詩詞語言詞彙切分與語意分類標記之系統設計與應用

羅鳳珠・元智大學中國語文學系
gef julo@saturn.yzu.edu.tw
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

摘要

本研究分析詩詞文體的語言特性，輔以詞譜、典故、人名、地名等專有名詞語料庫，建立詩詞語言詞彙切分之規則，設計詞彙切分及語意標記分類自動控制系統，所產生的詩詞詞彙及各類領域詞彙，提供文學研究使用，為引用資訊科技作為文學研究輔助工具探索可行的研究方向。

關鍵詞：
唐詩、宋詞、詞彙切分、語意標記、自動控制系統

壹、 研究動機與背景

以電腦作為文學研究的輔助工具，在電腦只能分辨字形，無法分辨字義的限制下，應用上受到很大的侷限，與人的判斷存在很大的距離，如何使電腦的判斷接近人腦，是否能判斷字詞義是其中的重要關鍵。

詞義的判斷需仰賴語意標記，語意標記需以詞彙切分為基礎，中國文字是單字單音，構成詞組的文字從一字到十六字都有。以《辭源》為例，收單字詞 12,890 條，多字詞 82,802 條，共計 95,692 條，雙字詞 66,087 條佔 69.02%，數量最多，其次是佔 13.47%的單字詞，再其次是佔 9.63%的三字詞（註一）。單字可以成詞，二個以上的字也可以組合成新的詞彙，其組合的字數、方式，所產生的詞義千變萬化，即便是如人名、地名、動物、植物、人造物等專有名詞，其描述的詞彙也有本名、別名、通俗名的不同，這些都造成詞彙切分與語意標記的困難。

詞彙的組合有各種不同的變化，詞義也非單一固定不變，每一組詞彙的詞義或一種，或二種以上，人類會隨著文明的進展創造新的詞彙，或在舊有的詞彙延伸出新的詞義，文學創作者對詞彙的創新與詞義拓展扮演重要的角色，詩詞文體的詞義創新度與詞義變化度居各類文體之冠。

如何讓電腦具備人的知識，具備學習的能力，使之在詞彙切分及語意標記方面能更接近人腦，如何設計一套有學習能力的詩詞語言處理系統，使詞彙切分與語意標記更有效率，再將處理結果藉助電腦強大的統計、運算、分析能力，提供詩詞文學研究使用，這是本研究的主要目的。

在語體文的語言處理方面，無論是「詞彙分析技術，未知詞判別及其詞類詞義猜測，句剖析，知識抽取與表達等功能」（註二）中央研究院陳克健教授與北京大學計算語言學研究所俞士汶教授所領導的團隊已經有很好的成果。然而現有的成果無論是針對古漢語或近、現代漢語，多數研究仍集中在處理語體文，對韻文的處理比較少，已有的成果包含筆者與北京大學計算語言學研究所俞士汶、胡俊鋒合作的「唐代名家詩語文標記系統」（2000 年），從其後發表的〈唐宋詩之詞彙自動分析

及應用〉(註三)論文可知是針對四百八十一萬字的全唐詩及一百六十萬字的宋代名家詩為範圍，利用「共現度」、「結合強度」等統計參數的計算方法，與傳統的「互信息」法進行比較，進行詞彙之自動提取、自動分析（註四)，並將句法的因素列入考慮（註五)。北大的研究已建立很好的基礎，但北大在沒有任何基礎語料可資比對的限制之下，很多三字詞的專有名詞被切割為二，例如《全唐詩》於 23 首詩之 23 句使用「鸚鵡洲」，而北大研究所抽取的詞彙，「鸚鵡」出現 220 次，但沒有「鸚鵡洲」；全唐詩於 22 首詩之 22 句使用「黃鶴樓」，而北大研究所抽取的詞彙，「黃鶴」出現 117 次，但沒有「黃鶴樓」。《全唐詩》所使用的典故，二至十字詞都有，北大研究所抽取的詞彙為一、二、三字詞，四字以上的典故詞彙如「傅粉何郎」、「二桃殺三士」、「少壯幾時奈老何」、「魏王臥內藏兵符」、「死生元有命，富貴本由天」等，都被切分為一、二、三字詞而消失了典故用語的原貌與詞義。

本研究嘗試分析詩詞文體的特性，以新的方法探討詩的詞彙抽取與語意標記的方法，並從詩擴充到詞（唐宋詞)，希望所提出的方法可以彌補前人研究尚未顧及的部分，得出更完整的結果，以提供文學研究使用。

