LANGUAGE AND LINGUISTICS 4.3:631-647, 2003
2003-0-004-003-000064-1

唐宋詩之詞匯自動分析及應用*

俞士汶　　胡俊峰
北京大學

本文介紹了唐宋詩之詞匯的自動提取、自動分析技術及其在古代詩詞電腦輔助研究中的一些應用的實例。

文章著重介紹了利用統計的方法對唐宋詩語料進行詞匯獲取的具體演算法及步驟。給出了「共現度」、「結合強度」等統計參數的計算方法，並與傳統的「互信息」方法進行了比較。

在對全唐詩（481 萬字）、宋代部分名家詩（160 萬字）的語料完成切分及詞性標注的基礎上，系統又對唐宋詩詞匯的使用情況進行了統計分析。提取了詞匯共現信息、詞匯對仗信息、作者詞匯特徵信息以及詞匯引用的時代變遷分布等多種統計信息。

在此基礎上，進一步對唐宋詩詩句的相似性檢索、詩人風格檢索、意象索引技術等研究課題進行了探索。

關鍵詞：詞匯自動提取，搭配，詩句的相似性檢索，意象索引，中國古代詩歌

1. 引言

在人們談論「數位化生存」(Being Digital) 的今天，中華古籍的數位化似乎已經算不上新潮。但回想 10 多年前，已有一批學者爲中華古籍的整理、研究、出版的現代化奔走呼號、身體力行，筆者由衷地欽佩他們的先知先覺。正是受海峽兩岸從事古籍電子化研究的學者的影響和鼓勵，北京大學計算語言學研究所自 1993

* 本文涉及的研究工作是北大計算語言學研究所開發的「中國古代詩詞電腦輔助研究系統」的一部分。「中國古代詩詞電腦輔助研究系統」的開發得到 1998 至 1999 年度中國國家社科基金項目「古詩電腦輔助研究系統及其應用」（項目號：98BYY022）的支援，也結合了北大計算語言所同北大古文獻研究所、同台灣元智大學合作項目的成果。這個項目現在正得到北大 985 項目的有力支援。筆者僅向給予支援的部門和單位及給予過指導的陸儉明、孫欽善、周先慎、張鳴等老師表示衷心的感謝。在典籍數位化研究過程中，我們同台灣元智大學羅鳳珠老師、台灣中研院謝清俊老師等學者進行了較多的交流。從交流中我們強烈地感受到炎黃子孫珍愛中華文化的拳拳深情，他們的學識與奉獻精神對我們是鞭策，是鼓勵。筆者對他們在許多方面給予的幫助致以誠摯的謝意。

年以來一直將「中國古代詩詞的電腦輔助研究」作爲研究所的重要研究方向之一（劉岩斌等 1997）。

中國的詩歌藝術源遠流長，詩歌作爲一種最接近口語的大眾化文學形式，在漢語文化的成長、演變與傳播中佔有著極其重要的地位，因而對中國古代詩詞的研究歷來是漢學研究的熱點之一。不過，關於詩歌的傳統研究往往著眼於作品的人文及藝術特徵，研究者多半依靠自身良好的文化修養與「強聞博記」，憑感悟直接把握作品的內涵，這種研究用於詮釋作品的美學及人文意義自有其無可替代的優勢，但在對作品的語言全貌進行同時代橫向或歷史縱向的細密分析時，往往就顯得力不從心。雖然也有一些文章對某些詞匯、典故的使用及意義進行分析和研究，通常只能是及其一點不及其餘，很難以時代或作者爲單位來分析其語言的風格、特徵及演變情況，更不要說從中總結出可信的規律了。

筆者一直認爲，古籍整理對我們電腦專業工作者來說是一個新天地，又是一個引人入勝的可以大有作爲的天地。運用在現代漢語信息處理研究中建立的計算語言學的理論、方法與技術，可以對古代漢語的語言現象進行深層次的研究。同時也會促進古代漢語語法研究與現代漢語語法研究的縱向結合，推動現代漢語語法研究的深入和語言信息處理技術的發展。正是基於這樣的理念，北大計算語言所在開發「中國古代詩詞的電腦輔助研究系統」時，既不滿足於紙本到電子文本的轉換和簡單的檢索、統計功能，也不追求多媒體的外在形式或眼前的商業價值。而是把力量集中於能發揮自己優勢的深加工和知識發現的領域。

