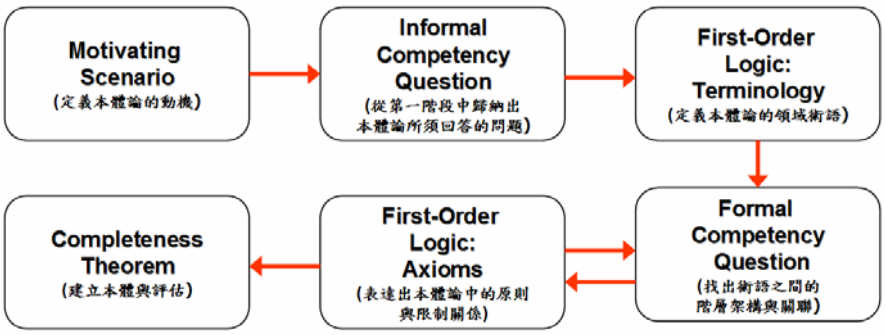

圖 1：TOVE 本體論流程

(1) 定義本體論的動機：

此階段定義本體論的動機，描述企業面臨的問題，以及此本體論所須解決的問題，和可能的解決方案。這個階段可以了解本體的用途，幫助後續本體的維護動作。

(2) 從第一階段中歸納出本體論所須回答的問題

從第一階段中歸納出本體論所須回答的問題，描述本體的需求。此階段也成為將來評估本體論的重要條件。

(3) 定義本體論的領域術語

此階段定義本體論會使用到的詞彙和詞彙之間的關聯性，也就是定義本體論的領域術語。TOVE 採用一階邏輯（First-Order Logic）來表達詞彙和關聯性。

(4) 找出術語之間的階層架構與關聯

將第二階段的問題轉成標準的表示式，並找出術語之間的階層架構與關聯。標準表示式將成為評估本體的準則。

(5) 表達出本體論中的原則與限制關係

使用一階邏輯表達出本體論中的原則、公理（Axiom）與限制、邏輯關係。

(6) 建立本體與評估

建置本體，並評估所建立的本體是否符合第一階段的動機。

2.3. 相關研究

2.3.1. 語意網與知識本體

Tim Berners-lee曾提到他對網路有二個夢想[4]：一是所有的人都可以透過全球資訊網（www）來共享知識；二是未來的網路將是語意網（Semantic Web），電腦藉此可以了解人的語言。而源起於哲學領域探討存在關係的本體知識，近來頗受資訊界的重視，將之視為知識、訊息的底層架構，進而用來描述特定領域中的概念，與概念之間的關係。簡言之，語意網就是由知識本體定義網頁上的語意，讓機器（代理程式）能夠閱讀語意，幫助我們獲取資訊。

組成語意網三大要素：

(1) XML（eXtensible Markup Language）

(2) 資訊描述架構（Resource Description Framework，RDF）+ URI [5]

(3) 知識本體[6]

藉知識本體（Ontology）定義關鍵詞，再利用 RDF（資源描述架構）與 URI（通用資源標誌碼）連結到相關網頁或資源，透過超連結找到關鍵詞，進行邏輯

的知識推理。

有鑑於此，IEEE標準上層知識本體工作小組建置了SUMO建議上層共用知識本體（Suggested Upper Merged Ontology）[6]，發展標準的上層知識本體，以促進資料互通性、資訊搜尋和檢索、自動推理和自然語言處理。並希望能藉由最高層次的知識本體，鼓勵其他特殊領域知識本體以其為基礎衍生出其他特殊領域的知識本體，並為一般多用途的術語提供定義。

中研院領域知識本體[7]以SUMO為上層，建立了兩個與韻文較相關的中文本體知識：唐詩三百首知識本體，以及蘇軾詩知識本體。這兩個本體都是針對詩領域建置。

韻文學習已變成缺專家不可，若能利用知識本體描述宋詞詞彙的涵意，就能夠解決專家不能隨時在旁提供建議的問題。而要建置宋詞詞彙的本體知識之前，需要從宋詞中擷取出詞彙。為了讓機器幫助我們完成這個工作，避免大量的人力負擔，首要工作就是進行機器自動斷字處理，以便萃取出宋詞詞彙，進行本體的建置。

2.3.2. 中文斷詞

中文斷詞在自然語言領域是一件基礎但是又相當重要的工作，例如應用在處理文件檢索、語音辨識、機器翻譯、中文輸入等等。雖然中文斷字處理蓬勃發展，但絕大多數以研究語體文（白話文）為主，針對韻文分詞、判別、語義、句剖析、知識萃取與表達研究少。同時，詞彙切分與語意標記困難，也造成對韻文斷詞的精確度不高。

目前中文斷詞研究中最常遇到的問題包含：

(1) 要如何進行斷詞?

中文的詞和英文不太一樣，詞與詞之間沒有空白以便區分，增加斷詞的困難度。目前中文斷詞普遍採用的斷詞方式約可分為：法則式、統計式與結合兩者特色之混合式斷詞法。法則式的缺點在於難以歸納所有的斷詞規則；統計式斷詞方法需要大量的語料庫以統計出數學模型，這些語料不易取得，需花費大量的時間與人力準備與整理。混合式斷詞法則是取兩者之優，成為比較通用的斷詞方法。

(2) 歧義性的問題

一個中文的句子，可能會有許多不同的斷詞組合，舉例來說「上鎖窗」這一個句子要斷成「上鎖/窗」或是「上/鎖窗」？要如何挑選出正確的斷字組合也是中文斷詞處理的一大挑戰。

(3) 未知詞問題

未收入在辭典中的詞稱之為未知詞，辭典很難收錄所有的詞彙。若辭典太龐大會降低斷詞效能，辭典太小又會降低斷詞精確度。

除了上述中文斷詞面臨的斷詞問題之外，針對宋詞部份還需考慮到：

(1) 宋詞無法完全根據詞類、文法斷詞

就語言的角度來看，學習英文時可以透過文法搭配句中的詞類，如主詞、動詞、受詞，以便了解英文句子的含意；但宋詞詞句結合優美與精鍊的特色，常以簡潔的字詞蘊含無窮意義，如連續出現三個形容詞「淒淒慘慘淒淒」，因此沒有任何填詞的詞句文法，或詞類的使用規則可以參考。

(2) 詞的倒裝、隱喻、典故手法，使斷詞與語意解析困難

宋詞常藉倒裝、隱喻、典故手法增加其風味，如：

翠貼蓮蓬小，金銷藕葉稀

－李清照《南歌子》

「翠貼」、「金銷」兩句皆是倒裝句，指的是貼翠和銷金的兩種工藝，描述以翠羽貼成蓮蓬樣，以金線嵌繡蓮葉紋。而蓮小藕葉稀，又暗喻離別容易會面難，愛情生活的短暫。簡短兩句詞中，包含許多內容、許多感情。這些都不是單純從字面中可以看的出來的。

(3) 專家對斷詞結果自有不同見解

同樣的詞句在不同的專家眼中自有不同的看法，舉例來說：

醉後明皇倚太真。

－李清照《瑞鷓鴣》

此句詞描述唐明皇與楊貴妃的愛情故事，若以典故為觀點，它是一個詞，不宜斷之；若以中文斷詞的正確性觀之，又應斷開為「醉後/明皇/倚/太真」。這些問題都造成斷詞上的困難。

在中文斷詞方面，中研院詞庫小組已設計出中文斷詞系統，並對斷完的字詞進行詞類的標註，也能自動尋找未知詞。比較可惜的是它主要是針對語體文處理，對韻文的處理較不精確[8] 。比如說：

藤牀紙帳朝眠起

－李清照《孤雁兒》

這一句輸入斷詞系統，這個系統幾乎將所有的字斷成單字詞：「藤/牀/紙/帳/朝/眠/起」。就語文學家的觀點而言，應該斷為：「藤牀/紙帳/朝眠/起」又如：

共賞金尊沈綠蟻

－李清照《漁家傲》

此系統斷為：「共/賞金/尊沈綠/蟻」，而實際上「金尊」是指酒器，而「綠蟻」是一種酒。較正確的斷法應該是：「共賞/金尊/沈/綠蟻」。

和韻文較相關的研究還包括了中央研究院語言所建置的「宋詞全首閱讀」網站[9]，提供宋詞斷詞與閱讀的功能。但收入的中文詞不夠多，像上述的綠蟻中的「蟻」字與「紙」字就不存在語料庫之中，因而無法進行斷詞。

中國南京師範大學張成等人建置「唐宋金元詞文庫及賞析系統」[10]，提供唐、宋、金、元朝詞的賞析與按詞牌、詞序等組合檢索功能。此外也提供了《全宋詞》詞頻統計資訊，收錄共 21,085 首，共 1,417,695 個字，6167 漢字的資訊。但缺乏語意、或典故的查詢。

中國北京大學中文系李鐸主持開發的「全唐詩電子檢索系統」[11]能按原書順序、按作者、按體裁等方式檢索以及瀏覽。元智大學羅鳳珠與北京大學計算語言學研究所俞士汶、胡俊峰合作的「唐代名家詩語文標記系統」(2000年)[11][12]，以全唐詩及以及宋代名家詩為研究範圍，利用統計方法「共現度」、「結合強度」，以便自動提取詞彙。這個多維的統計抽詞模型，抽詞精確度（Precision）達到 56%，召回率（Recall）達到 89.3%，就整體的滿意度而言還有改善空間。

北大計算語言學研究所與元智大學合作的「宋代名家詩網路系統」[13]，則可對宋詩資料進行全文檢索、自動注音。另外還有「唐宋詩電腦輔助研究系統」[12]，以 640 萬字的唐宋詩語料為基礎，可進行詞彙的檢索、統計以及相關的分析功能。透過電腦輔助的方法來深入地了解唐宋詩語言中描寫的宋朝社會生活，以語言的角度來解析當時各層次社會生活的風貌。不過根據北大的研究發現，許多三個字的專有名詞被切割為二個詞，以致喪失原意，如「鸚鵡洲」，被切分成「鸚鵡/洲」；而「黃鶴樓」，抽取出「黃鶴」詞彙，但

不含「黃鶴樓」。而一些典故，如「傅粉何郎」，卻被切分為一、二字詞。

元智大學羅鳳珠第四屆數位典藏技術研討會中提出「詩詞語言詞彙切分與語意分類標記之系統設計與應用」[13]，對宋詞的切分提出了獨到的見解。但是尚未提及對含領字句[14][15]、專有名詞、典故的句子在切分出這些字詞之後，後續的斷詞方法；再者文中建議以常用的句法（節奏）進行斷詞，但並未提及斷詞的召回率與精確度等相關數據資料，和三字詞的切分規則。

針對以上的問題，本研究以專有名詞、典故模組，切分出詞句中的專有名詞和典故資訊，以解決這些詞彙被切割而喪失原意的問題。再者，我們建置了領字資料庫，做為切割詞句中領字的基礎。除了利用詞句的字數進行斷詞之外，我們提出了根據讀[5]的字數，搭配詞的節奏做為切分的規則，以解決含領字句、專有名詞、典故的句子之斷詞方式。此外，我們利用構詞規則來解決詞庫中不收入定詞、量詞、定量複合詞、疊詞的問題。最後，以宋詞相鄰詞句經常使用對仗的特色，建立對仗資料庫，利用對仗規則來輔助切分三字詞。

2.4. 詞庫與資料庫的搜集和整理

我們使用的斷詞系統參考以下詞庫與資料庫，以下分別說明這些詞庫的規格和用途。

2.4.1. 中央研究院詞庫〔八萬目詞〕

在斷詞器的測試過程中，採用的詞庫是中央研究院建置的中文詞庫〔八萬目詞〕[16]，一共收錄了 78,409 筆記錄，再整合實驗室自行搜集的中文詞彙共 107,738 個詞。

中文詞庫〔八萬目詞〕收錄的詞包含一般用詞、常用專有名詞、成語、慣用語、常用派生詞、異體詞、合併詞以及少數特殊領域用語和古漢語詞語。每個詞項包含的有：詞語、詞頻、注音、詞義分類、詞類標記等相關的資訊。此外，針對通常不單獨出現，需和其它字結合在一起的附著語素（Bound Morpheme，簡稱 BM），詞庫將之標為b以識別[16]。

[5] 一個完整的句子稱為句，半句為讀。

2.4.2. 專有名詞資料庫

中央研究院詞庫〔八萬目詞〕所收錄的資料中雖然有專有名詞的資訊共一千六百七十六個，但這些專有名詞幾乎都是語體文，如中華民族、牛頓、中國國民黨。雖然有記錄古代的專名，如項羽、諸葛亮、楚辭等，但為數不多。本實驗室自行收集專有名詞資料並進行標註，記錄唐宋專有名詞資訊，含：人名、地名、建築物專名、動物專名、植物專名等。

2.4.3. 典故資料庫

宋詞用典的情況相當普遍，用以擴大詞的意境與表現力。不知詞中的典故，可能會誤解原詞的含意。

本實驗室收錄唐詩、宋詞典故資料庫，記錄了典故、典故別、相關的人物、相關典故、同義典故、參見典故、典故出處、出處內容等資訊共 10,517 筆[13][17]。

2.4.4. 領字資料庫

領字又名虛字、襯字、領句、領調、領格字；單字領字又名一字豆、二字領字又稱二字豆、三字領字又稱三字豆，甚至有些詞人認為可以有四至五個字。它是宋詞特有的句法與修辭技巧。利用領字承上句，領下句，讓詞句配樂歌唱時有高低起伏，不至於單調刻板。

領字句的字數歷代的詞人多有爭議，有些人認為領字有一字、二字、三字等三種；有些詞人則認為有四字、五字領字。舉例來說，有些詞人認為「爭忍」、「何計」為兩字領字；「更消愁」、「又莫是」、「最無端」為三字領字。有些人認為這些詞是可以切割成一個領字加一個詞彙，如「更/消愁」。但是有些詞人則抱持不同看法，如沈義甫《樂府指迷・句上虛字》、元代陸輔之《詞旨》、周濟《宋四家詞選目錄序論》、施蟄存《詞學名詞釋義・領字》等書中認為領字應該只有一個字。

本研究以王兆鵬、劉尊明等主編的《宋詞大辭典》為基礎，標註領字在各詞牌出現的位置，建立宋詞領字資料庫，只收錄一字領字。根據元代陸輔之《詞旨・單字集虛》將常用領字：「任、看、正、待、乍、怕、縱、問、愛、奈、似、但、料、想、更、算、況、悵、快、早、儘、嗟、憑、歎、方、將、未、已、應、若、

莫、念、甚。」收錄至資料庫。

2.4.5. 宋詞對仗資料庫

對仗也稱對偶，指的是在兩個長短相同的句子，在相同的地位，它的語義要相當（虛實相當），字調要相反（平仄相反）[18]。利用「奇偶相生、輕重相權」八字對仗法則，讓詞更為和諧以加強藝術效果。舉例來說，三言對：

柳絲長，桃葉小

－晏幾道 《更漏子》

「柳絲」對「桃葉」，「長」對「小」。

五言對：

落花人獨立，微雨雙燕飛。

－晏幾道 《臨江仙》

宋詞詞句中的對仗，有固定的，有一般用對仗的，有自由的[19]。

(1) 固定的對仗

某些詞牌固定在詞中的句子使用對仗，如《西江月》、《蘇幕遮》、《漁歌子》、《鵲橋仙》、《卜算子》、《南歌子》等詞牌多用於前闋的頭兩句和後闋的頭兩句，稱之為「蝦鬚格」。

(2) 一般

一般用對仗代表使用對仗是慣例，但是也可以不用對仗。如《念奴嬌》前後闋第五六兩句；《浣溪紗》後闋頭兩句。

(3) 自由的

詞並不一定要求要使用對仗，詞的某一些句子中，只要相連兩句字數相同，作者經常用對仗來表現，因此只要是前後兩句字數相同的，都有使用到對仗的可能。

簡言之，詞用到的對仗多在相同字數的相鄰句子上，而且是同一意義段落上[20]。不僅字數相同的相鄰句子能用到對仗，而且帶領字的句子，如果除了領字的字數，和後一句的字數相同，也可以用到對仗。例如：

但苔深韋曲，草暗斜川。

－張炎《高陽臺》

「但」是領字，除此字之外，「苔深」對「草暗」；「韋曲」對「斜川」。

本實驗室搜集全宋詞中對仗的資訊，記錄採用固定、一般對仗的詞牌、以及詞牌中哪些詞句互為對仗等資訊整合至對仗資料庫。至於自由對仗的詞與句子較難以評估，故不在收錄範圍。

2.4.6. 同義詞詞林

《同義詞詞林》是由上海外語學院梅家駒、竺一鳴、高蘊琦、殷鴻翔等人共同編輯整理的工具書，以便讓文字工作者在創作過程中，可以查閱同義資訊，避免在寫作時遇到詞窮的問題。從名稱就可猜測它是以中文詞義進行分類，總共分 12 大類，94 中類，1428 小類，參考**表 1**。在小類之下又根據同義的原則劃分不同的詞群，共計有 3,925 個詞群，包含將近七萬個詞義項目[21]。除了詞義的資訊之外，也兼顧詞類的資訊。