貳、詞彙自動切分控制系統設計

詩，分為近體詩和古體詩，古體詩有四言、五言、七言、雜言（長短句皆有)，近體詩分為五言絕句、五言律詩、七言絕句、七言律詩，還有極少數的六言詩，近體詩有固定的格律。唐宋詩詩作之中，除了典故及專有名詞之外，以雙字詞最多，單字詞次之；古近體五言詩詩句句首通常使用雙字詞，古近體七言詩詩句句首通常使用雙字詞，3、4 字通常也是一組雙字詞，末三字是 2、1 或 1、2 的可能性各半，也可能是三字詞。以上規則可能出現的例外是句首使用專有名詞時，就有可能出現以三字詞為句首，例如：白居易〈和劉郎中傷鄂姬〉:「不知月夜魂歸處，鸚鵡洲頭第幾家。」李白〈峨眉山月歌送蜀僧晏入中京〉:「黃鶴樓前月華白，此中忽見峨眉客。峨眉山月還送君，風吹西到長安陌。」其中的「鸚鵡洲」、「黃鶴樓」、「峨眉山」都是句首三字詞；或是使用典故時，也可能在句首出現三字以上的詞彙，例如盧綸〈送黎燧尉陽翟〉:「潘縣花添發，梅家鶴暫來。」,「梅家鶴」典出《漢書梅福傳》、又如韓翃〈贈別華陰道士〉:「賣鮓市中何許人，釣魚坐上誰家子。」,「釣魚坐上」典出《後漢書方術傳・左慈傳》；此外古體詩之樂府詩有時以「君不見」起始，例如李白〈將進酒〉:「君不見黃河之水天上來」；有時以單字詞起始，例如徐堅〈送考功武員外學士使嵩山置舍利塔歌〉:「伊川別騎，灞岸分筵。對三春之花月，覽千里之風煙。望青山兮分地，見白雲兮在天。寄愁心於樽酒，愴離緒於清弦。共握手而相顧，各銜悽而黯然。」詩句中之「對」、「覽」、「望」、「見」、「寄」、「愴」、「共」、「各」都在句首，都是單字詞。

詞的格律比較複雜，歷代留下來的詞調共有八百七十五調（註六)，每一詞牌有固定的總字數、總句數，每一句的字數也是固定。王力《漢語詩律學》第三章第四十一節〈詞字的平仄〉，論及詞最短的句子是一字句，最長是十一字句。使用一字句的詞牌只有〈哨遍〉後段起句「噫」以及陸游〈釵頭鳳〉:「錯！錯！錯！」、「莫！莫！莫！」、呂渭老〈惜分釵〉:「重！重！」、「忡！忡！」。二字詞出現在〈調笑令〉、〈如夢令〉、〈醉翁操〉、〈河傳〉、〈南鄉子〉、〈戚氏〉等詞牌，十字句只有〈摸魚兒〉前闋第六句和後闋第七句，是上三下七句法；或將〈攤破浣溪沙〉後闋的末二句認為一句，是上七下三句法。十一字句只出現在〈水調歌頭〉，其句法有上六下五及上四下七二種。三字句至九字句佔的數量最多，其句法分別是：三字句：上二下一、上一下二；四字句：上二

下二；五字句：上二下三、上三下二、上一下四；六字句：上二下四、上四下二、上三下三；七字句：上三下四、上四下三、上一下六；八字句：上三下五、上四下四、上一下七、上二下六；九字句：上三下六、上四下五、上五下四、上六下三。這些上下句之上句若是奇數字句，多數句子之第一個字是領字，尤以五字句之上一下四，七字句之上一下六為最（同註六）。

從以上的分析可知，唐宋詞每句總字數是偶數句的句子比奇數句多，各句之句法，「一」是單字詞，「二」是雙字詞，殆無疑義，字數為偶數之四、六字句，除使用專有名詞或典故之外，多數可切分為2、2及2、2、2字詞，字數為奇數之三、五、七字句，除使用專有名詞或典故之外，三字句為2、1或1、2的機率各半，五、七字句可切分為2、3及2、2、3，末三字為2、1或1、2的機率各半。唐宋詞句首是單字詞者，幾乎都是領字，那些詞牌的那些句子使用領字是固定的，以柳永〈八聲甘州〉：「對瀟瀟、暮雨灑江天，一番洗清秋。漸霜風凄緊，關河冷落，殘照當樓。是處紅衰翠減，苒苒物華休。惟有長江水，無語東流。不忍登高臨遠，望故鄉渺邈，歸思難收。歎年來蹤跡，何事苦淹留。想佳人、妝樓顒望，誤幾回、天際識歸舟。爭知我、倚闌干處，正恁凝眸。」為例：「對」、「漸」、「望」、「歎」是領字，其詞譜為：仄／＋平仄仄仄平平，＋＋仄平平（平韻）。仄／平平＋仄，＋平＋仄，＋仄平平（平韻）。＋仄平平＋仄，＋仄仄平平（平韻）。＋仄平平仄，＋仄平平（平韻）。＆＋仄＋平＋仄，仄／＋平＋仄，＋仄平平（平韻）。仄／平平＋仄，＋仄仄平平（平韻）。仄平平、＋平平仄，仄＋平、＋仄仄平平（平韻）。平平仄、仄平平仄，＋仄平平（平韻）。（以「／」表示領字位置。以「＋」表示該字平仄皆可，以「＆」表示分片）因此可在每一詞牌之格律標記領字的位置，以電腦作詞彙切分時優先將領字切分。