經過我們七八年之努力，終於取得了一些階段性成果，「古詩自動注音軟體」是其一例（穗志方等 1998）。本文介紹的是另一項成果，即使用統計方法對唐宋詩的詞匯進行分析所取得的一些結果。「詞」（這裡的「詞」相當於英語的 word，有別於「宋詞」的「詞」）在漢語中是一個難以嚴格定義的模糊的概念。類型語言學將漢語歸入「孤立語」，並認爲以單音節語素（其書面記錄符號即漢字）構成的「詞」（以下簡稱「單字詞」）成爲詞匯的主體是漢語的主要特徵之一。對於古漢語，這個論斷是接近實際的。隨著社會生活的發展，漢語要保持和豐富其表達能力而又要避免不斷擴充漢字的數量，必然要引進由兩個以上語素或漢字構成的「詞」（以下簡稱「多字詞」）。「多字詞」在漢語的發展史上究竟是什麼時代開始出現的顯然是一個饒有興趣的問題。不過，本文研究唐宋詩中的「多字詞」並非僅僅出於理論研究的興趣。筆者從研究過程體驗到，從唐宋詩中提取詞匯（詞匯包括「單字詞」和「多字詞」，對於自動提取技術研究來說，只考慮「多字詞」），並以詞匯爲基礎進行深層次的分析，相對於以「字」或「字串」爲基礎的簡單匹配，可以取得更多的有實際應用價值的成果。

北大計算語言學研究所自 1998 年至 1999 年建造了「中國古代詩詞電腦輔助研究系統」(原型系統),在這個系統中已收納全唐詩(481 萬字)、宋代部分名家詩(160 萬字)共計 640 多萬字的語料。以這個語料庫爲實驗材料,本文介紹了詞匯的自動提取、自動分析技術及其成果在詩句、風格相似性檢索等領域的應用。除引言外,以下還有 4 節。第二節到第四節是本文的核心內容,分別介紹唐宋詩詞典的電腦輔助構造、唐宋詩語料的自動切分與詞匯分析、基於詞匯的詩歌相似性分析。第五節是結語,討論了詞匯分析在相關領域的應用和古代詩詞電腦輔助研究的若干新課題。

2. 唐宋詩詞典的電腦輔助構建

本文避開「詞」或「多字詞」的定義,但從實際應用的需要出發,根據具體的目標確定了「詞」特別是「多字詞」在統計意義上的提取標準。

詞的提取與分析當然離不開對詞義的理解。領域專家對詞義的理解自有優勢。但是,許多現代漢語中的詞(如:可以、上學等)在古詩詞中還不是詞,而古詩中的一些詞(如:弱冠、小槽等)由於社會環境的變化,在現代漢語中已經很少這樣使用了。即使同爲古漢語,不同歷史時代的詞也會有很大的區別。僅僅依靠領域專家是很難進行大規模調查與分析的。統計手段的引入,就能夠有一個相對客觀的標準來判定古漢語中的詞。

2.1 唐宋詩之詞匯的自動提取

2.1.1「互信息」統計抽詞模型的考察與「共現度」概念的提出

「互信息」常被用於未登錄詞的統計發現,其公式如下:

$$I(xy) = \ln \frac{P(xy)}{P(x) * P(y)}$$

其中,P (x)、P (y) 分別爲漢字 x、y 在語料庫中出現的概率;P (xy) 是漢字 x、y 在語料庫中的鄰接同現概率。需要說明的是,在原始的互信息公式中,隨即變數 x,y 是無序的,而用於詞匯提取時則必須考慮其序關係。在具體應用時這個公式可以轉化爲以下形式:

$$(1)\ I(xy) = \ln \frac{N * f(xy)}{f(x) * f(y)}$$

其中，N 爲一個與語料庫大小有關的常數，f (x)、f (y) 則表示漢字 x、y 在語料庫中出現的次數；f (xy) 表示漢字串 xy 在語料庫中出現的次數。直觀的來講，如果兩個字在一起出現的頻度越高，這兩個字組成的二字組是一個詞的可能性就越大。

在中文裡，有許多字使用時都是在合成詞中出現。在這種情況下，「互信息」就不能很好地反映漢字串的成詞特性。

舉例來看：「縉雲」在全唐詩中出現了 11 次，「縉」出現了 32 次，「雲」出現了 12698 次。如果單計算「縉雲」的互信息值是很低的。考察發現，在「縉」出現的 32 次中「縉紳」出現了 21 次。這意味著，「縉」這個字在全唐詩詩句中以雙字詞出現的概率很大。如果在確認「縉紳」是一個詞的前提下，「縉」的另外 11 次出現都是同「雲」在一起共現。在這種情況下有理由認爲「縉雲」很可能是一個詞。