表 1：同義詞詞林大類

大類	說明	大類	說明
A	人	G	心理活動
B	物	H	活動
C	時間與空間	I	現象與狀態
D	抽象事物	J	關聯
E	特征	K	助語
F	動作	L	敬語

由於同義詞詞林已經將詞彙分門別類，並建立了詞彙的同義資訊，本研究參考同義詞詞林對詞彙的分類，來建立詞詞彙本體概念階層和詞彙的同義詞資訊。

2.4.7. 常用詞首、詞尾字資料庫

常用詞首、詞尾字資料庫[22]是由中央研究院詞庫小組所建置，此資料庫是根據中研院平衡語料庫為基礎，收集常用的名詞詞首字 1,135 個（含歧義為 1,197 個）、名詞詞尾字 1,427 個（含歧義為 1,610 個）、動詞詞首字 735 個（含歧義為 918 個）、動詞詞尾字 282 個（含歧義為 300 個）、總計 4,025 筆資料。

第三章 宋詞斷詞器與本體論設計

本章首先介紹宋詞斷詞系統的規劃與建置的步驟。此系統結合了宋詞特有的領字、依聲填詞、按節奏停頓的特色，再輔以詞庫、構詞規則進行斷詞。接著說明如何利用斷詞系統擷取出的詞彙建立描述宋詞語意、詞彙之間的關聯，以建立宋詞本體語彙知識網路。

3.1. 系統架構

系統是以宋詞斷詞器為基礎架構，你可以把一闕宋詞輸入斷詞系統，進行斷詞。參考圖 2為本論文的系統架構圖，斷詞系統將參考專有名詞資料庫、領字資料庫、典故資料庫、對仗資料庫、中研院詞庫和構詞規則進行斷詞。針對斷詞完的結果，再解歧義，最後由專家進行校正評估斷詞的成果。

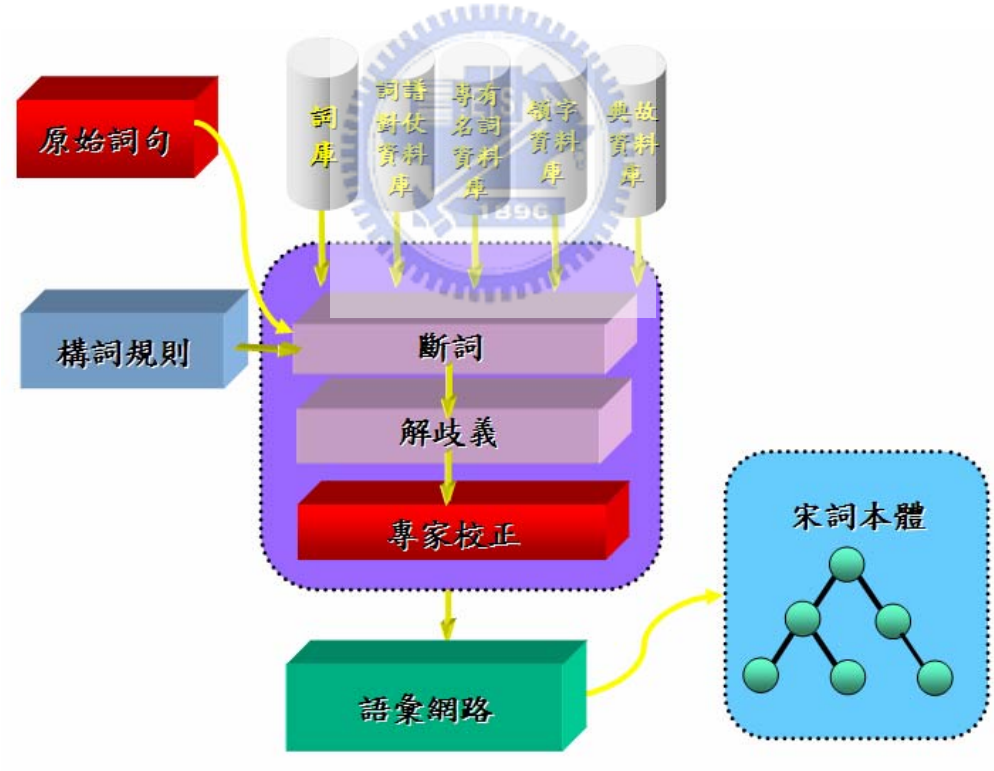

圖 2：系統架構圖

有了從宋詞詞句中萃取出的基本詞彙之後，便可以進一步建立詞彙的相關資訊，如前後詞彙、同義詞、近義詞、詞類等等資訊，再利用這些資訊建立宋詞詞彙語意本體。

3.2. 宋詞斷詞器

3.2.1. 斷詞模組分析

在本研究中，我們首先搜集斷詞的相關文獻與研究，分析常用的斷詞方式的優缺與精確度，再與專家進行討論。根據專家的經驗與建議，建構出一個結合宋詞特色的規則式（Rule-Based）斷詞系統。利用規則中內含的知識，讓系統自動推論要採用的斷詞模組與順序。

參考圖 3，提供的預設六大斷詞模組為：

(1) 專有名詞模組：切分詞句中的專有名詞。

(2) 領字模組：切分出領字（虛字）。

(3) 典故模組：切分詞句中典故資訊。

(4) 構詞模組：切分出定詞、量詞、定量複合詞、複疊詞、詞綴。

(5) 節奏斷詞模組：根據宋詞節奏（句法）切分詞句。

(6) 對仗模組：根據詞句對仗，切分三字詞。

圖 3：宋詞斷詞器

3.2.2. 規則式斷詞方法

這六個斷詞模組的順序若變動，會影響精確度，根據專家建議與觀察詞的句式，我們使用兩種方式來決定這六大模組的順序：規則式斷詞法、使用者自訂斷詞法。

規則式斷詞法的斷詞順序由下面的 Meta Rule 與規則決定：

- **Meta Rule 1：if 詞牌已知 then try Rule 1**

領字模組、對仗模組是根據詞牌建立，因此要採用 Rule 1 斷詞時，這些必需是已知的。

- **Meta Rule 2：if 詞牌未知 then try Rule 3**

若詞牌未知，則可以跳過領字模組、對仗模組的比對。

斷詞模組的順序一開始是參考專家的建議，再從實驗結果找尋出精確度和召回率較佳的組合，我們定義以下三條斷詞規則：

- **Rule 1：if 詞牌已知 and 詞牌是《好事近》、《憶少年》、《醉太平》、《沁園春》、《風流子》 then 順序為 2，1，3，4，5，6**[6]

由於宋詞中的領字句在《好事近》、《憶少年》、《醉太平》、《沁園春》、《風流子》詞牌中已成定式，這代表這幾闕詞使用領字的機率高於其它詞，若能先切分出領字句，就可以降低領字和專有名詞衝突時切分錯誤的機率。

這條規則定義的順序允許進行修改，唯一不可修改順序的是（6）對仗模組，因為對仗模組是參考其模組的斷詞結果，和詞的對仗關係進行斷詞，所以它一定要在所有的模組最後進行。

- **Rule 2：If 詞牌已知 and 詞牌不是《好事近》、《憶少年》、《醉太平》、《沁園春》、《風流子》then 斷詞順序為 1，2，3，4，5，6**
- **Rule 3：If 詞牌未知 then 斷詞順序為 1，3，4，5**

領字資料庫和對仗資料庫根據詞牌建立，若不知詞牌，就跳過這兩個模組以增加斷詞處理的效能。

此外系統提供使用者自訂斷詞法，可視需求自行決定斷詞順序以及要挑選哪一些模組進行斷詞。唯一要注意的是若使用到對仗模組，則對仗模組一定要在所有模組之後進行。

以下小節將分別說明各斷詞模組的設計方式。

3.2.3. 專有名詞模組

專有名詞指的是屬於人、地、事、物所專用的名稱，比如說何遜（南朝梁詩人）、陽州（地名）、黃鶴樓（物）等等。這些專有名詞若沒有正確識別出來，容易造成語意上的混淆。比如：

[6] 1 代表專有名詞模組；2 代表領字模組；3 代表典故模組；4 代表構詞模組；5 代表節奏斷詞模組；6 為對仗模組。

黃鶴樓頭月午。

－吳文英《水龍吟》

其中的黃鶴樓若切成「黃鶴/樓」可能會被誤解所指的是黃鶴這種生物，而非一棟高樓。

專有名詞模組主要是比對專有名詞詞庫，從宋詞詞句中斷出專有名詞。參考圖4是專有名詞模組的比對流程，首先輸入一闋詞，系統將輸入的詞句切割成 1 至 7 個字的字串組合，共 28 種（即 7+6+5+4+3+2+1）排列組合，再跟詞庫進行比對。將詞句切分成一到七個字的原因是詞的句數通常為三到七字句，八字句以上的句子大多可以事先拆成兩個句子。幾個字要拆成一句，這些相關的資訊都可以從書中直接查詢。

接下來把所有組合的字串和專有名詞資料庫進行比對，若詞庫中收錄相同的專有名詞，則將專有名詞斷出。若比對結果含兩個以上的專有名詞，則採用長詞優先法，取詞長最長者斷出。

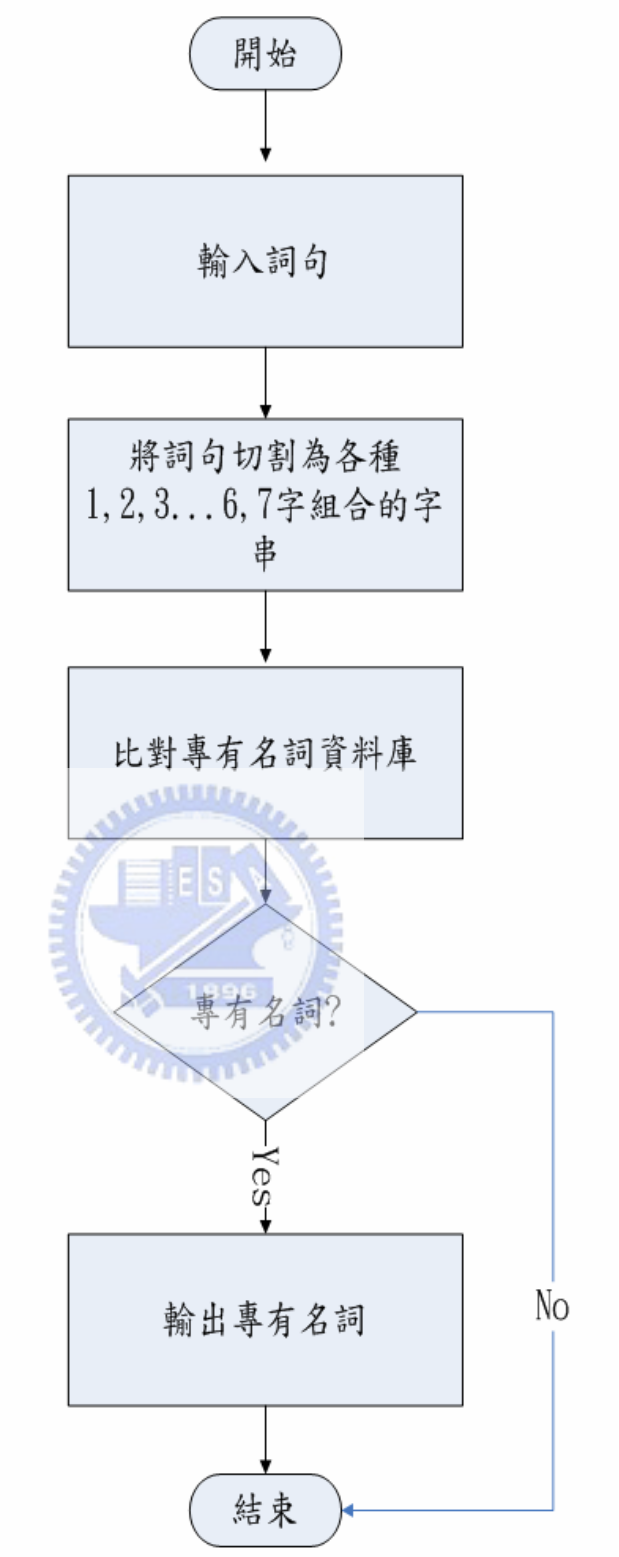

圖 4：專有名詞模組比對流程

專有名詞模組斷字演算法是根據圖 4的流程而撰寫的，參考**演算法 1**。每一個Ci物件代表一闕詞；一闕詞中的每一個句子以Sentence物件表示。SentenceCollection

物件則代表一個句子中所有的詞句所成的集合。Segment物件代表詞中某一個句子的某一個詞彙。SegmentCollection則是一個句子中所有Segment物件所成的集合。

演算法 1：專有名詞模組斷字演算法 （針對一首詞）

Symbol Definition:

Ci: a Ci object

Sentence: a sentence object

SentenceCollection：a sentence collection of a Ci object instance

Segment：a segment object representing a segmented word

SegmentCollection :a collection of all segment object in a Sentence instance

Input:

Ci: a Ci object instance for segmenting

Output:

newCi: a Ci object, it contains all segmented words collection ,and tags with particular parser

Algorithm :

Begin:

for each $Sentence_i$ of a ***Ci*** {

for each $Segment_j$ of $Sentence_i$ {

step1: segment the $Sentence_i$ into { Segment $_0 \cup$ Segment $_1 \ldots \cup$ Segment $_n$} according to the combination of any length of words

step2: if each $Segment_n$ in $Segment_j$ exists in corpus {

step2.1 : extract the longest $Segment_n$ from $Sentence_i$

step2.2 : tag $Segment_n$ with ProperNounParser

/ /ProperNounParser 代表專有名詞模組

step2.3 : create a new *SegmentCollection* for the $Sentence_i$,

create one to many *Segment* objects representing all words in $Sentence_i$

add all *Segment* objects into *SegmentCollection*

```
            }
        }
}
End:
```

此演算法一開始輸入欲進行斷詞的宋詞詞句，經專有名詞資料庫比對之後，視比對結果切分詞句，最後會輸出一個新 Ci 物件。若比對到專有名詞，新 Ci 物件中含專有名詞的 Sentence 之 SegmentCollection 物件必需重建，將句子轉換成一至多個 Segment 物件，以表示此句子是由多個詞彙（Segment 物件）所成的集合。舉例來說，「何遜在揚州」這一句經演算法斷出「何遜」與「揚州」這兩個專有名詞，那麼 SegmentCollection 物件中將包含三個 Segment 物件，一個 Segment 物件代表「何遜」；一個 Segment 物件代表「在」；一個 Segment 物件代表「揚州」。

最後將代表「何遜」「揚州」這兩個專有名詞的 Segment 物件進行標註，代表這兩個詞彙是由專有名詞模組斷出的。

3.3.4. 領字模組

領字又名虛字、襯字、一字豆、一字逗、一字領、領格字，它是宋詞特有的句法與修辭技巧。詞中的領字句，是在特殊的詞句中，以第一個領字帶領本字後面的幾個字，有領下一句，至兩、三句。它通常用來修扮演著語意轉折，讀此字時，須稍作停頓，用以帶動下文。例如一字領三字句：

似黃梁夢，辭丹鳳，明月共，漾孤篷。

－賀鑄《六州歌頭》

其中的「似」是一個領字，領本句「黃梁夢」和其下幾句。

一字領四字句的例子：

更誰家橫笛，吹動濃愁。

－李清照《滿庭霜》

其中的「更」就是一個領字，領本句「誰家橫笛」和下一句「吹動濃愁」。領字句「更」需排除在外，從「誰家」兩字開始計算節拍。節拍指的就是節

奏[7]，因為宋詞始於民間的歌謠，能搭配音樂吟誦，節奏是讓音樂旋律更為優美的必備條件。節奏可以做為宋詞斷詞的識別單位之一，若能正確地將領字識別出來，這一句就可以斷成「更/誰家橫笛」，而不可能斷成「更誰/家橫笛」，如此將有助於提升斷詞的精確度。

領字句在一些詞調中已成定式。如小令中的《好事近》、《憶少年》、《醉太平》等調中上下片的結句，都是領字句式；長調中的《沁園春》、《風流子》，調中的上下片，都有一個領字領四個四字句的[23]情況。由於同一個詞牌領字位置的字，並非每一位詞人在創作時都會遵守這個規則使用領字，因此在比對領字資料庫時，還需要比對詞中標註為領字位置的字是否是領字常用字。領字模組的比對流程參考圖 5。