從以上詩詞文體之特性，設計詩詞詞彙自動切分控制系統，詩詞詞彙切分之步驟如下：

一、 以唐宋詩詞等詩詞資料庫為基礎，適用範圍包含現代詩之外的所有韻文。

二、 建立唐宋詞詞譜八百七十五種：

（一） 以王兆鵬、劉尊明等主編的《宋詞大辭典》為底本，以龍沐勛所著《唐宋詞格律》增補，建立宋詞詞譜資料，並標記領字的位置。

（二） 從已經標記領字位置的唐宋詞詞譜與唐宋詞資料庫比對，抽取領字資料，建立宋詞領字語料資料庫。

三、 建立各種專有名詞詞彙資料庫

筆者於2002年主持之國科會「以XML(eXtensible Markup Language)可延伸式標注語言建立文章標誌(Content Markup)系統研究：以蘇軾詩詞為範圍」之研究計畫，以蘇軾詩為範圍，進行詞彙切分及語意標記工作，所產生之專有名詞詞彙資料，總計有人名（本名與別名字號）、神鬼仙人專名、地名（古今地名、通俗簡名）、閬苑仙境專名、天文專名、建築物專名、官銜職稱專名、動物專名、植物專名、神物專名、器物專名、朝代專名、國名專名、書篇名專名、樂曲專名、舞曲專名、戲曲專名、稱謂專名等十八類。這些資料作為全面切分詩詞語料之專有名詞基本詞彙資料庫。（註七）

四、 建立詩詞典故詞彙資料庫

筆者於參與清華大學「下一世代資訊通訊網路尖端技術及應用--網路教育園區及其社會影響研究」卓越計畫(2000年1月至2004年12月)，負責古典文學館之執行，建置詩詞典故網站（註八），已建立唐詩典故詞彙9,698筆，宋詞典

故詞彙10,518筆，總計20,216筆，詩詞裡重複使用同一個典故的現象很普遍，因此可以列為優先比對之條件。

五、詞彙切分方法及步驟：

（一）第一優先比對專有名詞資料庫，以專有名詞做為詞彙切分之最優先順序。

（二）第二優先之切分順序是領字：依據所建立的宋詞詞譜優先將領字切分，其餘詞句依「宋詞詞譜」之步驟切分，例如：「漸／霜風／淒緊／，關河／冷落／，殘照／當樓／。」（柳永〈八聲甘州〉）先將領字「漸」切分，其餘字數依「（四）之1、2、3」步驟切分。

（三）第三優先比對典故詞彙資料庫

（四）宋詞除了專有名詞之外，以雙字詞佔絕大多數，因此，未使用領字的詞句以及扣除領字之外的詞句，以下列步驟切分：

1. 句子字數為偶數句之句子，兩兩切分為雙字詞，例如：「離恨／恰如／春草／，更行／更遠／還生／」（李煜〈清平樂〉）。
2. 句子字數為奇數句之句子，末三字之外的句子，兩兩切分為雙字詞，例如：「菡萏／香銷／翠葉殘／，西風／愁起／綠波間／，還與／韶光／共憔悴／，不堪看／。」（李璟〈浣溪沙〉）因為除了領字及三字詞、五字詞等奇數字的專有名詞之外，宋詞詞句少有以奇數字為句前詞彙。
3. 奇數句句子之末三句，其句法或作1、2，或作2、1，並無準則，切分的方法需仰賴詞庫以提高切分之正確率。切分時以前述步驟所獲得之雙字詞比對句末之三字詞，若這三個字之前二字或後二字是前述步驟所獲得之雙字詞，則優先取為雙字詞，予以切分，例如：「春到小園春草綠」（辛棄疾〈朝中措〉）從前述「離恨／恰如／春草／，」已取得「春草」之詞彙可知「春草綠」三字，「春草」是詞彙的機率比「草綠」高，因此優先將「春草」認定為詞彙，予以切分。不過也有末三字之前二字或後二字都是詞彙的句子，此時便需以人工校正，例如：「歸心正似三春草。」（蘇軾〈虞美人〉）之「三春」、「春草」都是宋詞常用的詞彙，遇到這種情況，便需以人工校正。

六、唐宋詩之詞彙切分方法與步驟，除了沒有領字之外，其餘均與唐宋詞相同。

七、從前面幾個步驟建立「唐宋詩詞詞彙資料庫」，所建立的詞彙資料，匯入詞彙資料庫，作為切分時的基本比對資料庫。

參、詞彙語意自動標記、分類及控制系統設計

筆者於「以XML(eXtensible Markup Language)可延伸式標注語言建立文章標誌（Content Markup）系統研究：以蘇軾詩詞為範圍」之研究計畫，根據蘇軾詩所切分之詞彙，考量文學研究之需要，對於可以區分知識類別之詞彙予以分類，無法區分者暫時保留，分類表及所含詞彙元素如附表。本研究將以此分類表作為唐宋詩詞語意標記及分類的基礎，再透過詩詞全面之詞彙切分與語意標記，重新修訂分類表，希望能得出更適當、更完整

的分類標準。

以蘇軾詩語意分類表為基礎，設計詞彙標記分類自動控制系統，其功能如下：

一、 蘇軾詩語意分類表為基礎，大類之下有次類、次次類、元素。

二、 容許分類的增減等修訂功能：文本內容之文體，從詩擴充到詞，從一位作家擴充到唐宋詩詞全部作家，所包含的內容必然增加，所以系統需要容許分類類別修訂的功能。

三、 以蘇軾詩的詞彙及語意標記之資料作為「唐宋詩詞基本語料庫」。

四、 新加入的文本所切分的詞彙，遇到「唐宋詩詞基本語料庫」已有的詞彙，可以自動加上已有的標記，並容許修改。自動加上已有的標記可以節省人力，並避免標記不統一，相同的詞彙有可能因為語意不同而分在不同類別，所以需要提供修訂的功能。

五、 顯示未知詞、新詞：「唐宋詩詞基本語料庫」還沒有的詞彙，提供顯示該筆詞彙是新詞的功能，由人工加上語意分類及標記，標記之後自動匯入基本語料庫。由於詞彙自動切分時，電腦有可能判斷錯誤，以人工加上語意分類及標記的過程中，可以檢查該未知詞、新詞是不是正確的詞彙。

六、 如果是需要對仗之詞句，能顯示其對應位置之詞彙。

七、 每一個詞彙均能顯示在同一作品中共現之詞彙。

八、 能顯示每一類別之所有詞彙。

九、 專有名詞語料能同時匯入專有名詞資料庫。

十、 建立各種專有名詞資料庫：除了從蘇軾詩所產生的專有名詞語料庫，本研究還建立了「詩詞典故語料」、「古今地名語料」、「廣群芳譜」植物名稱語料，以供比對。

肆、人工輔助之校正

一、 典故詞彙部分，電腦比較難以判斷：有些詞彙含有多種語意，有些是典故，有些不是，例如「此君」，可解作「竹子」（典出《晉書・王徽之傳》）；又可解作「酒」（典出白居易〈效陶潛體詩十六首（并序）之六〉：「清光入杯杓，白露生衣巾。乃知陰與晴，安可無此君。」）；也可能單純指「這一個人」。

二、 專有名詞優先比對的結果，使得典故詞彙若含有專有名詞，將被優先切分及歸為專有名詞類別，需要人工校正，例如：「元亮秫 」、「元亮井」、「元亮徑 」、「陶潛菊 」，這些典故都與陶潛有關。「元亮」、「陶潛」是專有名詞，將優先被切分歸入專有名詞語料庫，而造成錯誤。

三、 詩詞裡提到人名時，有時以官銜代表人名（蘇軾＝蘇內翰），有時以地名代表人名（柳宗元＝柳柳州），詞彙切分及語意標示時容易造成錯誤。（註九）

四、 作品若不合格律，尤其是該使用領字而未用領字之唐宋詞，詞彙切分時容易造成錯誤。

五、 三字詞之 2、1 與 1、2 如果都是正確的詞彙時，自動控制之詞彙切分系統便難以判斷，而造成錯誤，雖說這二種都是正確的詞彙，但是會影響到作品正確的語意。

電腦畢竟不是人腦，無法判斷的部分需要輔以人工校正。

伍、文學研究之應用

本研究將建立唐宋詩詞全文資料庫、唐宋詩詞格律資料庫，依據格律譜資料，輔以唐宋詩詞文體的句法及詩詞語言特性，開發「唐宋詩詞詞彙自動切分控制系

統」，以建立「唐宋詩詞語料庫」，開發「語意自動標記系統」，以「唐宋詩詞語料庫」進行語意標記工作，從中產生語意分類表項下的「領域詞彙資料庫」，如「植物詞彙資料庫」、「動物詞彙資料庫」、「器物詞彙資料庫」等等，以及「同義詞資料庫」、「唐宋詩詞典故資料庫」、「唐宋詩詞本體知識資料庫」。另以「唐宋詩詞語料庫」建立讀音資料庫以及音韻資料庫，作為標記唐宋詩詞讀音及平仄音韻的基本資料，也可作為「智慧型倚聲填詞系統」、「智慧型依韻作詩系統」，提供自動檢查平仄音韻格律、依語意標記資料庫搜尋各種語意分類語料的功能。

所建立的工具平台（系統）以及各種資料庫，在唐宋詩詞之研究、創作方面，可提供下列的功能，架構圖如附表二：

一、 研究方面

提供古典詩詞風格研究、古典詩詞彙語意學研究、古典詩詞地理資訊研究、古典詩詞意象研究、古典詩詞虛詞用法研究、唐宋詞領字研究、古典詩詞典故研究。

二、創作方面

智慧型倚聲填詞、依韻作詩系統：除了提供自動檢查並協助修改平仄押韻等格律之外，前述各種資料庫的資料可以協助創作時豐富詞彙、提升修辭技巧之功能。

此外，在語意標記的基礎之下，建立可以分辨字義的智慧型全文檢索系統，作為研究時查檢資料的工具；以及所有唐宋詞文本的讀音標記、典故出處含義標記、語意標記、詞類標記等功能，作為解讀唐宋詞的輔助工具，進而建立唐宋詩詞本體知識資料庫。