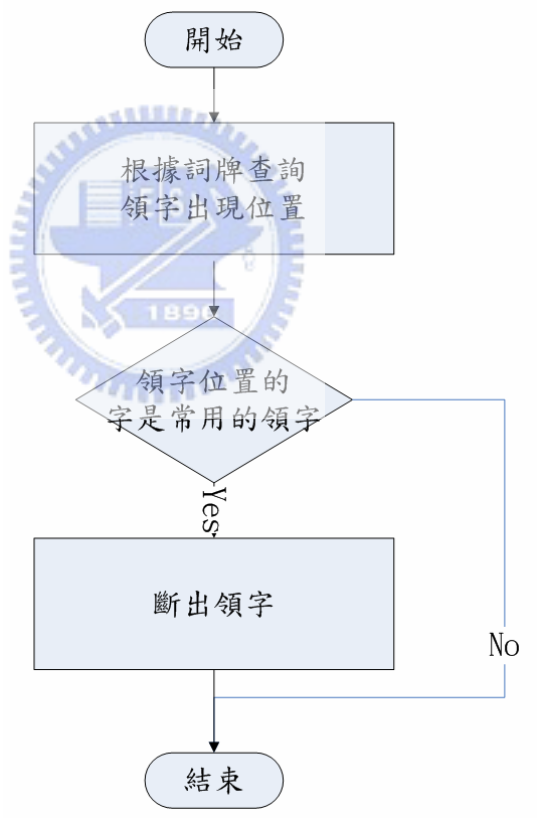

圖 5：領字模組比對流程

3.2.5. 典故模組

詞在宋朝時從原來帶點民歌性質的質樸氣息，漸漸走向述事詠懷，進而愈來愈莊重典雅。為了擴大詞的內容，增強詞的表現力與生命力，詞人使用典故情況

[7] 節奏指語言上的停頓，稱為頓，通常為一個具完整語意的詞彙。

也愈來愈多。若要了解一闋詞的含意，了解詞中包含的典故是相當重要的。不了解典故，可能會把詞的原意誤解而貽笑大方。例如：

楚天千裡清秋，水隨天去秋無際。

遙岑遠目，獻愁供恨，玉簪螺髻。

落日樓頭，斷鴻聲裡，江南游子。

把吳鉤看了，闌干拍遍，無人會，登臨意。

－辛棄疾《水龍吟》

從字面上看把「玉簪」當名詞，它是一種植物。把「玉」當形容詞則「玉簪」可以解釋為玉製的束髮夾。但無論解釋成哪一種，都失去了原意。「玉簪螺髻」指的是山，此典出自於韓愈〈送桂州嚴大夫同用南字〉詩：「江作青羅帶，山如碧玉簪。」和皮日休〈縹緲峰〉詩：「似將青螺髻，撒在明月中」。同樣地，「把吳鉤看了」此句中的「吳鉤」從字面上看指的是吳王闔閭的寶刀，但實際上是引用自杜甫〈後出塞〉詩：「少年別有贈，含笑看吳鉤。」，指有報國封侯的志向和能力[24]。

典故模組的設計方式和專有名詞模組類似，將詞句切成字串之後，跟典故資料庫進行比對，輸出含典故的字詞。參考演算法 2。

演算法 2：典故模組斷字演算法 （針對一首詞）

Symbol Definition:

Ci: a Ci object

Sentence: a sentence object

SentenceCollection：a sentence collection of a Ci object instance

Segment：a segment object representing a segmented word

SegmentCollection :a collection of all Segment object in a Sentence instance

Input:

Ci: a Ci object instance for segmenting

Output:

newCi: a Ci object, it contains all segmented words collection ,and tags with particular parser

Algorithm :

Begin:

for each $Sentence_i$ of a ***Ci*** {

for each $Segment_j$ of $Sentence_i$ {

step1: segment the $Sentence_i$ into { Segment $_0 \cup$ Segment $_1 \ldots \cup$ Segment $_n$} according to the combination of any length of words

step2: if each $Segment_n$ in $Segment_j$ exists in corpus {

step2.1 : extract the longest $Segment_n$ from $Sentence_i$

step2.2 : tag $Segment_n$ with LiteraryQuotationParser

/ / LiteraryQuotationParser 代表典故模組

step2.3 : create a new *SegmentCollection* for the $Sentence_i$,

create one to many *Segment* objects representing all words in $Sentence_i$

add all *Segment* objects into *SegmentCollection*

}

}

}

End:

3.2.6. 構詞模組

以詞典比對法為主的斷詞器面臨的最大問題就是：詞庫中永遠會包含未知詞，不可能將所有的詞都完整地收入詞庫。如此詞庫將會過大，也會降低比對的效能。有些詞具備規律性，可以利用構詞的規則來判斷是否為詞，尤其是定詞、量詞與定量複合詞。

因這些詞無法窮舉，且我們採用的中研院詞庫〔八萬目詞〕中並不收列這些定量複合詞，本研究採取利用構詞模組來處理這些定量複合詞。採用的構詞規則包含三大類：（1）定詞、量詞、定量複合詞，（2）複疊詞，與（3）詞綴、接頭、接尾詞。

(1) 定詞、量詞、定量複合詞：

定量複合詞的組合有無限多種，無法完全收入詞庫，舉例來說：

柔腸一寸愁千縷。

－李清照《點絳脣》

離鉤三寸無生路。

－黃庭堅《漁家傲》

「一寸」與「三寸」其中的「一」是定詞，「寸」是量詞，我們利用構詞規則將定詞結合量詞切分出定量複合詞出來。

由於定量複合詞構詞具規律性，中研院詞庫小組陳克健、黃居仁就曾整理了定詞、量詞與定量複合詞的構詞規則[25]，我們將利用部份的規則進行構詞的判斷。以下就其中的規則舉例說明：

- **詞庫小組複合定詞構詞規則 1：IN1 → NO1*;**

IN1 這條規則的意思是某一個字組（詞彙）是由NO1 中的集合所組成，且NO1 可以有一到多個。NO1 為以下字所成的集合：「一，二，兩，三，四，五，六，七，八，九，十，廿，卅，百，佰，千，仟，萬，億，兆，零，幾，０，１，２，３，４，５，６，７，８，９」[26]。如「一兩」這個字組中的兩個字，都是由NO1 這個集合中的字組成，符合此項構詞規則。「八十九」這三個字，也是從NO1 這個集合中的字組合而成的。從這一條規則中就可以切出前述詞中的「一寸」與「三寸」這兩個複合定詞。

- **詞庫小組複合定詞構詞規則 2： IN → NO2***

這條規則和上一條規則類似，某一個字組是由 NO1 中的集合所組成，NO2 可以有一到多個。NO2 集合中包含的字為：「壹、貳、參、肆、侮、陸、柒、捌、玖、拾、佰、仟、萬、億、兆、零、幾」。

- **詞庫小組定量複合構詞規則 1：DQ2 → Dfa {多，少};**

DQ2 這條規則是由 DQ2 集合中的字，和「多」或「少」這個字組成的。舉例來說，Dfa 所含的字串集合中包含了「很，挺，怪，真，好，極，滿，更，再，頂，

最，太，忒，多，夠，非常，異常，十分，尤其，有點，略為，稍為，比較，不大，過份，過分，這麼，那麼」這幾個字，因此以此規則便可以切出「多少」這個字組，如：

多少春情意。

－李清照《孤雁兒》

- **詞庫小組定量複合構詞規則 10：NOP2 → DESC {半};**

此規則描述詞彙是由 DESC 集合中的任一字和「半」字組合而成的。DESC 的集合中包含「大、小、半」，因此可以切出如「小半」「大半」等詞彙。

- **修改自詞庫小組定量複合構詞規則 13：RD13 → {一} M;**

此規則描述由「一」這個字，和M所組成的複合字。M可能是Nfa、Nfb、Nfc、Nfd、Nfe、Nfg、Nfth、Nfi[26]這幾個集合中的任一個字。舉例來說Nfd集合中含「縷」這個字，那麼這個規則就可以切出「一縷」這個詞彙。

(2) 複疊詞：

複疊詞指的是由重複的字組合而成的字組，如李清照《聲聲慢》一詞中大量使用的疊字：

尋尋覓覓，冷冷清清，悽悽慘慘戚戚。

－李清照《聲聲慢》

以語言學家的觀點而言，詞句的節奏是以每兩個音節（即兩個字）作為一個節奏單位的[19]。因此上述的詞句就不能以四字疊詞[26]的方式斷之，而要斷為「尋尋/覓覓，冷冷/清清，悽悽/慘慘/戚戚」。由於節奏通常是兩個字一單位，在複疊詞的切分方面就變得比較簡單，只要判斷出AA型[8]就可。

(3) 詞綴、接頭、接尾詞：

詞綴是一種附著語素（Bound Morpheme，簡稱BM），這些附著語素必須依附其它的成份，才能出現[16]。根據經濟部中央標準局定義的分詞規範中提及：「附著語素盡量和前後詞合為一個分詞單位」[27]。

[8] AA型：指兩個字相同的複疊字。

附著語素是有獨立意義，但無法獨立扮演一個語法功能的語言成份，因此可以用來當做斷字判斷的依據。例如：

甚霎兒晴，霎兒雨，霎兒風。

－李清照《行香子》

句中的「甚」是領字句，而霎在中研院詞庫小組〔八萬目詞〕中標記為 b（bound），代表附著語素。

參考圖 6的詞綴、接頭、接尾詞斷詞比對流程，首先從詞句中比對詞庫，找出詞綴所在位置，再根據分詞標準[27]，比對此詞綴是為前詞綴或後詞綴，此詞綴和前一個字以及後一個字是否可以合成一個詞。「兒」這個字為多功能後綴，代表可以將「霎兒」合併成一個詞。這三句詞就可以斷成：「甚/霎兒/晴，霎兒/雨，霎兒/風。」。

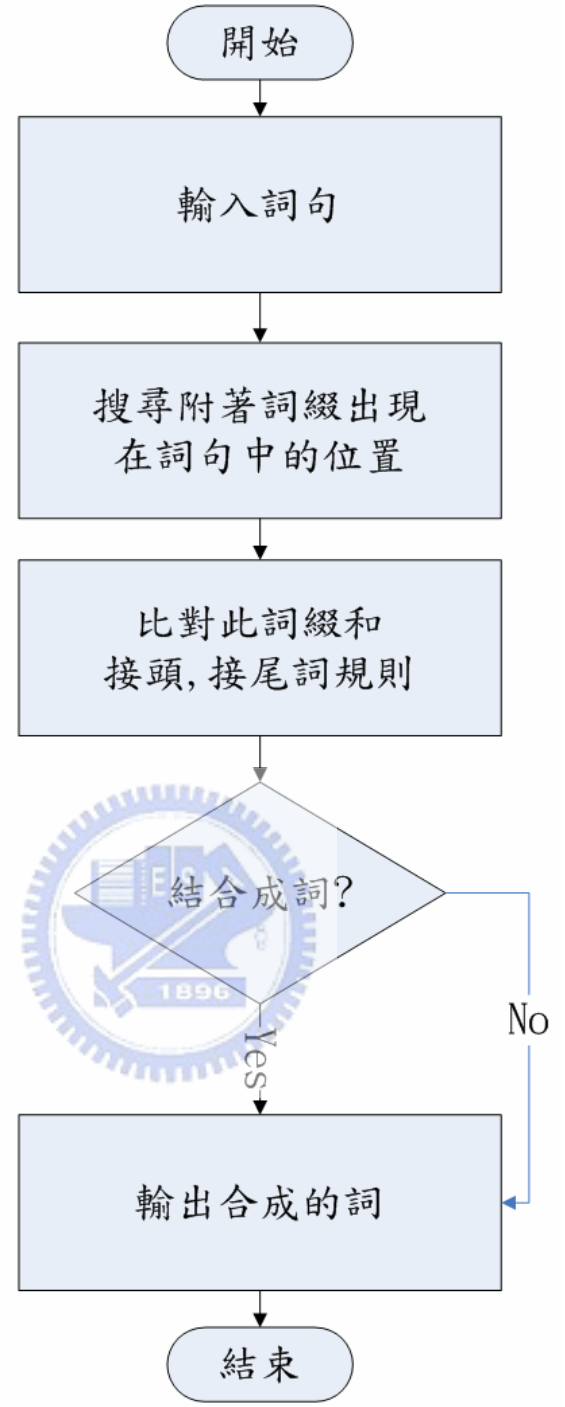

圖 6：詞綴、接頭、接尾詞比對流程

3.2.7. 節奏斷詞模組

節奏是音樂旋律中必備的條件，指語言上的停頓，也稱之為頓。而宋詞為長短句，配合音樂依聲填詞，為了讓詞句吟詠時能有高低起伏、優美悅耳，填詞時非常注重節奏。節奏約分為兩大類：聲律節奏與意義節奏。舉例來說，4 字句的詞句，常切分為（2，2），或（1，3）[28]：

音律節奏： 彩舟/雲淡 （王安石《桂枝香》）

意義節奏：紅蓼花/繁 （秦觀《滿庭芳》）

詞的節奏代表詞的聲律單位，而詞的意義單位常常是和聲律單位結合的很好的[19]。一般來說，意義單位就是一個詞、一個詞組、一個介詞的結構，或一個句子形式，因此節奏點通常就是斷詞的識別記號，也可以幫助我們用來解決詞句中未知詞（Unknown Word）的問題。

詞中三字句至九字句佔的數量最多，我們整理出**表 2**詞的常用節奏[13]。四字句以上，若節奏以 1 開始的多是領字句。而六有些字句、七字句，或十字句可以明確地斷成兩個句子，如六字句可以事先斷成兩個三字句，這些特殊的情況，在詞譜中標註為「豆」以作識別[28]。

表 2：詞句的常用節奏

詞句	節奏
三字句	(1，2) (2，1)
四字句	(2，2) (1，3) (3，1) (1，2，1)
五字句	(2，3) (3，2) (2，2，1) (1，4) (1，2，2)
六字句	(2，4) (4，2) (1，5) (5，1) (2，2，2) (3，3 詞譜標豆)
七字句	(4，3) (2，5) (5，2) (3，4 詞譜標豆) (1，6)
八字句	(3，5) (5，3) (1，7) (2，6) (4，4) (6，2)
九字句	(3，6) (5，4) (6，3) (7，2) (1，8) (2，7) (4，5) (6，3)
十字句	(3，7) (7，3) (4，6) (6，4) (5，5 詞譜標豆)
十一字句	(4，7) (6，5)

八字以上的句子，都可事先切分為兩個讀（按前人說法，整句為句，半句為讀或豆），經由專家的建議，我們從常用的節奏歸納出斷詞的規則為：

(1) 規則 1：四字句或讀，切為（2，2）

(2) 規則 2：五字句或讀，（2，3）

(3) 規則 3：六字句或讀，（2，2，2）

(4) 規則 4：七字句或讀，（2，2，3）

(5) 規則 5：三字句或讀，（1，2）或（2，1）

節奏斷詞模組針對常用的節奏進行斷詞。不過以五字為例，可能切為（2，3）、（3，2）、（2，2，1）、（1，4）、（1，2，2）這幾種組合，如果直接切成（2，3）難免有以偏概全之疑慮，此時就有賴於其它模組的輔助。如：

念武陵春晚

－李清照《鳳凰臺上憶吹簫》

「念」是領字，若領字模組正確切割出此字，那麼這個五字句就可能就是（1，4） 或（1，2，2） 的組合；而「武陵」是專有名詞，同樣地，若專有名詞模組正確地切分出此字組，那麼五字句的組合就為（1，2，2）。如果其它模組都無法輔助進行切分的動作，就參考四字讀斷為（1，2，2）。

針對三字句或讀沒有一定的切分準則，切分的方式首先根據詞庫進行比對，找尋（2，1）或 （1，2） 之中的二字詞在詞庫的詞頻較高者為優先。若無法從詞庫進行比對，則參考詞句是否對仗，再進行切割。

綜合以上所述，參考圖 7是節奏斷詞模組的比對流程。首先輸入詞句，判斷詞句中每個詞彙的長度。若為一或二個句或讀，直接斷出不必再行處理，因為所有詞的詞句中僅有一個字的只有陸游《釵頭鳳》一闕，此外都是以兩個音節為一個單位。若為四、六字句或讀，則按常用規則切為雙字詞。五字句或讀則切為（2，3）；七字句或讀則切為（2，2，3）。而三、五、七字句或讀中可能會包含三字詞，其解法是將三字詞先切為（2，1）與（1，2），再查詢兩個二字詞的詞頻，以詞頻較高者為斷字原則。若無法取得詞頻，則留待對仗模組處理。

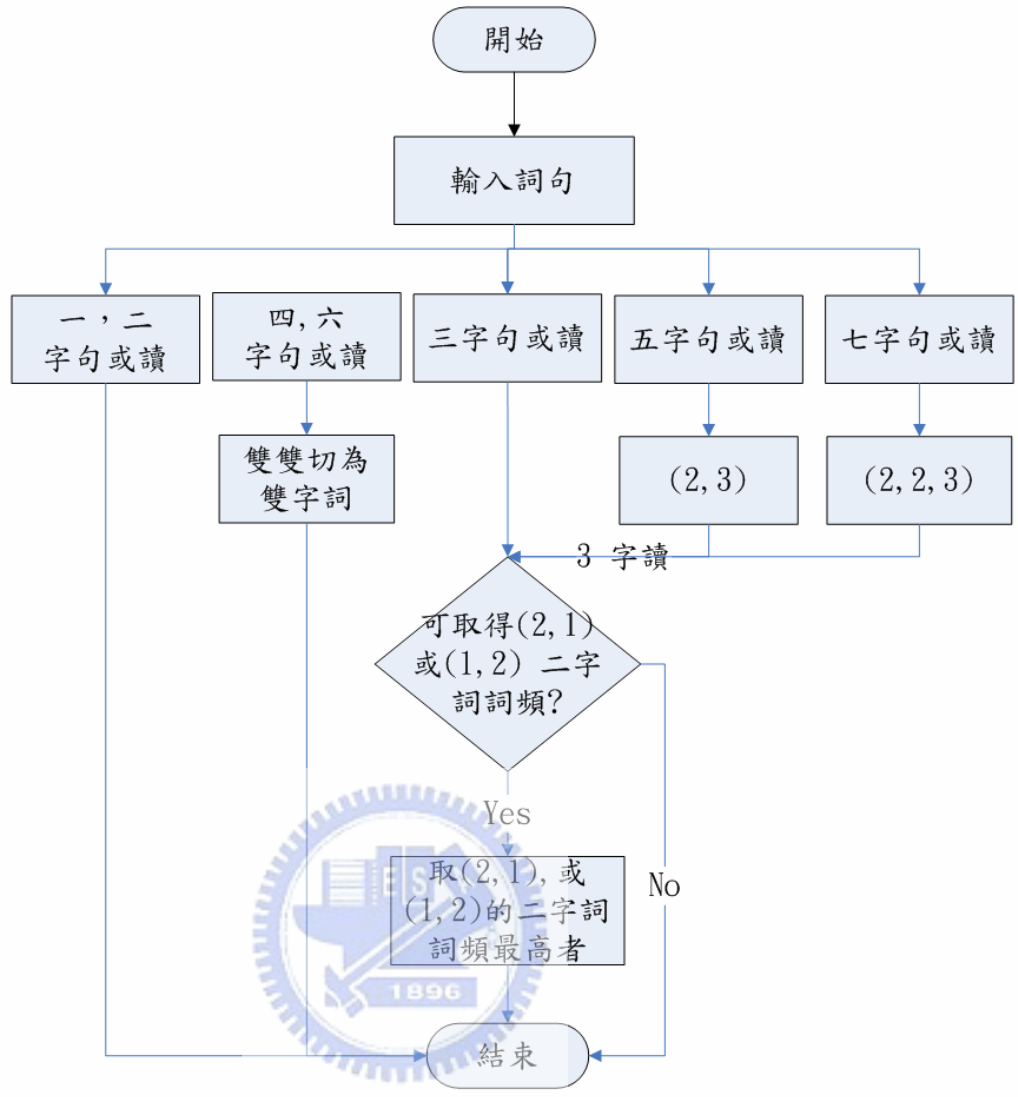

圖 7：節奏斷詞模組流程

根據**圖 7**的流程，定義出節奏斷詞模組的斷字演算法，參考**演算法 3**。

演算法 3：節奏斷詞模組斷字演算法 （針對一首詞）

Symbol Definition:

Ci：a Ci object

Sentence：a sentence object

SentenceCollection：a sentence collection of a Ci object instance

Segment：a segment object representing a segmented word

SegmentCollection :a collection of all Segment object in a Sentence instance

t1：temp variable

t2：temp variable

Input:

Ci: a Ci object instance for segmenting

Output:

newCi: a Ci object, contains all segmented words collection ,and tags with particular parser

Algorithm :

Begin:

for each $Sentence_i$ of a ***Ci*** {

for each $Segment_j$ of $Sentence_i$ {

switch length of $Segment_j$ **{**

case 1: goto **End:**

case 2: goto **End:**

case 4: slice $Segment_j$ into two subsegments with length two.

case 6: slice $Segment_j$ into three subsegments with length two.

case 5: slice $Segment_j$ into two subsegments with length two and three for each.

goto **case 3**

case 7: slice $Segment_j$ into three subsegments with length two ,two, and three for each.

goto **case 3**

case 3:

search the frequency of first two words of $Segment_j$ from corpus into *t1*

search the frequency of last two words of $Segment_j$ from corpus into *t2*

if ***t1*** **>** **t2 {**

step 1: extract first two words from $Segment_j$ into $Segment_k$

step 2: tag $Segment_k$ with WordParser

/ / WordParser代表節奏斷詞模組

step 3: create a new *SegmentCollection* for the $Sentence_i$,

step 4: create one to many *Segment* objects representing a word in $Sentence_i$

add all *Segment* object into *SegmentCollection*

}

else {

step 1：extract last two words $Segment_j$ into $Segment_k$

step 2: tag $Segment_k$ with WordParser

/ / WordParser代表節奏斷詞模組

step 3：create a new *SegmentCollection* for the $Sentence_i$,

step 4：create one to many *Segment* objects representing a word in $Sentence_i$

add all *Segment* object into *SegmentCollection*

}

}

}

End:

3.2.8. 對仗模組

對仗是中國古典詩詞中的重要手段之一，但詞和詩不太相同，詞句並不要求一定要對仗，但是詞的某一些句子中，只要相連兩句字數相同，經常用對仗。許多的作者經常利用這樣的句式來加強詞的藝術效果。如：

秋已盡，日猶長。

－李清照《鷓鴣天》

舉例來說《鷓鴣天》詞牌的上闕第三四句、下闕第一二句一般要求對仗，若已斷出「秋/已盡」，我們就可以根據對仗斷出「日/猶長」。若沒有對仗資訊，則跳過此步驟，不做任何處理。

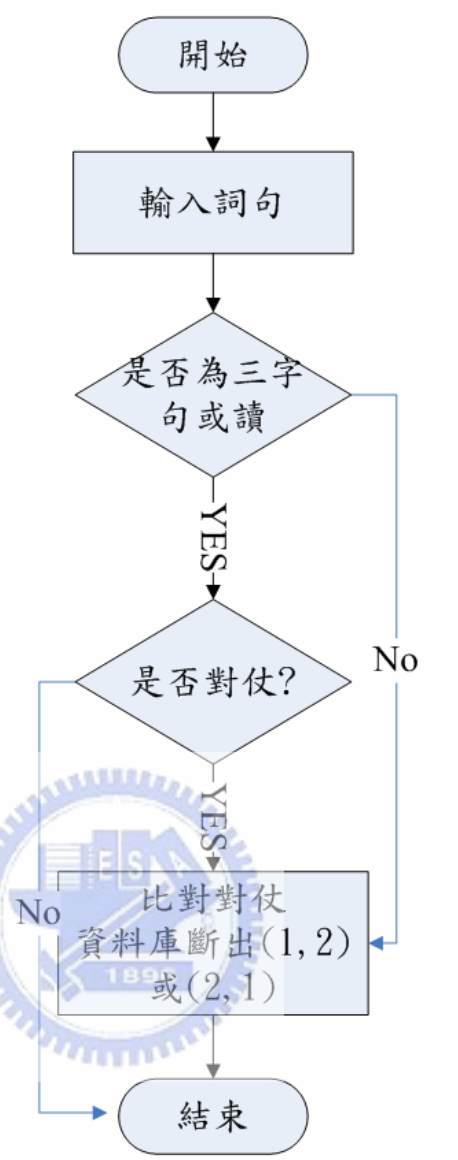

圖 8：對仗模組比對流程

參考圖 8是對仗模組比對流程，此模組只針對三字句或讀進行處理。若詞句不為三個字，則直接結束。若為三字句或讀，則根據詞牌查詢資料庫中是否包含此詞牌的對仗資料。若資料庫含對仗資訊，則根據對仗切割三字詞，若不含對仗資訊，則直接結束。演算法請參考**演算法 4**。

演算法 4：對仗模組斷字演算法 （針對一首詞）

Symbol Definition:

Ci: a Ci object

Sentence: a sentence object

SentenceCollection：a sentence collection of a Ci object instance

Segment：a segment object representing a segmented word

SegmentCollection :a collection of all Segment object in a Sentence instance

Input:

Ci: a Ci object instance for segmenting

Output:

newCi: a Ci object, it contains all segmented words collection ,and tags with particular parser

Algorithm :

Begin:

for each $Sentence_i$ of a ***Ci*** {

for each $Segment_j$ of $Sentence_i$ {

if length of $Segment_j$ equals to 3 {

step 1: extract the $Segment_k$ from $Segment_j$ *according to the Pair database*

step 2: tag $Segment_k$ with PairParser

/ / PairParser 代表對仗模組

step 2: create a new *SegmentCollection* for the $Sentence_i$,

create one to many *Segment* objects representing a word in $Sentence_i$

add all *Segment* object into *SegmentCollection*

}

}

}

End:

3.3. 解歧義

中文斷詞面臨的一個難解的問題便是歧義的處理。比如說：

寒日蕭蕭上鎖窗

－李清照《鷓鴣天》

其中的「上鎖窗」三個字要切分為「上鎖/窗」還是「上/鎖窗」?我們所採取的方式是利用詞頻針對三字句來解歧義。

關於詞頻的取得方式比較特殊，由於中研院詞庫〔八萬目詞〕中有記錄詞頻的資訊，而我們自行搜集的詞彙有些無詞頻可參考。為解決無詞頻問題，我們以中研院詞庫〔八萬目詞〕詞頻資訊為基礎，每次斷詞器斷出的詞彙，經由專家校正之後，將之儲存起來，系統會自動累計所有詞彙的詞頻資訊。

解歧義的動作則分為兩階段，首先利用雙向最大匹配法找尋歧義（Max Match），以上述的例子而言，步驟為：

(1) 使用正向最大匹配法比對：由左至右根據詞庫比對，得到「上鎖/窗」，屬於 AB/C 型。

(2) 使用反向最大匹配法比對：由右至左根據詞庫比對，得到「上/鎖窗」，屬於 A/BC 型。

(3) 接著針對 AB/C ， A/BC 找尋這兩個類型的詞頻，若詞頻庫中無此字組，則詞頻為 0，取詞頻累加後最高者為解歧義的手段：

AB/C　　X = Freq（AB）+ Freq（C）

A/BC　　Y = Freq（A）+ Freq（BC）

(4) 若詞頻相等，則從 AB/C、A/BC 中亂數挑選一種斷詞方式。

3.4 宋詞詞彙本體論

在這個小節中，我們參考 TOVE 本體論工程的建置步驟，建立宋詞的詞彙本體知識，以下說明建立的步驟。

3.4.1. 本體論

「知識」這個名詞是一個抽象的概念，很難以用語言或文字來具體的陳述。尤其是在資訊科技融入生活的時代，要讓電腦能夠了解所謂的知識，進而能幫助我們進行學習更是一件困難的事情。為了讓電腦能夠了解這些知識，必需使用一些特殊的表達方式。

「本體論」（Ontology）起源於哲學，一開始是用於探討「存在（being）」。常使用樹狀結構及關聯的方式來描述事物，並描述事物之間的規則，以便讓知識更容義分享和共用。在 90 年代，本體論一詞被應用在人工智慧的領域之中，描述

某個特定領域的知識，以及該領域相關的各式物件、物件屬性與物件彼此之間構成要素和關係。它提供了一個讓人與人之間及不同的應用系統之間，彼此可以分享、溝通的一個知識交換的媒介。

本體論的研究主要方為兩個方向，一是針對特定領域建置的本體論；二是研究本體論的建構與表達方式。

3.4.2. RDF（S）

W3C提出了RDF（ Resource Description Format，簡稱RDF）[5]標準用來表達資源之間的關聯，希望能夠利用此一共通的資料通訊協定標準，讓網路環境中不同系統能夠互通資料。它是一個可以攜帶多種不同的元資料（Metadata）透過網路傳送與交換的工具。RDF為資源這種資料模型提供了一種簡單的語義，並以XML語法來表示之。

RDF的基本資料模式包含了三種物件型態：資源（Resource）、特性（Property）與敘述式（Statement），參考圖 9。

- 資源：RDF 能夠描述的任何事物都稱之為資源。如一個人、一本書。
- 特性：描述資源的特徵或屬性。特性可以用來定義值，所描述的資源型態，以及資源和特性之間的關係。
- 敘述式：以特性和特性的值來描述資源的描述稱敘述式。一個敘述式中包含主詞（Subject）、述詞（Predicate）和受詞（Object）三大部份，分別表示資源、特性和特性值。

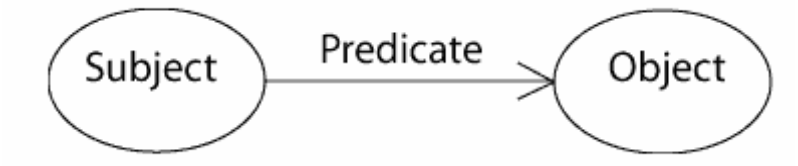

圖 9：RDF

如果要使用RDF來表達某本書的擁有者是Mary，則可以使用圖 10的有向圖來表達。

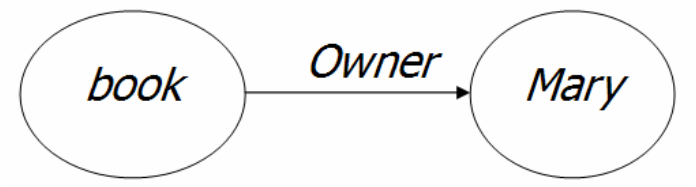

圖 10：RDF 示例

RDF schema[29]是一個檔案，可以讓使用者在其中定義一些資源，以及這些資源可以有哪一些屬性。由於RDF這種資料模型，主要是用來描述物件（資源） 與彼此之間的關係，以及RDF（S）表達的方式不夠豐富，隨後便發展了DAML+OIL與OWL以擴展資源的描述方式。

3.4.3. OWL

OWL[30]繼RDF成為W3C所推薦的Web本體描述語言，結合了DAML與OIL[31]的特色，能以XML、RDF語法透過描述邏輯來描述語意，參考圖 11為本體論語言階層之間的關係。

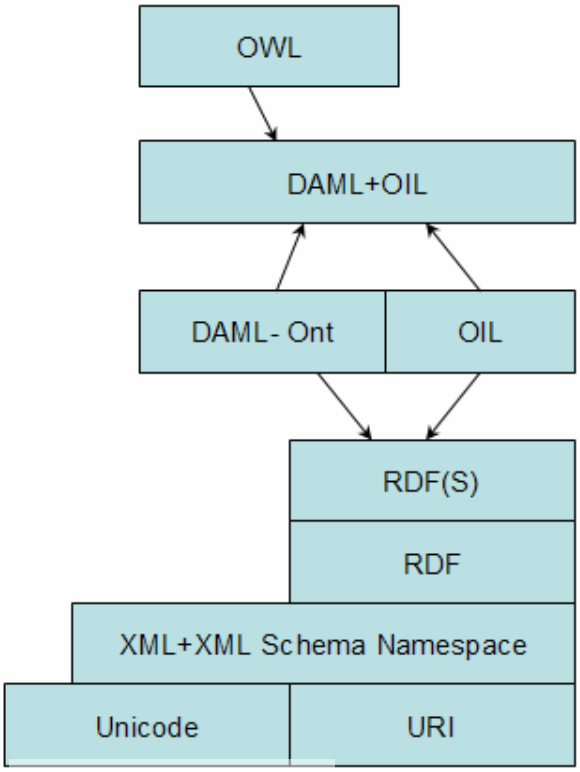

圖 11：本體論語言階層

（資料來源：Deborah. L.，2004）

OWL 新增了許多的語彙來描述屬性與類別：其他之中（among others），類別之間的關係（relations between classes 如：解體 disjointness），基數（cardinality 如：剛好一個 exactly one），相等（equality），屬性的多元型態（richer typing of properties），屬性的特徵（characteristics of properties 例如：對稱 symmetry），與列舉類別（enumerated classes）。

3.4.4. 建置宋詞詞彙本體論

我們根據 TOVE 本體建立的方法論，建立了宋詞詞彙本體論，步驟如下：

第一步：確定本體論的領域以及研究的範圍

我們以宋詞為本體論的建置範圍，其目的地是希望建立詞彙與詞彙之間的前、後關聯性，以及詞彙的同義關係，和詞彙的近義關係與反義詞。透過此本體論，可以了解某一詞彙在語意上所屬的概念階層、比如說「玉簪」是屬於植物還是「器物」，還是兩者皆是。此外，我們也記錄了此詞彙曾被使用在哪一個詞牌

上，以及詞的作者，和使用的時空背景、此詞彙同義、反義或近義的詞彙等等。這些資訊就成為未來在研究或學習詞時的參考，也可以從本體知識中獲得詞彙相關的語意資訊。

在這一個步驟中，整理以出下資訊：

- 本體論涵蓋的範圍?