附表一：語意標記分類表

大類	小類	所含元素
人名	本名	實際人名（含略稱）
人名	別名字號	東坡、子由、太白、子美
人名	他稱	東方先生（東方朔）、嵇中散（嵇康）、玉川子（盧仝）
人名	法號	定慧師、慧能、德長老、芝師、釋了元、德雲
行政體系	部門名稱	中書省
朝代君王	朝代民族名號	六朝、北狄、西戎、西羌、西涼、西蜀、西漢、西號、周、契丹、唐、秦、軒轅、商、周、楚、漢、趙、衛、魯、鄭
朝代君王	君王帝號	晉惠帝、秦始皇
朝代君王	君王年號	元嘉、永嘉、貞觀、開元、嘉祐、皇祐
朝代君王	君王陵墓	裕陵、永昭
神鬼仙人	神鬼專稱	上帝、山神、天人、天女、天公、天王、江神、河伯、河神、龍神、織女、神巫、蚩尤、女媧、伏羲、夷羿、嫦娥
神鬼仙人	神鬼泛稱	仙人、仙子、百神、老仙、神、神人、神女、神仙、神兵、神怪、神翁、鬼神、群仙
神鬼仙人	菩薩尊號	摩詰、文殊、佛、應真、佛祖

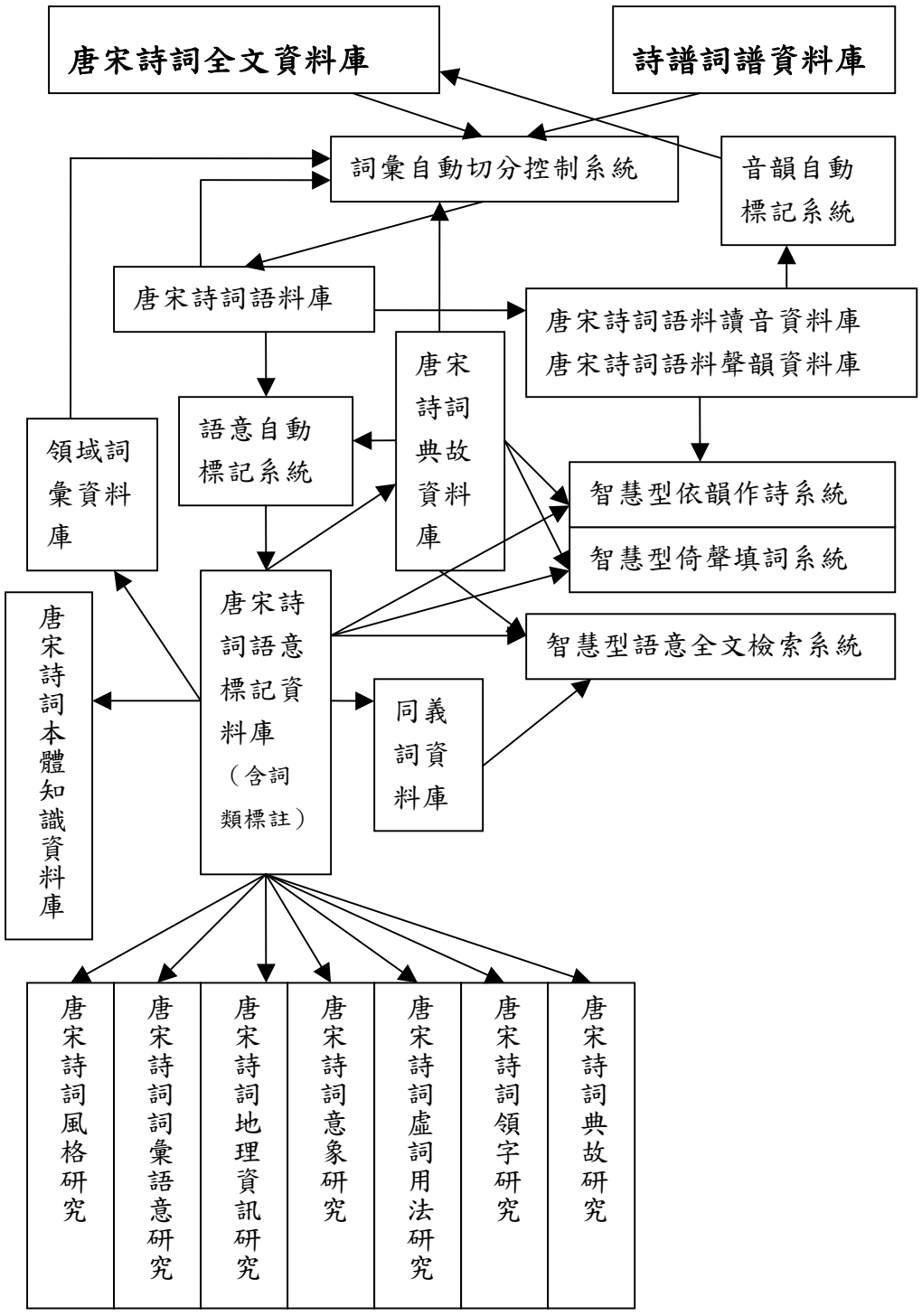
身體器官	五官顏面	耳、眉、眼、鼻、瞳、頸、頷、臉、額、顏、頰、頸、口、舌、睫、唇、齒
身體器官	鬚鬢髮膚	鬒、鬟、鬢、髻、髭、髮、髯、鬚、膚
身體器官	身體四肢	手、肘、足、身、肩、膝、指、腳、股、臂、頭、拳、脛
身體器官	五臟器官	肝、肺、膽、腸、腦、腹、牙、骨、舌、腹、齒
地名	行政地名	實際地名
地名	山嶺地名	九仙山、九曲嶺、九華、九疑、卞山、天山、天台山
地名	河湖地名	中泠水、中泠泉、五溪、六一泉、天池、太湖、巴峽
地名	關津地名	蘇公隄、玉門關、白馬津、白馬關、白鷺洲、朱雀橋
地名	地名代稱	岱宗（泰山）、東武（密州）、東都（揚州）、浮玉山（金山）
地名	地名合稱	三吳（蘇州、常州、湖州）、五嶺（大庾嶺、始安嶺、臨賀嶺、桂陽嶺、揭陽嶺）
地理	區域範圍	河西、中原、西秦、秦、楚鄉、州縣、巴東、荊吳、楚地、衡湘、楚境、江南、西蕃、巴蜀、江南、九州、六州
地理	邦國都城	國、邦、城、邑、都、郡、鄉國、鄉縣、畿
地理	郊原村野	野市、陌、野、路、村、郭、郊、塞、道、平陸、郊、屯、皋、郊原、江淮、淮上、城外
自然景觀	山峰崖嶺	山、峰、壁、嶺、陵、崖、山曲、岡、嶂、山阿、巖、壑
自然景觀	水澤湖泊	水、江、泉、海、峽、潭、湖、池、浪、溪、河、潮、灘、湍、洲、沙、井、淵、澗、灣、渠、川、波、塘、溪谷
自然景觀	山川泛稱	山川、山水、溪山
自然景觀	關津渡口	隄、古塞、小橋、曲港、江口、江驛
建築物	建築專名	岳陽樓、大明寺、大明宮、大槐宮、北固樓
建築物	宮室屋廬	軒、宮殿、堂、舍、室、齋、屋、廬、菴、館、房、塢、店、府、闕、第、宅
建築物	亭臺樓閣	亭、臺、樓、閣、榭、坊、驛、齋、門、闕
建築物	園林院落	園、庭院、莊、苑、廊
建築物	寺院道觀	寺、廟、庵、觀、菴、僧舍、精舍
建築物	碑塔陵墓	碑、塔、陵、墓、祠、坊、壇、冢
建築物	官署建築	監獄（南冠）、東府（宰相府）
建築物	建築部件	椽、窗、壁、階、簷、門楣、瓦、牆、檻、欄、門、籬、廊、欄、檻、梁
閬苑仙境	閬苑	仙府、仙宮、仙壇、瑤臺、閬苑、龍宮、瓊館、瓊宮
閬苑仙境	仙境	天池、仙山、仙都、蓬萊島、緱山、芙蓉城、桃花源
天候氣象	風霜雪露	冰雪、冰雹、風、風雷、風霜、霜風、風雷、風雨、雪
天候氣象	雲霧煙霞	雲、煙、煙雨、霞、霧、霧雨霾、靄、嵐、霰
天候氣象	雷電雹霓	虹、雷、電、雹、霓、雷