  宋詞使用到的詞彙，某詞彙和其它詞彙是否有前後關係。如：「紅藕香殘玉簟秋」一句的「玉簟」出現在「秋」之前；「紅藕」之後接「香殘」這個詞彙。使用此本體論，可以協助了解詞彙的使用，做為將來個人學習或機器學習的參考。

- 是否可沿用已存在的本體論?

  IEEE 標準上層知識本體工作小組建置了 SUMO（Suggested Upper Merged Ontology），作為建議的上層共用知識本體。由於本研究其中一個目的是要描述同義詞的資訊，而同義詞詞林已對同義詞進行了相當完善的分類，因此本研究以同義詞詞林的概念階層為主，再參考 SUMO 本體來建立宋詞詞彙本體論。

第二步：歸納出本體論所須回答的問題

在這一個步驟中，整理以出下資訊：

- 本體論需提供哪些解答?

  哪位作者曾使用此詞彙?哪個詞牌中曾使用此詞彙?此詞彙的同義字、近義字、反義字是哪些?此詞彙和哪個詞彙曾有對仗關係?此詞彙所屬的語意概念階層為何?

- 本體論的使用的對象?

  任何對於學習詞的使用者都可以使用這個建置出來的詞彙本體論，透過表達知識的 OWL 語言進行描述，也可以幫助在不同系統與平台之間交換知識，以方便讓機器進行自動學習。

第三步：列舉領域中的詞彙

本研究將從宋詞斷詞器斷出，且經專家校正的詞整理成為本體論的候選詞彙。表 2、3 是部份的詞彙以及定義的列表。

表 3：一些類別(Class)的示例

類別 (Class)	定義
人	與人相關的概念之父類別
天體	與天空、星球相關概念之父類別
植物	物的子類別
動物	物的子類別
地貌	物的子類別
氣象	物的子類別
抽象事物	如外貌、意識、社會、經濟
特徵	描述外形、表象、顏色、性質
菊花	一種植物
蘭花	一種植物
薔薇	一種植物
織女星	天體的子類別
北極星	天體的子類別

表 4：屬性（Property）

	Property 類型	定義
hasCiPaiName	Datatype Property	詞牌的名稱
hasAuthor	Datatype Property	作者的名稱
hasFrequency	Datatype Property	詞頻
hasWordType	Datatype Property	詞類
hasNotation	Datatype Property	詞譜中標註的平仄資訊
hasPrevious	Object Property	前一個詞彙
hasNext	Object Property	後一個詞彙
hasAntonym	Object Property	反義詞
hasNearSynonym	Object Property	近義詞
hasPair	Object Property	對仗詞
hasSynonym	Object Property	詞彙的同義詞

以下說明這些屬性：

- hasCiPaiName（詞牌）

  記錄詞彙在哪個詞牌中曾經使用過。一個詞彙可以在多個詞牌中使用。

- hasAuthor（作者的名稱）

  記錄哪個作者曾經使用此詞彙。一個詞彙可以關聯到多個作者。

- hasFrequency（詞頻）

  詞彙的使用頻率。

- hasWordType（詞類）

  記錄詞彙的詞性，如動詞、名詞、複合詞等等。

- hasNotation（平仄）

  標示此詞彙所在的詞譜中的讀音資訊，如平平、平仄，可以有多種組合，代表一字多音。

- hasPrevious（前一個詞彙）

  描述此詞彙的前方曾經出現的詞彙，對照到多個詞彙。

- hasNext （後一個詞彙）

  描述此詞彙的後方曾經出現的詞彙，對照到多個詞彙。

- hasAntonym（反義詞）

  指和詞彙相反或相對的詞，反義詞之間必須具有同一性，以便表現出相反或相對的情形，如：男女是在性別上有語義的對立；長短在長度上對立；老幼則是年齡上的對立。互補詞也可以視為反義詞，譬如：生與死，非生即死，沒有兩者同時存在的情況。具有對立關係的也可以稱為反義詞，譬如：買賣，售貨員對購物者而言說是賣，購物者對售貨員而言是買。

  一個詞可能會有一個以上的反義詞，如「大」可對應到「細」、「小」；「生」可對應「死」、「熟」。

- hasNearSynonym（近義詞）

  對於描述一件具體事例或一個概念，在語意相近的語詞稱為近義詞，妨礙、妨害這兩個詞，在意義上接近，但在詞義有輕重之別，「妨礙」的詞義比「妨害」較輕。同樣的「侵佔」、「侵犯」、「侵略」這三個詞彙也有近義關係，但顯然「侵犯」比「侵佔」層度上稍為嚴重，而「侵略」又比「侵犯」這個字眼互為強烈。

- hasPair（詞彙對仗詞）

記錄曾和此詞彙成對仗一起使用的詞。晏幾道《更漏子》中有一句「柳絲長，桃葉小」，「柳絲」的對仗詞便是「桃葉」，「長」的對仗詞便是「小」。

- hasSynonym （同義詞）

在意義相同的詞彙稱為同義詞。比如說針對「植物」這個語意概念而言，「梅花」、「菊花」、「木蓮」、「木樨」、「薔薇」、「牡丹」等詞彙都是屬於同義詞；而對於「表情」這個概念來說，「強笑」、「竊笑」、「熱淚盈眶」都是屬於同義詞詞林。

第四步：定義領域中的類別，以及階層關係

本研究為了描述詞彙的同義資訊，以同義詞詞林的階層架構為主體，再參考SUMO自行建立本體論的階層架構，參考圖 12。

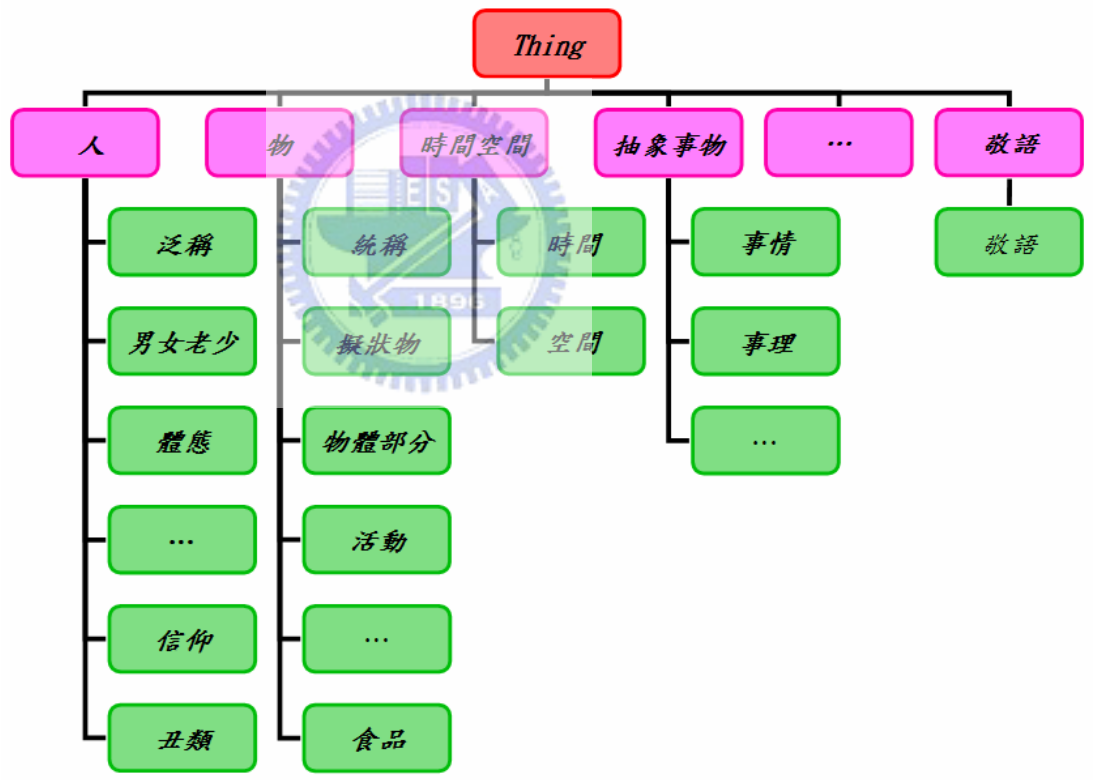

圖 12：宋詞詞彙本體論階層

第五步：設定定理與屬性

為了要描述資源的關係，可以利用本體論的定理（Axiom）進行限制。比如說「秋」一詞在李清照《一剪梅》詞牌中曾出現：「紅藕香殘玉簟秋。」，也在蘇易簡《滿江紅》詞牌中出現：「秋知否」，這代表 hasPrevious 和 hasNext 兩個限制條件中可以使用 hasPrevious>=1 與 hasNext>=1 加以限制。同樣地，hasSynonym

也可以根據需求設定為 hasSynonym >= 0，代表同義字可以有多個或沒有同義字，如此將來就可以協助我們了解某詞彙的替代詞彙。

接下來是設定資料的屬性，屬性包含兩類，分別是資料型態屬性（data property）與物件屬性（object property）。我們使用的資料型態屬性包含以下資訊：

- 領域 （Domain）：限制屬性可以使用的類別，如「吃動物」這個屬性限制只有肉食動物可以使用，而抽象事物如空氣和水就不能夠使用。
- 範圍（Range）：有點類似程式設計領域中變數的型別，包含 any、string、integer、boolean、float、symbol。

以「梅花」這個詞彙為例，其資料型態屬性參考**表 5**。

表 5：資料型態屬性

屬性名稱	**領域（Domain）**	**範圍（Range）**
hasCiPaiName	owl:Thing	String
hasAuthor	owl:Thing	String
hasFrequency	owl: Thing	Integer
hasWordType	owl:Thing	String
hasNotation	owl:Thing	String

其中「owl:Thing」代表的含意是本體中所有類別的聯集。這也就是說，所有在本體中的詞彙不管分在哪一個概念階層（類別）都有 hasCiPaiName 、hasAuthor、hasFrequency、hasWordType、hasNotation 的屬性，以描述詞彙的詞牌、作者、詞頻、詞類、平仄資訊。

我們使用的物件屬性包含了以下資訊：

- 領域（Domain）：限制哪些類別可以使用。
- 範圍（Range）：使用物件連結的方式關聯到其它類別。

參考**表 6**為「梅花」詞彙的物件屬性示例。

表 6：梅花物件屬性

屬性名稱	領域（Domain）	範圍（Range）
hasPrevious	owl:Thing	owl:Thing
hasNext	owl:Thing	owl:Thing
hasSynonym	owl:Thing	植物
hasAntonym	owl:Thing	動物
hasNearSynonym	owl:Thing	植物
hasPair	owl:Thing	owl:Thing

第六步：建立及表達本體知識

這一步驟將根據第三步驟的分類，將宋詞的詞彙加到本體之中。在建置過程中，為避免大量的人工處理，我們透過自行設計的語意編輯工具，將本體直接匯出成 OWL。

首先，使用者可以將斷詞完成的結果，匯入語意編輯工具，進行概念階層，以及相關屬性的編輯和設定。為了讓使用者在此階段建立的資料能夠重複利用，又不希望使用者需從無到有建立本體，此工具提供一個功能，能夠自動根據詞彙語意相關的資訊，產生表達本體的 OWL 文件。到此階段，一個描述宋詞詞句中語意的本體就建置完成了。

第四章 實驗成果

本章節藉由一些實驗來評估斷詞的成果，以及實驗成果的分析。最後將展示本體論的實作與建置的成果。

4.1. 斷詞評估指標

在評估一個斷詞系統的運作情況時，最常用到兩個指標：召回率與精確度。這兩者的定義如下 ：

召回率 （Recall） = N3/N1 （1）

精確度 （Precision） = N3/N2 （2）

N1：正確的詞彙數目

N2：斷詞器辨認出的詞數

N3：斷詞器正確辨認出的詞數

召回率是指斷詞器正確辨認出的中文詞彙數目占正確的詞彙數目之比率，精確度是指在所有斷出的中文詞彙數目當中，斷詞器正確辨認出的中文詞彙數目所占的比率。當召回率與精確度的值愈高時，表示斷詞的品質愈好。通常召回率和精確度不能兩全其美，若召回率上升，則精確度就會下降。這是因為在一次的斷詞程序中，N1 的數目是固定不變的，為了要使召回率增高，代表斷出的詞數 N2 必需要增多。一旦 N2 變多，就（2）而言，代表分母變大了，因此反而會使精確度下降。

在目前資訊擷取（Information Retrieval）的領域中，也常使用召回率和精確度來判斷取得的資訊是否有意義，此外，還會利用召回率與精確度的調和平均數（Harmonic mean）來判斷資訊的有效性（Effectiveness）定義如下：

效度 （Effectiveness） = 2PR / (P+R) （3）

R：召回率

P：精確度

效度的數值愈高，代表取得的資訊愈有意義。在後續的章節中，我們以宋朝李清照所寫的五十首詞為測試資料，以召回率、精確度和效度三個指標來評估斷詞系統的運作。

4.2. 系統實作架構圖

本論文實作了三個系統，參考圖 13分別為：

- 宋詞斷詞器：針對一至多闋詞進行斷詞處理。
- 語意編輯工具：編輯詞彙的語意，如詞彙的同義詞、反義詞資訊。
- 絕妙好詞網站：提供使用者透過網頁查詢詞與詞彙的相關資訊。

圖 13：系統實作架構圖

「宋詞斷詞器」把斷完的詞之詞彙提供給「語意編輯工具」，以編輯詞彙的相關語意資訊，這些資訊又提供給「絕妙好詞網站」，讓使用者從網頁進行檢索。

4.3. 宋詞斷詞器實作

4.3.1. 物件導向的 Ci 物件模型

以物件導向的方式進行開發的好處有兩個：一為容易重複使用，一為容易維護。因此，為了減少後續維護的成本，設計宋詞斷詞器時，我們採用物件導向的方式設計一個代表詞（Ci）的物件模型。

參考圖 14，物件模型中包含三個主要的類別：Ci、Sentence、Segment類別。SentenceCollection代表一到多個Sentence所成的集合，SegmentCollection則代表一到多個Segment所成的集合。

在實際運作的過程中，根據物件導向所定義的類別會轉換成物件實體（Object Instance），由系統配置相關的記憶體。這也就是說，每一個 Ci 類別會轉換成一個 Ci 物件，代表一闋詞。一闋詞中的每一個句子會自動轉換成 Sentence 物件來表示。句子中的每一個詞彙則自動地轉換成 Segment 物件。

針對這些類別說明如下：

- Ci 類別：代表一闋詞。除了記錄詞牌（Ci Pai）資訊之外，還包含一個 Sentences 屬性（Property），這個屬性代表詞中所有句子（Sentence 物件）的集合，因此它的資料型別是一個 SentenceCollection 物件。
- SentenceCollection 類別：詞句所成的集合。包含一個 Count 屬性，記錄了詞中句子的數目。在實際執行時，每一個句子將自動變成一個 Sentence 物件。
- Sentence 類別：代表一個句子。它包含 SentenceNo 與 Segments 屬性。SentenceNo 記錄詞句的號碼，Segments 屬性的資料型別為 SegmentCollection，代表某一個句子中所有詞彙所成的集合，也就是指 Segment 物件所成的集合。
- SegmentCollection 類別：代表一個句子中詞彙所成的集合。包含 Count 屬性，記錄一個句子中詞彙的數目。
- Segment 類別：代表一個詞彙，它包含 Word 屬性，記錄詞彙。

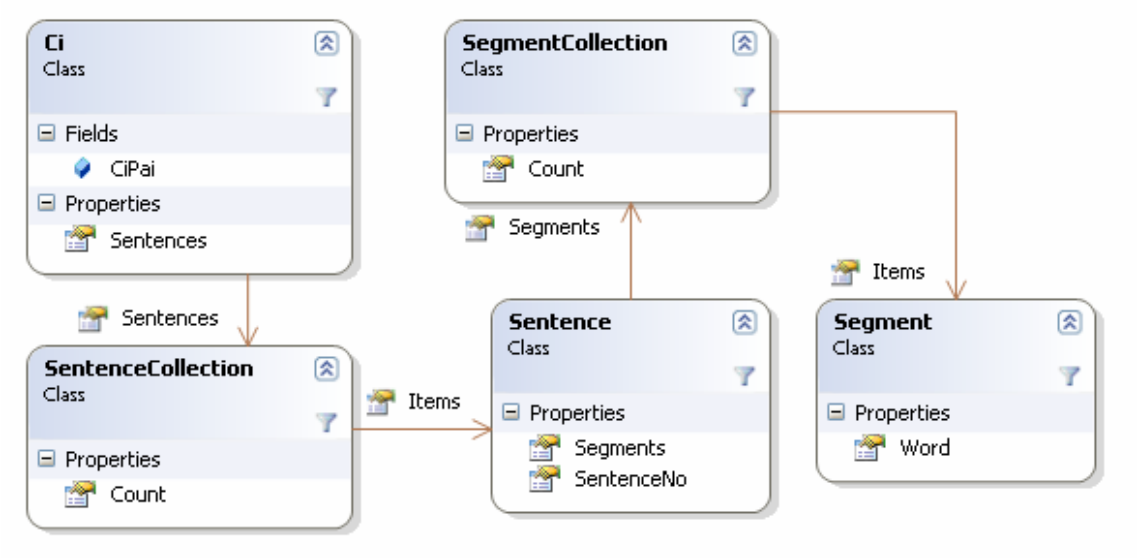

圖 14：Ci 物件模型類別圖（Class Diagram）

4.3.2. 宋詞斷詞器

為了進行斷詞測試，我們實作了宋詞斷詞器，可以讓使用者選擇要斷一或多闕詞。參考圖 15，斷詞器左方為欲斷的詞之資訊，中間為斷完的結果，右方是每一個句子正確識別出的詞彙數目、精確度與召回率。為了避免浪費大量人力進行召回率與精確度的計算，只要將詞的正確斷法輸入到資料庫，斷詞器會自動比對斷詞的結果，計算出召回率與精確度。