天候氣象	冷寒	冷、凍、涼、寒、
天候氣象	熱旱	旱、熱、
天候氣象	晴天	晴、新晴、晴時、晴空、晴和
天候氣象	陰雨	雨、雨雪、陰晴
天文	日	赤日、日腳、出日、跳丸、落日、扶桑、夕烽、斜陽、日
天文	月	嬋娟、月上、月明、月滿、秋月、斜月、霜月
天文	星	星辰、太白星、星河、德星、牛斗、流星、疏星
天文	天空	青天、雲漢、碧霄、如天、抹坤、天、空碧、雲天
天文	銀河	銀河、河漢、銀漢
自然資源	玉石	玉、石、雲母、瑪瑙、琉璃、琥珀
自然資源	火	火、烈火、祝融、野火、陰火、新火
自然資源	金屬	黃金、鐵、鉛、銀、銅、錫
自然資源	礦物	煤、辰砂、丹砂
自然資源	燃料	炭、束薪、珠煤、濕葦、薪炭
飲食	飲品	酒品：杯中物、白玉泉
飲食	飲品	茶品：奇茗、雪湯、貢茗
飲食	食品	菜品、羹、粥、米食
飲食	食材	蔬菜、魚肉、水果、五穀、鋪粟、蓴、鱸、米、瓠葉
飲食	零食	零食、餅食、羊羹、柿霜、涼餅、槐芽餅、湯餅
飲食	調味品	蜜、鹽豉、桂醑醋、糖霜、醬、鹽
飲食	宴席	席上、華筵、歌筵、綺席、綺筵、翠筵
四時節氣	節氣	立春、雨水、驚蟄、春分、清明、穀雨、立夏、小滿、芒種、夏至、小暑、大暑、立秋、處暑、白露、秋分、寒露、霜降、立冬、小雪、大雪、冬至、小寒、大寒
四時節氣	節慶	中秋、重陽、除夜、七夕、上元、寒食、新年、臘日
四時節氣	四季	春、夏、秋、冬
時間	明確時間	詩詞中明確記載之時間詞
時間	範圍時間	清晨、終日、終年、終夜、終朝、終歲、連夜、竟夕
文藝	書籍	詩經、尚書、楚辭
文藝	文具	硯、筆、翰墨、紙、印
文藝	書篇名	楚辭、四庫書、道藏
文藝	文體	書信、詔書、詩、賦、簿書、史、跋、行、回文、公文書
文藝	引文	韋曲花誣賴（杜甫〈陪鄭駙馬〉詩）、他鄉各異縣（古詩）
技藝	畫	丹青、畫、畫品、翰墨、丹墨、三峰圖、畫水、畫圖、丹青、畫骨、畫馬、丹青、書畫、畫佛、《六幅圖》、凌煙圖
技藝	書法	書法、嶧山碑（李斯作）、草書
技藝	碁	碁、棋

技藝	舞蹈	歌舞、魚龍舞、歌舞、楚舞
技藝	戲劇	魚龍
生物	神物	混沌、龍、麒麟、鵬、鳳凰、鯤、鵬、蛟、鴟、鸞
生物	動物	馬、猿、豺虎、鳥、鷹、隼、鹿、龜、兔、鯨、魚、蛟、鶻、烏鴉、虎狼、鼉、鱉、鸕鷀、鷺鷥、蛤蟆、蛇、猱
生物	植物	杏、木瓜、水仙、蓮、瓜、甘棠、白芡、白藤、白蘋、石楠樹、石榴、禾、麻、黍、江蘺、竹、杜若、松、杞菊
稱謂	自稱謙稱	予、奴、老子、老身、臣、臣子
稱謂	尊稱	孝婦、聖人、聖賢、賢人、賢子、賢達
稱謂	貶稱	酷吏、醜廝、驩兜（惡人）
稱謂	親人眷屬	父母、兒孫、姊妹、妻妾、伯叔、舅姑、兄弟、女婿
稱謂	指代稱謂	泛稱，但此處特指某人或某些人
稱謂	一般稱謂	女子、老圃、老翁、老婆、老農、老儒、至人、舟人
稱謂	職官稱謂	中郎、中書、公侯、公相、公卿、勾漏、太史、太守
稱謂	皇室稱謂	君王、公主、王子、王孫、妃子、王公
稱謂	宗教稱謂	真人、僧、天師、佛祖、法師、道人、道士、禪老、禪客、上人、比丘
器物	生活用品	衾被、帳帷、廚具、餐具、燈燭、枕席、擺飾、梳子、鏡子、氈帳、床、酒器、鐘漏、脂粉、屏風、香料、熏爐
器物	交通工具	車、船、轎
器物	工具用品	砧、杵、磨、尺、箠、籠、繩索、轆轤、槊、矢、甲、鋏、劍、畚、鍤、刀、箭、檄、斧、弓、矛、叉、剪刀、刀圭
器物	玉帛服飾	衣裳、布料、巾帽、鞋襪、配飾、盔甲、羅帕、腰帶
器物	宗教用品	符、龍車、虎駕、丹砂、龕燈、木魚、仙藥、僧巾、佛燈、禪榻、金丹、符命、篆符、丹鼎、衲裙、蒲團
器物	喪葬祭祀	棺、篷蓨、金棺、瓦棺、桐馬
器物	娛樂器具	鞦韆、花板（鞦韆）、鬥草、綵繩、骰兒、選仙
器物	金帛錢幣	百金、錢數、一錢、千金
器物	公堂器物	笏、旂、符、旗（旌、旆）、鉞
器物	神器	紫雲車
樂部	樂器	琴、鐘、弦樂器、簫、笛、笙、竽、簧、笳、琵琶、瑟、鼓、角、箏、瑟、箜篌、鼗
樂部	樂曲	樂曲、歌曲、舞曲、歌嘯、樂府、舞
空間	方向	東、西、南、北
空間	位置	上、下、左、右、前、後、內、外
顏色	原本顏色	千紅、小紅、已丹、已白、天紅、半紅、玄、玄黃、白
顏色	借代顏色	雪色、玉、寒鴉、卵色、霜、點酥、未絲、玉色