預設各斷詞模組的使用順序是根據規則自動決定，但系統設計上保留了彈性，使用者也可以自行設定斷詞模組的順序，或勾選想要使用的模組。

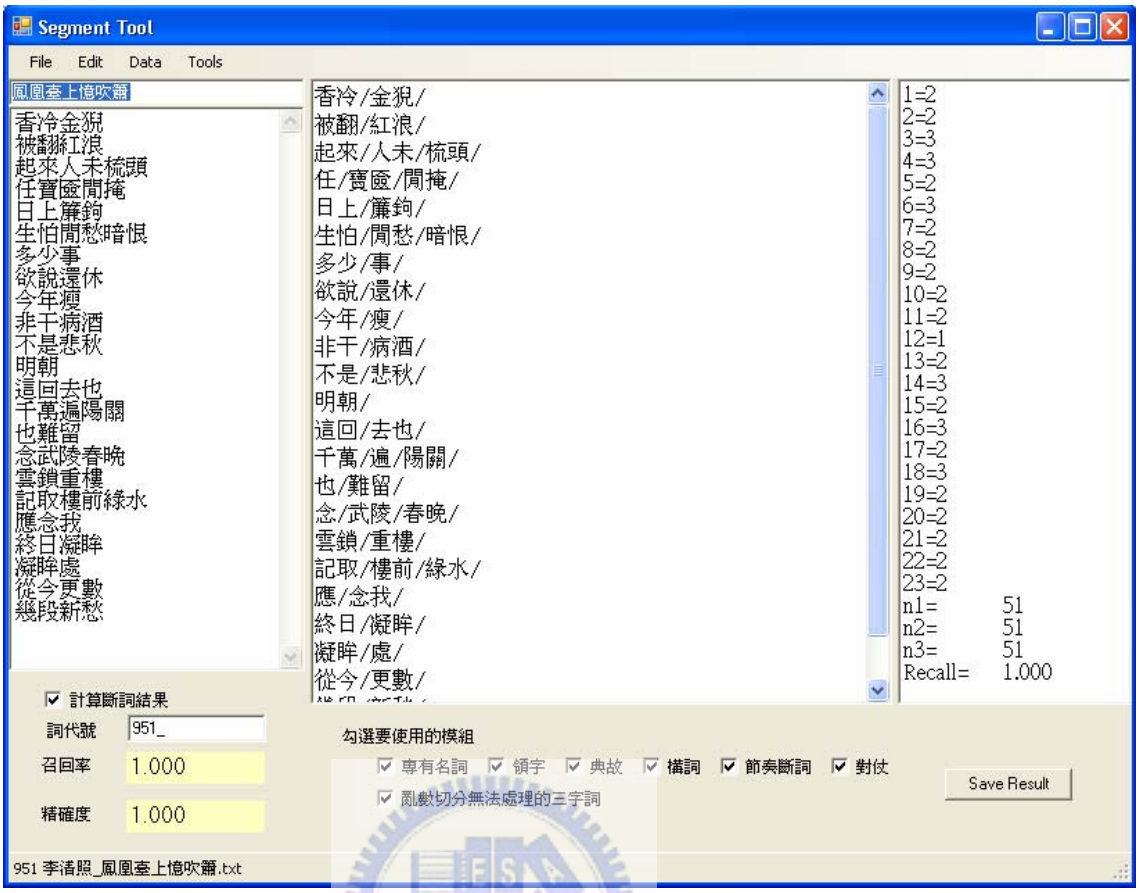

圖 15：宋詞斷詞器

由於斷詞的結果不一定百分之一百的正確，這時就需要專家的介入。專家可以利用斷詞工具修訂斷詞完的結果、將辨識出的新詞、專有名詞、典故資料新增到詞庫當中。必要時可以並把斷詞結果匯出成XML，供其它程式參考或使用。參考圖 16為匯出的XML檔案。

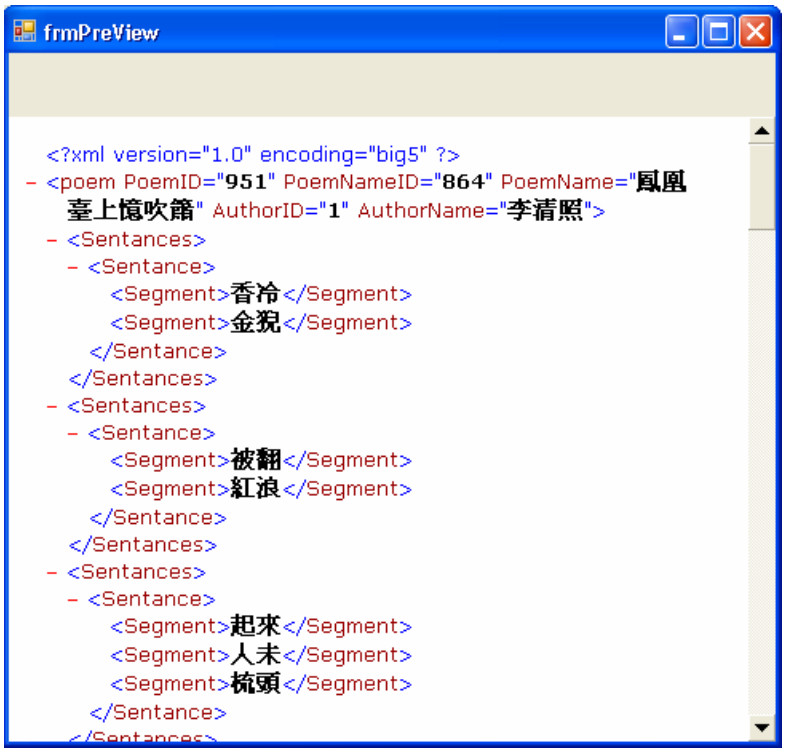

圖 16：將斷詞結果匯出成 XML

4.4. 斷詞實驗

這個小節我們將透過宋詞斷詞器進行斷詞的實驗，評估各斷詞模組的必要性，並分析斷詞過程中影響召回率與精確度的原因。

宋詞斷詞器中包含了六大斷詞模組，分別是：

（1）專有名詞模組：切分詞句中的專有名詞。

（2）領字模組：切分出領字（虛字）。

（3）典故模組：切分詞句中用典資訊。

（4）構詞模組：切分出定詞、量詞、定量複合詞、複疊詞、詞綴。

（5）節奏斷詞模組：根據宋詞節奏（句法）切分詞句。

（6）對仗模組：根據詞句對仗，切分三字詞。

後續的小節，我們設計了幾個實驗，以評估這些模組對於斷詞的貢獻度。

4.4.1. 僅以詞庫斷詞

在這個斷詞的實驗中，採用中央研究院詞庫〔八萬目詞〕再整合實驗室搜集的詞彙而成的詞庫（簡稱 KDE 詞庫）進行測試，目的在評估此詞庫對於斷詞系統的貢獻度。

採取的比對方式是正向最大匹配法（Maximum Matching method），將詞句中的字由左至右切割為一、二、三字詞，一一和詞庫中的詞進行比對，若比對到相等的字組便將此字組切割出來。

表 7：僅使用詞庫斷詞的結果

	中央研究院詞庫〔八萬目詞〕	中央研究院詞庫〔八萬目詞〕+ KDE 詞庫
召回率（Recall）	39.7%	69.7%
精確度（Precision）	58.4%	79%
效度（Effectiveness）	47%	74%

參考**表 7**，從召回率與精確度來看，單純使用中央研究院詞庫〔八萬目詞〕斷詞的效果並不是很好。這樣地結果是可以事先預測得知的，因為中央研究院詞庫雖然收錄了一些專有名詞、古漢語詞語，但仍是以語體文（白話文）為主，對於韻文詞彙的收錄較少，所以無法正確斷出一些詞語。舉例來說，以下詞彙就無法切分出來：何遜（人名）、孫壽（人名）、韓令（人名）、綠蟻（酒）等等。

而使用中央研究院詞庫和實驗室收集而成的 KDE 詞庫進行斷詞，其召回率和精確度有明顯的提升，但在數量上仍不能滿足需求，詞庫無法收錄所有的中文詞彙，因而無法處理未知詞問題。

4.4.2. 節奏斷詞模組實驗

這個實驗只採用節奏斷詞模組，再搭配KDE詞庫進行斷詞測試，主要是驗證利用宋詞節奏斷詞的可行性。由於三字句或讀[9]按慣例常切分為(2,1)或(1,2)，因此只採用節奏斷詞模組無法判別要如何進行切分。針對三字詞的處理，本實驗採用亂數方式切為（2，1）或（1，2）的策略。以李清照五十首詞為測試資料，每首詞進行十回合測試，再求十次平均的召回率與精確度做為斷詞結果。

此次實驗，參考**表 8**召回率和精確度分別達到 88%，88.3%，此外資料的效度也達到 88%，這代表的含意是宋詞的節奏（句法）為斷詞的識別項這個假設的可信度很高。

[9] 一句為句，半句為讀。

表 8：僅使用節奏斷詞的結果

	百分比
召回率（Recall）	88%
精確度（Precision）	88.3%
效度（Effectiveness）	88%

但只考慮到通用的節奏進行斷詞，會讓一些可識別的詞彙被切分為二，以致造成召回率與精確度下降。舉例來說，五字句的節奏可能為（2，3）（3，2）（2，2，1）（1，4）（1，2，2）等，但最常切分為（2，3），只要詞中含領字，單一首詞的精確度就變低：

錯誤：任寶/匜閒/掩　（系統切分的結果）

正確：任/寶匜/閒掩

－李清照《鳳凰臺上憶吹簫》

若僅使用常用節奏進行斷詞，那麼就無法切分出五字句中（3，2）（2，2，1）（1，4）（1，2，2）等組合的詞彙。

4.4.3. 使用所有斷詞模組以及標準斷詞順序

這個實驗中，採用斷詞系統中所有的斷詞模組（專有名詞、領字、典故、構詞、節奏斷詞、對仗），搭配 KDE 詞庫，並根據第三章 3.2.2.節定義的斷詞規則為斷詞順序進行實驗。同樣針對無法以任何模組切分的三字句或讀，以亂數方式切分為（2，1）、（1，2），再將每首詞由斷詞系統處理十次，求其平均的精確度與召回率。

由於採用亂數方式切分斷詞器無法斷出的三字句會影響斷詞的結果，我們也以不處理系統無法切分的三字句、讀來進行實驗，以做比較。

表 9：使用所有斷詞模組以及標準斷詞順序的斷詞結果

	使用所有斷詞模組（亂數切分三字句、讀）	使用所有斷詞模組（不處理系統無法切分的三字句、讀）
召回率（Recall）	88.8%	84%
精確度（Precision）	89.4%	88.4%
效度（Effectiveness）	88.7%	86.1%

參考表 9，整體的召回率是 88.8%，精確度為 89.4%，資料的效度達 88.7%。觀察斷詞的結果發現，影響精確度、召回率與資料效度的主要問題在於有些三字句、讀並沒有辦法透過任何斷詞模組正確地切分。例如：

錯誤：西風/留舊/寒　（系統切分的結果）

正確：西風/留/舊寒

－李清照《菩薩蠻》

「舊寒」可視為一個詞彙，但目前的採用的斷詞方式並沒有辦法可以萃取出這個未知詞。

如果不處理系統無法切分的三字句、讀，則召回率與精確度百分比分別為 84.03%與 88.4%，效度為 86.1%。很明顯地斷詞器辨認出的詞數與斷詞器正確辨認出的詞數皆變少，因此召回率和精確率都稍微下降。

4.4.4. 除專有名詞模組，使用標準斷詞順序

這個實驗中，採用斷詞系統中除專有名詞模組之外的所有的斷詞模組（領字、典故、構詞、節奏斷詞、對仗），搭配 KDE 詞庫，並且使用標準的斷詞順序進行測試，以分析專有名詞模組的運作情況。同樣以李清照五十首詞進行測試，亂數切分無法處理的三字詞，求每首詞斷詞十次平均的召回率與精確度做為斷詞結果。

表 10：除專有名詞模組，使用標準斷詞順序的結果

	所有模組	不含專有名詞模組
召回率（Recall）	88.8%	88.1%
精確度（Precision）	89.4%	89%
效度（Effectiveness）	89.1%	88.5%

參考表 10，在召回率、精確度和資料的效度上比前一節的實驗（使用所有模組）稍為下降一些。無庸至疑，專有名詞模組對於斷詞的結果有正面的影響。統計李清照五十首詞中，總共有 25 個專有名詞，這些專有名詞恰巧都是兩個字。以這個實驗而言，若專有名詞詞庫中未收入這些專有名詞，這些名詞也可能被其它模組斷出，如被節奏斷詞模組正確地切分。不過若專有名詞是三字詞，那麼沒有專有名詞模組的輔助，就可能被其它模組誤斷。就理論而言，專有名詞模組對於長度為三以上的詞彙對於斷詞是有絕對的助益的。

4.4.5. 除領字模組，使用標準斷詞順序

這個實驗中，採用斷詞系統中除領字模組之外的所有的斷詞模組（專有名詞、典故、構詞、節奏斷詞、對仗），搭配 KDE 詞庫，並且使用標準的斷詞順序進行測試，以分析領字模組的運作情況。同樣以李清照五十首詞進行測試，亂數切分無法處理的三字詞，求每首詞斷詞十次平均的召回率與精確度做為斷詞結果。

表 11：除領字模組，使用標準斷詞順序的結果

	所有模組	不含領字模組
召回率（Recall）	88.8%	87%
精確度（Precision）	89.4%	87.7%
效度（Effectiveness）	89.1%	87.3%

參考表 11，排除領字，與 4.3.3 節的實驗節果比對，召回率和精確度就分別下降了 1.8%與 1.7%，而資料效度下降了 1.8%。

在測試的資料當中，李清照所寫的《滿庭霜》、《鳳凰臺上憶吹簫》、《醉花陰》、《好事近》、《行香子》、《念奴嬌 春情》、《長壽樂》、《蝶戀花》、《慶清朝》等詞牌皆使用到領字。從統計資料可發現，對這些詞而言，使用領字模組進行切割，其平均召回率與精確度可達 93.14%與 93.76%，效度為 93.4%參考**表 12**。對照之下，未使用領字模組的召回率與精確度，分別為 85.76%與 86.79%，效度為 86.3%，相距甚遠。

表 12：使用領字、未用領字個別詞的結果

	使用領字模組			未使用領字模組		
詞牌	召回率	精確度	效度	召回率	精確度	效度
滿庭霜	92.3%	92.3%	92.3%	88.5%	88.5%	88.5%
鳳凰臺上憶吹簫	100%	100%	100%	95.3%	95.3%	95.3%
醉花陰	88.1%	88%	88%	79%	79%	79%
好事近	97.6%	97.6%	97.6%	96.8%	96.8%	96.8%
行香子	94.7%	97.3%	96%	78.9%	85.7%	82.2%
念奴嬌 春情	88.7%	88.7%	88.7%	80%	80%	80%
長壽樂	90%	93%	91.5%	75.3%	77.8%	76.5%
蝶戀花	92.65%	92.65%	92.65%	86.8%	86.8%	86.8%
慶清朝	94.25%	94.25%	94.25%	91.2%	91.2%	91.2%
平均	93.14%	93.76%	93.4%	85.76%	86.79%	86.3%

4.4.6. 除典故模組，使用標準斷詞順序

這個實驗中，採用斷詞系統中除典故模組之外的所有的斷詞模組（專有名詞、領字、構詞、節奏斷詞、對仗），搭配 KDE 詞庫，並根據第三章 3.2.2.節定義的斷詞規則為斷詞順序進行測試，以了解典故模組對斷詞的影響。

實驗方法同樣以李清照五十首詞進行測試，亂數切分無法處理的三字詞，求每首詞斷詞十次平均的召回率與精確度做為斷詞結果。

表 13：除典故模組，使用標準順序的斷詞結果

	所有模組	不含典故模組
召回率（Recall）	88.8%	90%
精確度（Precision）	89.4%	90%
效度（Effectiveness）	89.1%	90%

參考**表 13**，這個實驗的結果讓召回率、精確度與效度都達到 90%。那麼為何除去典故模組反而提高了斷詞的召回率與精確度呢? 分析實驗的結果發現，原因就在於：就某些專家而言，有些典故的詞彙較長，大於兩個以上的字，因此有些專家認為這些詞彙可以再行切割成兩個以上、長度較短的詞彙。舉例來說：

笛聲三弄，梅心驚破，多少春情意。

－李清照《孤雁兒》

「笛聲三弄」一句，以漢代橫吹曲[10]中的《梅花落》照應詠梅的命題，同時還聯想到園中的梅花，好像一聲笛曲，催綻萬樹梅花帶來春天的消息[32]。因此，有些專家認為此句可以當成典故，直接斷成「笛聲三弄」；但有些專家則以為此句僅是描述窗外傳來一陣陣悠揚的笛聲樂曲，而以為應該斷成「笛聲/三弄」。

同樣地，在相同的詞牌中另一句：

吹簫人去玉樓空

－李清照《孤雁兒》

這一句可以斷成「吹簫/人去/玉樓/空」四個詞彙，也可以斷成「吹簫人去/玉樓/空」三個詞彙。「吹簫人去」出自《列仙傳》秦穆公女弄玉與其夫蕭史的典故，「吹簫人」指的是蕭史，而李清照以「吹簫人」暗喻其夫趙明誠。感慨趙明誠已過世，人去樓空。若斷詞器斷成三個詞彙，而專家認為應該是四個詞彙較精確，那麼這樣的情況會讓加入典故模組的召回率和精確度下降。類似的例子比比皆是，此問題是宋詞斷詞的兩難，每個專家見解都不相同，不易論斷。

另一項造成典故模組召回率和精確度下降的原因是在於典故詞庫的搜集是否

[10] 橫吹曲為一種音樂的品種，相傳在漢武帝時由張騫從西域傳入，以鼓角爲伴奏，在馬上橫吹，於行軍時使用。

完備。我們收錄的典故資料庫中雖然已含大量的典故資料，但仍難以詳盡。再者，典故資料難以識別也增加比對的困難度。舉例來說：

徐娘傅粉

－李清照《多麗》

其中「傅粉」出自典故南朝宋劉義慶《世說新語》記載三國時何晏臉色白淨，魏明帝以為他抹粉，後世用以稱美男子。唐朝李端《贈郭駙馬詩》中描述：「熏香荀令偏憐少，傅粉何郎不解愁。」但由於我們的典故資料庫收錄的是「何郎傅粉」，而非「傅粉」、「傅粉何郎」，因此要如何斷定「傅粉」是出自典故，還是只是美男子的通稱?

目前關於典故的斷詞問題就有賴專家進行校正。

4.4.7. 除構詞模組，使用標準斷詞順序

這個實驗中，採用斷詞系統中除構詞模組之外的所有的斷詞模組（專有名詞、領字、典故、節奏斷詞、對仗），搭配 KDE 詞庫，並根據第三章 3.2.2.節定義的斷詞順序進行測試，以了解構詞模組對於斷詞的影響。

構詞模組的產生是因為我們採用的詞庫並沒有收集定詞、量詞、定量複合詞等這些可以使用規則切分出來的詞彙，簡單的說，構詞模組的作用是補詞庫不足之處。

實驗的方法同樣以李清照五十首詞進行測試，並以亂數切分無法處理的三字詞，再求每首詞斷詞十次平均的召回率與精確度做為斷詞結果。

表 14：使用標準順序的斷詞結果

	所有模組	不含構詞模組
召回率（Recall）	88.8%	88.9%
精確度（Precision）	89.4%	89.5%
效度（Effectiveness）	89.1%	89.2%

構詞模組的加入所得到的測試結果和 4.4.3 節的實驗進行比對，相形之下，三個評估指標都減少 0.1%，如**表 14**。這代表使用了構詞模組，召回率、精確度和效

度這三個評估指標卻降低了。分析實驗結果，發現構詞模組雖然如預期般切出定詞、量詞、定量複合詞、複疊詞、詞綴，也輔助斷出一些三字詞，但卻造成搶詞的情況。舉例來說：

錯誤：醉莫插/花花/莫笑

正確：醉莫/插花/花/莫笑

－李清照《蝶戀花》

構詞模組將「花花」這個複疊詞正確切分，卻造成語意上的錯誤。另一個例子：

錯誤：惜/春春/去

錯誤：惜春/春去

－李清照《點絳脣》

不過這種情況並不多見，在實驗的五十闋詞中因複疊詞規則產生的錯誤只有這兩處。

4.4.8. 除對仗模組，使用標準斷詞順序

這個實驗中，採用斷詞系統中除構詞模組之外的所有的斷詞模組（專有名詞、領字、典故、構詞、節奏斷詞），搭配 KDE 詞庫，並根據第三章 3.2.2.節定義的斷詞順序進行測試，以了解對仗模組對於斷詞的貢獻。

同樣以李清照五十首詞進行測試，亂數切分無法處理的三字詞，求每首詞斷詞十次平均的召回率與精確度做為斷詞結果。

表 15：除對仗模組，使用標準順序的斷詞結果

	所有模組	不含對仗模組
召回率 （Recall）	88.8%	88.3%
精確度（Precision）	89.4%	88.9%
效度（Effectiveness）	89.1%	88.6%

參考**表 15**，不採用對仗模組，斷詞的召回率（88.3%）、精確度（88.9%）、效度（88.6%）比 4.3.3 節斷詞方式召回率稍為差一些，顯然對仗對斷詞成果有一

定的助益。唯不能大幅提升評估指標的原因在於：對仗資料庫尚未能完整地建置出來。目前對仗資料庫還在持續地建置當中。

4.4.9. 解歧義

本實驗針對三字句、讀進行解歧義的動作，以 KDE 詞庫的詞頻資料進行測試。測試的條件為使用所有的斷詞模組（專有名詞、領字、典故、構詞、節奏斷詞），搭配 KDE 詞庫，並根據第三章 3.2.2.節定義的斷詞順序進行。同樣以李清照五十首詞進行測試，亂數切分無法處理的三字詞，求每首詞斷詞十次平均的召回率與精確度做為斷詞結果。

表 16：解歧義

	不解歧義 所有模組	解歧義 所有模組
召回率 （Recall）	88.8%	82.42%
精確度（Precision）	89.4%	82.94%
效度（Effectiveness）	89.1%	82.7%

從實驗的結果得知，使用詞頻總合解歧義並不能有效地提升精確度、召回率與效度，反而讓這些指標下降。分析下降的原因是因為，詞頻的資訊由斷詞器自動累計，但截至目前為止，系統處理過的詞數目還不夠多，因此累計的詞頻資訊還不夠豐富，大部份斷出的詞彙的詞頻都為零，讓解歧義的成果不彰。未來當處理的詞數目增多之後，詞頻的資訊也會更多，那麼對於召回率、精確度和效度就可能會有提升的效果。

4.4.10. 斷詞實驗小結

就整體斷詞的表現而論，從實驗的結果得知這些模組對於斷詞方面都是有相當的助益，滿意度方面是可接受的。雖然目前典故模組的實驗結果讓召回率和精確度稍微下降，但在未來在典故資料庫的收集若能更完整，相信對斷詞器而言具有加分的效果。相同地，在解歧義部份，目前受限於詞頻資訊不夠豐富而效果不是很好，未來在斷詞器自動累計的詞頻資訊更多時，應該可以得到較佳的效果。

至於構詞模組的部份，後續可從擴大實驗的詞下手，從多家詞人的作品中再進行規模更大的測試來看其成效。或者，可以斟酌採用更多的構詞規則以補其不足之處。

4.5. 本體論實作

為了藉資訊科技之力，讓宋詞的學習變得更為簡單，宋詞語意本體的建立格外的重要。利用本體描述詞彙之間的關係，可以協助學習者釐清語意，以便了解詞中的含意。接下來後續的小節中將介紹本體論的實作成果。

4.5.1. 實作架構

在本體論的建置上，首先我們參考 TOVE 本體論工程的方法學進行初步的資料搜集與分析的動作（參閱第三章），接下來依照此階段得到的資訊進行實作。首先我們先介紹本體論的基本實作架構。

參考圖 17我們規劃的基本架構為：先將宋詞斷詞器斷好的詞彙資料，以一闕詞為單位，匯出成XML檔案，然後將這個XML檔案匯入到我們設計的語意設計工具。

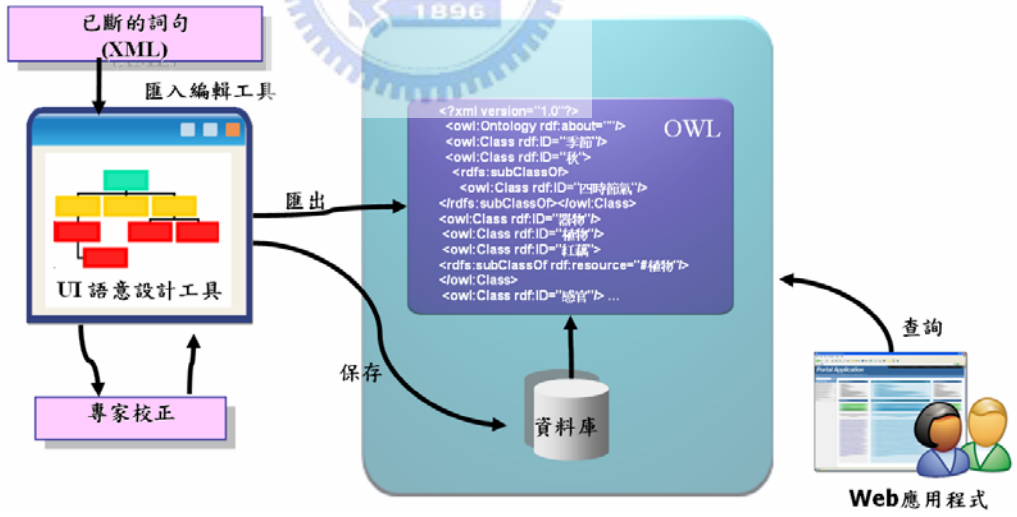

圖 17：本體論實作架構

在語意設計工具中設定詞彙的相關資訊，如概念階層、詞類、同義詞等等。經由專家校正之後，將這些詞彙相關的語意資料儲存到資料庫之中。最後，使用者可以將這些資訊匯出成表示知識的 OWL 文件。我們也設計了「絕妙好詞」Web

應用程式（網站）讓使用者可以透過網際網路進行資料，以及相關語意的檢索。

4.5.2. 語意設計工具

語意編輯工具可以讓使用者或專家建立宋詞詞句中每個詞彙語意相關的資訊，參考圖 18。

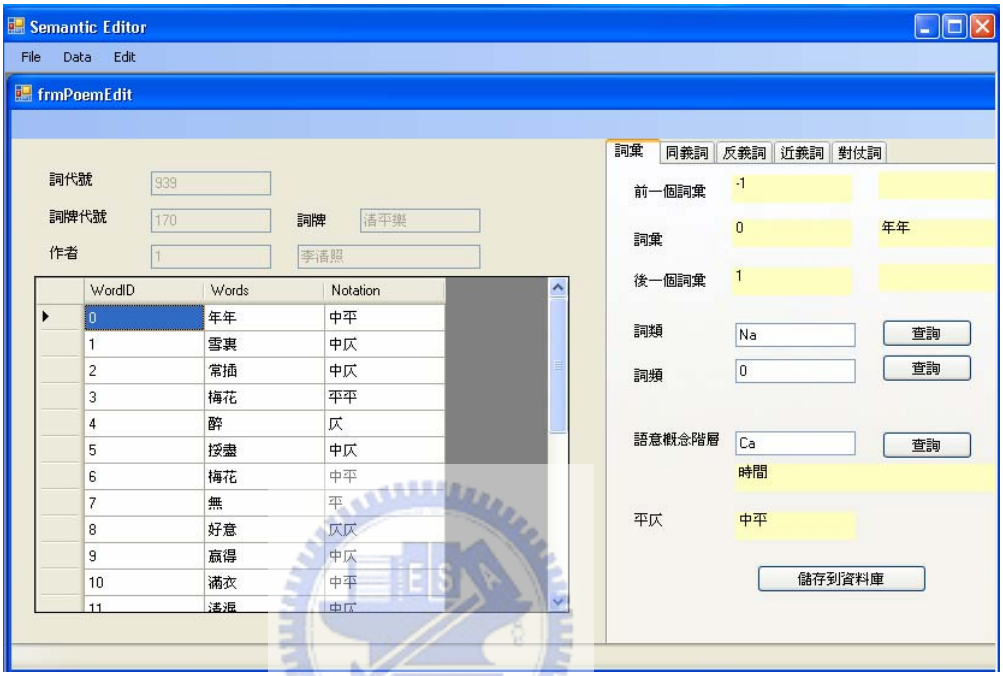

圖 18：語意編輯工具

描述的詞彙資訊包括：

- 詞代號：詞的編號，為唯一識別碼，此項資訊由匯入的 XML 檔自動建立。
- 詞牌：詞的詞牌，如「清平樂」，此項資訊由匯入的 XML 檔自動建立。
- 作者：詞的作者，如「李清照」，此項資訊由匯入的 XML 檔自動建立。
- 詞彙：詞句中的詞彙，如「年年」、「梅花」。
- 前一個詞彙：某一個詞彙前方接的詞彙。如詞句「常插梅花醉」中的「梅花」前一個詞彙為「常插」。此項資訊由系統自動建立。
- 後一個詞彙：某一個詞彙後方接的詞彙。如詞句「常插梅花醉」中的「梅花」後一個詞彙為「醉」。此項資訊由系統自動建立。
- 詞類：詞彙的詞性，如「Nab」為普通名詞。此項資訊由使用者輸入，但系統提供使用者可以查詢中研院詞庫〔八萬目詞〕中詞類的定義，以輔助

建立此項資訊。

- 詞頻：此詞彙在 KDE 詞庫中記載的頻率。此項資訊由斷詞系統自動累計，可以在語意編輯工具中進行查詢。
- 語意概念階層：此詞彙的概念分類，如「植物」。
- 同義詞：記錄和某詞彙同樣意思的其它詞彙，可以有零到多個。舉例來說，針對「植物」這個語意概念而言，「梅花」這個詞彙的同義字可以是「菊花」、「木蓮」。
- 反義詞：記錄和某詞彙有反義關係的其它詞彙，可以有零到多個。舉例來說，「大」這個詞彙反義詞可以是「細」、「小」。
- 近義詞：記錄和某詞彙有近義關係其它詞彙，可以有零到多個。舉例來說，「侵佔」這個詞彙近義詞可以是「侵犯」、「侵略」。
- 對仗詞：記錄和某詞彙以對仗方式呈現的詞彙，可以有零到多個。舉例來說：

  柳絲長，春雨細，花外漏聲迢遞。

  星斗稀，鐘鼓歇，簾外曉鶯殘月。

  －溫庭筠《更漏子》

  「柳絲」這個詞彙對仗詞可以是「星斗」；同理，「長」對「稀」；「春雨」對「鐘鼓」，依此類推。

為了讓使用者能夠很容易地設定此詞彙的語意概念階層，我們採用半自動的方式，列出此詞彙出現在同義詞詞林中的分類，協助專家建立這些資訊。當使用者按下圖 18「語意概念階層」後方的「查詢」按鈕，就會顯示出圖 19的查詢畫面，列出詞彙定義在同義詞詞林中的哪一些分類，讓使用者選取。舉例來說，「蕭蕭」這個詞彙在同義詞詞林中分類在「自然現象」、「擬聲」這兩個分類，但在出現此詞彙的詞句中，原詞到底是描述自然現象或是擬聲這就無從判斷起，需要專家輔助，進行設定。