方言俚語	方言	雞頭鶻、步、軟飽（浙人謂飲酒為軟飽）
方言俚語	俚語	黑甜（俗謂睡熟為黑甜）
典故	佛典	百結花、戒定慧、精明、湛然、叢林、一念、彈指
典故	其他	三徑、南柯、沐猴、三黜、折屐、食雞肋、畫蛇足
虛詞		哉、耳、乎、也
感嘆詞		太息、嘆息
情緒詞		喜、怒（嗔瞋憤）、悲（淒悽咽）、愁、怨、恨（遺憾）、憎、愛憐、憂、苦、惱、惆悵、感傷
感官詞	嗅覺	香、香少、香冷、香消、香留、香殘、清芬、暗香、塵香、餘香、濃香
感官詞	聽覺	聲、人聲、不啼、天籟
感官詞	味覺	苦、甜
感官詞	視覺	觀、覽、視、看
複疊詞		悽悽、戚戚、淒淒、綿綿、蒼蒼、赫赫、儼儼、凜凜
雙字詞		琥珀、珊瑚、蜈蚣、惆悵、淒涼、水仙、枇杷、芙蓉
同近義詞		蝴蝶＝蛺蝶、憔悴＝顦顇、彷彿＝髣彿、葫蘆＝胡蘆、朦朧＝蒙朧、踴躍＝踊躍、歡呼＝讙呼、辜負＝孤負

附表二：架構圖



感謝：本研究獲得國科會【以XML(eXtensible Markup Language)可延伸式標注語言建立文章標誌(Content Markup)系統研究－以蘇軾詩詞為範圍】計畫資助，計畫編號：91-2422-H-155-3401。

附註：

一、 參見拙作〈試論引用資訊科技作為詩學研究輔助工具的發展方向與建構方法〉,2000 年 6 月 29 日-7 月 1 日,2000 年第三屆國際漢學會議，中央研究院。

二、 參見陳克健個人網站之「研究簡介」:「如何建構具有自我學習能力的中文自然語言處理系統為近期的主要研究目標，期望系統能自動分析文件抽取新知。為了達成上述目標，已研究完成以下功能：詞彙分析技術，未知詞判別及其詞類詞義猜測，句剖析，知識抽取與表達等功能。因此本系統建立在語言基礎知識及處理功能上，有了基礎詞庫、語法及處理技術後，正透過分析網路上源源不絕的文件，辨識新詞、分析句結構、抽取詞與詞之間的語義關係，進而擴展到概念與概念之間的關係。剖析器利用詞彙和概念之間的關係強度搭配語法規律完成文句分析，也因為新知識的累積而日益精進表現更精確的文件分析能力。」,網址：http://www.iis.sinica.edu.tw/pages/kchen/cindex.html

三、 俞士汶、胡俊鋒,〈唐宋詩之詞彙自動分析及應用〉,《語言，文學與資訊》,169~192 頁，羅鳳珠主編，2004 年 3 月，清華大學出版社。

四、 此段文字改寫自原文之摘要。

五、 同註三之 179 頁:「在唐宋詩這類特定的語體中，五言的 2、3、七言的 2、3 字之間，4、5 字之間一般不會出現雙字詞。所以在計算結合強度時在相應位置的字串可以不認為是相鄰出現。」

六、 參見拙作,〈以資訊科技作為宋詞領字研究方法探討〉:「合計唐圭璋所編《全宋詞》及孔凡禮所編《全宋詞補輯》二書，共收 1493 家詞人 21055 闋詞。詞調的總數，依梁啟勳《詞學銓衡》記載:『康熙二十六年(1687)，萬樹撰《詞律》二十卷，所收共六百六十調。後來徐本立著《詞律拾遺》,續收一百六十五調。杜文瀾又著《詞律補遺》,又補收五十調。共計八百七十五調。杜著成於同光間，在最近八十年間，大概沒有能自度新腔創作新調的人了，八百七十五調的數字應是準確的。』(註二十三)王力:『《詞律》共載六百六十調，一千一百八十餘體，《拾遺》補載一百六十五調，一百七十九體，又補體三百十六，連《詞律》原書合計，共八百二十五調，一千六百七十餘體。」(同註十六，頁 692)《宋詞大辭典》收詞譜 1380 種(含同調異名)，扣除同調異名共計八百七十五調。』」,第六屆詞彙語意學會議，2005 年 4 月 21-22 日，廈門大學主辦，刊於 http://cls.hs.yzu.edu.tw/。

七、 該網站之網址：http://cls.hs.yzu.edu.tw/CM/login.htm

八、 該網站之網址：http://cls.hs.yzu.edu.tw/ORIG/

九、 參見拙作,〈文學地理時空資訊系統設計與應用：以蘇軾詩為例〉,第二屆數位地球國際研討會，2004 年 5 月 27-28 日，中央研究院、國家實驗研究院、中國文化大學主辦