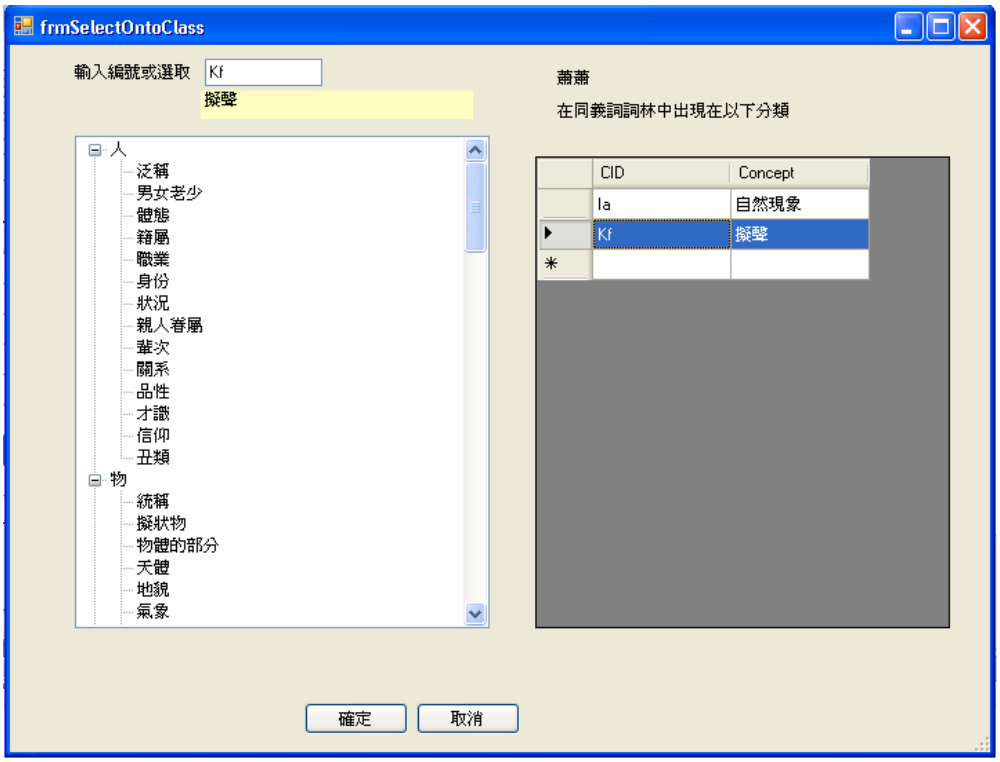

圖 19：自動對應詞彙概念階層

同時也可以利用這個工具來檢視語意概念階層，以及此階層下包含了哪一些詞彙。參考圖 20，左方列出詞彙分類的概念階層，右方則列出此階層下所有詞彙的列表。

圖 20：概念階層

最後語意查詢工具也提供了將這些包含語意資訊的詞彙、詞彙的相關資訊、概念階層等資訊，直接匯出成表達知識的OWL文件，匯出的格式參考圖 21。

圖 21：將詞彙資訊匯出成 OWL 文件以表達知識

4.5.3. 絕妙好詞網站-宋詞語彙網路

網際網路無遠弗界的特色，讓現代人能藉其助力，跨越時間、空間的限制，隨時隨地都能夠利用以網際網路為基礎的學習系統（Web-Based Application）進行學習或資料檢索。

基於上述的概念，我們設計了「絕妙好詞-宋詞語彙網路」網站提供宋詞語彙資訊的檢索，以便讓這些知識能夠更容易地重複利用。使用網站當做使用介面的原因在於它具有以下特色：

- 觸角廣、無時空的限制：只要任何使用者能夠使用到網際網路，電腦上有

瀏覽器就可以在任何時間、任何地點進行存取。

- 易部署：網站程式只需要部署到一台具備Web伺服器的主機上，就可以運作，具備零接觸[11]安裝的特色，降低散發應用程式的成本。和單機版（Windows應用程式，編譯成EXE檔案）相較之下，單機版程式需要程式安裝到每個使用者的電腦上，不易進行部署。
- 易管理：由於系統是建置在一台伺服器上，管理者只需要針對一台電腦進行管理，減輕管理上的負擔。
- 易維護：後續網站程式版本修改時，只要把修改後的二進位檔案更新到伺服器的網站目錄之中，新版本的程式馬上就可以運作。所有的使用者馬上可以使用到新版本的程式和功能。而單機版程式後續更新版本時，需要每一台電腦都重新更新，浪費大量的人力。

圖 22是絕妙好詞網站架構圖，終端使用者可以使用桌上型電腦或者是手持式裝置透過網際網路連結到絕妙好詞網站伺服器上，透過網站程式從後端資料庫伺服器中查詢詞彙相關的資訊。

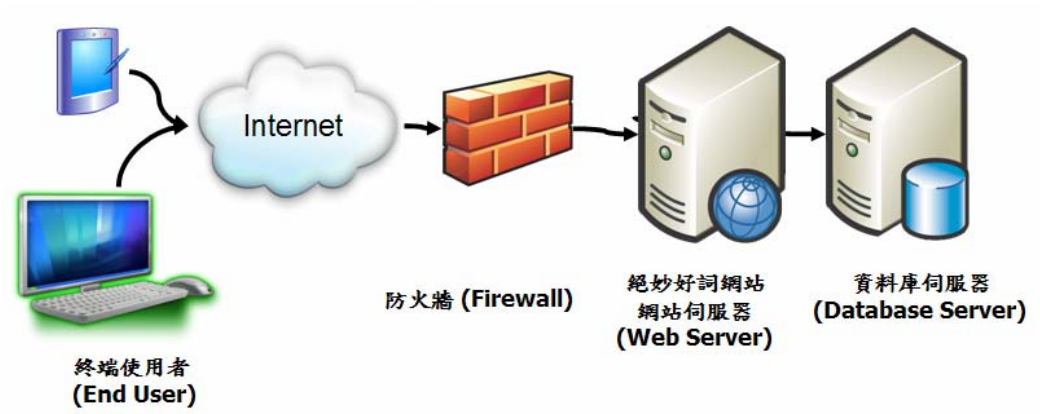

圖 22：絕妙好詞網站架構圖

學習者使用瀏覽器透過網站提供的使用者介面，就可以從網際網路存取相關的詞彙資訊。參考圖 23，只要輸入想查詢的詞彙，如「梅花」，就可以把使用到

[11] 零接觸安裝 (no touch deployment)，不需要接觸到使用者就可以安裝系統。

「梅花」這個詞彙的詞資訊查詢出來。

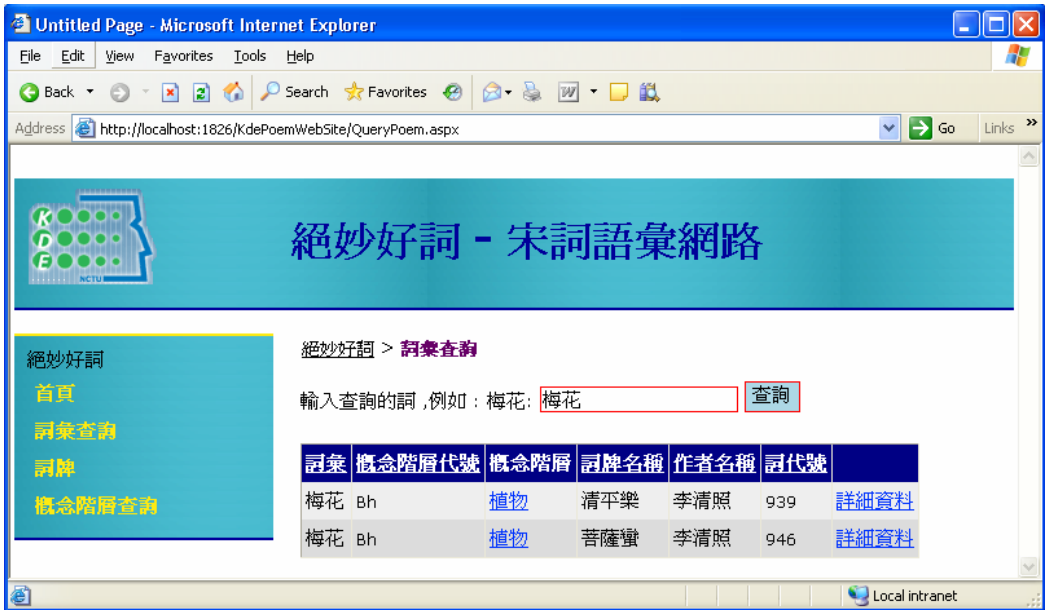

圖 23：絕妙好詞網站

網站提供的功能包含：

- 宋詞詞牌檢索：查詢詞牌相關資訊。
- 概念階層查詢：查詢詞彙概念分類。
- 詞彙查詢：查詢某詞彙相關語意資訊，包含哪些詞中使用到這個詞彙、詞牌、作者、同義詞、近義詞、反義詞、對仗詞，以及此詞彙的前後詞彙，此詞彙所屬的概念階層等相關的語意資訊。

第五章 結論與展望

現代化的電腦資訊科技與吟誦風雅情懷的曲子詞，本是兩個極端，一往前創新，一往舊懷古，此一新一舊各具不同風情。拜科技之賜，現今可以藉由資訊科技融合了具備藝術性質的詞學，賦予古典文學新的生命力，也能藉由科技之推手，讓更多的人能夠欣賞到詩詞之美。

本篇論文根據宋詞依聲填詞、按節奏停頓的特色設計了一個智慧型的規則式宋詞斷詞器，利用 Meta Rule 和規則自動決定使用的斷詞模組和模組順序，以萃取出詞句中的詞彙。透過專有名詞模組使專有名詞不致被切分為單字詞，以免造成語意的混淆或錯誤。領字模組可切分出宋詞專有的虛字，虛字不會和其它的字組合成一個詞。典故模組可保留典故資訊，不致因斷詞而喪失原意。構詞模組可以補詞庫之不足，以解決部份未知詞問題。節奏斷詞模組則根據音律、句式對詞句做切割，以萃取其中的詞彙。對仗模組則以詞句常用對仗加強氣勢的特色，輔助切分三字詞。就整體而言，所獲得的召回率與精確度兩者最佳可達 90%的滿意度，資料的效度最高也達 90%，表示我們提出以宋詞節奏做為切分詞彙的方法具有一定的成效。宋詞斷詞器是以節奏斷詞模組為中心，除節奏模組之外，以領字模組對斷詞的影響最為顯著，能夠切割出宋詞的領字，就能夠大幅提高精確度、召回率和效度。

在宋詞詞彙本體方面，為節省大量人力建置成本，我們設計了語意編輯工具，以便對宋詞斷詞器切分出的詞彙，進行概念階層和相關的語意描述。此外，還設計了「絕妙好詞」網站，藉網際網路無遠弗界的能力，讓更多人可以穿越時間、空間的藩籬，利用它進行詞彙語意的檢索與線上學習。這個網站可以輔助現代人學習古典詞學之美，了解詞彙的涵意、典故，也能讓學習者透過詞彙的同義詞、近義詞、反義詞、平仄、詞彙之間的關係，舉一反三，望「詞」生義，而不致對詞意有所誤解。

針對上一章的斷詞實驗成果，我們得到了以下的結論，以及未來發展方向：

- 需持續搜集詞庫：要能正確切分專有名詞、典故資訊有賴於詞庫的完整性，詞庫越完整，精確度和召回率就越高。同時，也需持續更新對仗資料庫的資料，

以幫助斷詞。

- 增加構詞規則：我們採用的詞庫中不收錄定詞、量詞、定量複合詞、複疊詞等詞彙，構詞規則的產生可以補詞庫之不足之處。挑選該增加哪些構詞規則以輔助斷詞，也是研究的重點。
- 以機器學習方式自動斷詞：未來對於任何斷詞模組都無法正確切分的三字詞部份，可以考慮尋找詞彙的構詞規則，或結合機率的方式計算字與字之間成詞的可能性。或者，也可以思考詞彙與詞彙是否經常在上下文中一起出現的共現關係，以抽取詞句中的詞彙，利用機器學習的方式，自動斷字詞。

在本體論方面，目前本體中的詞彙以及語意資料數量還不夠完備，未來以持續建置這些資料為重心，才能提供更多的應用。此外，本體資料的建置有賴專家的引導，才能保持資料的正確性。未來本體的語意資料夠豐富時，可以將本體的資訊應用在斷詞器上，當做解岐義的一種策略。

參考文獻

[1] Fernndez, M., Gmez-Prez, A. and Juristo, N. "Methontology: From ontological art towards ontological engineering", In Proceedings of Workshop on Ontological Engineering: AAAI-97 Spring Symposium Series, Stanford, CA, 1997.

[2] Michael Gruninger and Mark S Fox, "TOVE Methodology for the Design and Evaluation of Ontologies", Department of Industrial Engineering University of Toronto, Toronto, Canada, M S A, 1995.

[3] Mike Uschold, Michael Gruninger. The Knowledge Engineering Review, 1996.

[4] Tim Berners-Lee. , Weaving the Web: Origins and Future of the World Wide Web, Texere Publishing, US., 1999.

[5] Resource Description Framework (RDF). 9 Mar. 2006. World Wide Web Consortium. 25 May. 2006 < http://www.w3.org/RDF/>.

[6] Suggested Upper Merged Ontology. 11 Oct. 2005. IEEE Standard Upper Ontology Working Group. 25 May. 2006 <http://ontology.teknowledge.com/>.

[7] 中英雙語知識本體詞網。2003 年 10 月 1 日。中央研究院。2006 年 5 月 25 日。<http://BOW.sinica.edu.tw>.

[8] 中文斷詞系統。2004 年 9 月 1 日。中央研究院。2006 年 5 月 25 日。<http://ckipsvr.iis.sinica.edu.tw/>.

[9] 宋詞全首閱讀。2004 年。中央研究院。2006 年 5 月 25 日。<http://elearning.ling.sinica.edu.tw/SCPoemsframe.html>.

[10] 唐宋金元詞文庫及賞析系統。2001 年 8 月。南京師範大學。2006 年 5 月 25 日。<http://202.119.104.80/Ci_ku/ci_web/title2.htm>.

[11] 俞士汶、胡俊峰，「唐宋詩之詞彙自動分析及應用（Word-based Statistical Analysis of Chinese Ancient Poetry）」，語言暨語言學，第四卷第三期，2003。

[12] 胡俊峰，"基於詞彙語義分析的唐宋詩電腦輔助深層研究"，北京大學，博士

論文，2001 年 5 月 25 日。

[13] 羅鳳珠，「詩詞語言詞彙切分與語意分類標記之系統設計與應用」，第四屆數位典藏技術研討會，中央研究院主辦，2005 年 9 月 1-2 日。

[14] 羅鳳珠，「唐宋詞單字領字研究」，第七屆漢語詞彙語意學研討會，臺灣交通大學主辦，2006 年 5 月 22-23 日。

[15] 羅鳳珠，「以資訊科技作為宋詞領字研究方法探討」，第六屆詞彙語意學會議，廈門大學主辦，2005 年 4 月 21-22 日。

[16] 詞庫小組，「中文詞類分析」，中文詞知識庫小組技術報告#93-05，南港，中央研究院，1993

[17] 詩詞典故網站。2002 年 10 月 10 日。羅鳳珠。2006 年 5 月 25 日。<http://cls.hs.yzu.edu.tw/ORIG/>.

[18] 龍沐勛，倚聲學 （詞學十講），里仁書局，2003 年 9 月初版三刷。

[19] 王力，詩詞格律，中華書局，2004 年 2 月再版。

[20] 士會，詩詞挈領，萬里機構萬里書店，2001 年 7 月第二次印刷。

[21] 梅家駒、竺一鳴、高蘊琦、殷鴻翔編，同義詞詞林，上海：上海辭書出版社，1996 年第二版。

[22] 常用詞首、詞尾字資料庫查詢。中央研究院詞庫小組。2006 年 5 月 25 日。<http://140.109.19.103/affix/>.

[23] 吳丈蜀，詞學概說，中華書局，香港，2002 年。

[24] 陳振寰，讀詞常識，萬卷樓圖書公司，1990 年 3 月初版。

[25] Mo, Ruo-ping Jean, Yao-Jung Yang, Keh-Jiann Chen and Chu-Ren Huang., ”Determinative-Measure Compounds in Mandarin Chinese: Their Formation Rules and Parser Implementation.”, In Proceedings of ROCLING IV（R.O.C. Computational linguistics Conference）. , pp. 111-134.

[26] 唐大任，"中文斷詞器之研究"，國立交通大學，碩士論文，民國九十一年七月。

[27] 詞庫小組，「資訊處理用中文分詞標準草案」，經濟部中央標準局，1996 年。

[28] 陳弘治，詞學今論，文津出版社，1991 年 7 月。

[29] RDF Vocabulary Description Language 1.0: RDF Schema. 10 Feb. 2004. World Wide Web Consortium. 25 May. 2006 <http://www.w3.org/TR/rdf-schema/>.

[30] OWL Web Ontology Language Overview. 10 Feb. 2004. World Wide Web Consortium. 25 May. 2006 <http://www.w3.org/TR/owl-features/>.

[31] DAML+OIL (March 2001) Reference Description. 18 Decc 2001.World Wide Web Consortium. 25 May. 2006 <http://www.w3.org/TR/daml+oil-reference>.

[32] 高明，王熙元，陳弘治，張仁青，莊雅州，閔宗述，李周龍編，中國文學總欣賞，9，初版，錦繡文化企業，民國 81 年 8 月。