根據漢語這一特點，筆者提出「共現度」概念。

首先，要定義相對共現度 R (x|y) 和 R (y|x)：

$$(2)\ a.\ R(x|y) = \frac{f(x\ y) * \ln f(x\ y)}{F(x)}$$

$$b.\ R(y|x) = \frac{f(x\ y) * \ln f(x\ y)}{F(y)}$$

其中：f (xy) 爲字組 xy 在語料中的出現次數，F (x) 是 x 在語料中的「自由出現」次數。所謂「自由出現」次數 F (x) 是指：

如果已知語料中 xz_i (i =1, …, n) 爲詞，那麼：

$$(3)\ x \text{ 的「自由出現」} F(x) = f(x) - \sum f(xz_i)$$

顯然：R (x, y) 一般不等於 R (y, x)。在語料中字組 xy 的共現度定義爲：

$$(4)\ \text{共現度 } C(x, y) = R(x, y) + R(y, x)$$

這就是說，在一個二元字組中，如果有一個相對共現度達到了某個閾值，就可以認爲該字對另外那個字有著很好的親和度以至於這個二元字組有可能是詞！

共現度的計算是一個遞進的過程，運算過程相對複雜，下面僅通過一組實例來比較共現度與互信息的差別。

〈表 1a〉列出了全唐詩語料中含有「徊」字的頻度大於 2 的所有二字組。根據統計信息不難確定「徘徊」爲一個候選詞。再根據以上公式對其餘字串的共現度重新進行計算，得到〈表 1b〉，在確定「俳徊」、「裴徊」爲候選詞後得到〈表 1c〉。隨著「徊」字的自由出現頻度逐步遞減，一些原本統計特徵不十分明顯的詞就顯現出來。

字串	字串頻度	前字	前字頻度	後字	後字頻度	互信息	共現度
徘徊	143	徘	145	徊	**190**	1.29016715	4.25187981
俳徊	9	俳	13	徊	**190**	.936331482	.706768991
裴徊	17	裴	828	徊	**190**	-2.58174354	.125636521
低徊	12	低	1181	徊	**190**	-3.28515390	.075490433
從徊	2	從	4734	徊	**190**	-6.46532234	.008295742
得徊	2	得	8096	徊	**190**	-7.00192194	.008162814

〈表 1a〉按共現度排序的含有「徊」字的二字組

字串	字串頻度	前字	前字頻度	後字	後字頻度	互信息	共現度
俳徊	9	俳	13	徊	**47**	.936331482	1.9419006
裴徊	17	裴	828	徊	**47**	-2.58174354	1.0829494
低徊	12	低	1181	徊	**47**	-3.28515390	.65969309
從徊	2	從	4734	徊	**47**	-6.46532234	.02978846
得徊	2	得	8096	徊	**47**	-7.00192194	.02966684

〈表 1b〉減去「徘徊」中結合出現的「徊」字頻度後的共現度情況

字串	字串頻度	前字	前字頻度	後字	後字頻度	互信息	共現度
低徊	12	低	1181	徊	21	-3.28515390	1.4451954
從徊	2	從	4734	徊	21	-6.46532234	.06630685
得徊	2	得	8096	徊	21	-7.00192194	.06618524

〈表 1c〉減去「徘徊、裴徊」中結合出現的「徊」字頻度後的共現度情況

2.1.2「結合強度」的提出

「詞」的直觀意義常被解釋爲「使用頻繁，結合緊密」。可到底怎麼樣才叫「結合緊密」,「結合緊密」這一特徵是否能有一個統計學上的解釋呢？回答應該是肯定的。

在考察唐宋詩語料時注意到這樣的一個現象：如果兩個字能構成一個詞的話，這兩個字在一句詩裡同現時，一般會以緊密相鄰的方式出現。例如：如果「功名」是一個詞，那麼在實際語料中就很少會出現「功*名」兩個字隔開出現的句子。爲了考察這是否是一個普遍的語言現象，筆者作了如下的統計試驗：對語料庫中在同一句中同現的所有二字組進行統計（如果一句詩有 7 個字，則會生成 6+5+4+3+2+1 共 21 個字組，包含了相鄰與不相鄰的所有情況），對每一個二字組求出其總的出現次數 W 以及相鄰出現次數 M。

二字組的「插入率」S = 1 – M/ W

$$\text{(5) 成詞的「結合強度」} D = \left(\frac{M}{W}\right)^{2} * \ln (M)$$

需要說明的一點是，在唐宋詩這類特定的語體中，五言的 2、3 字、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爲是相鄰出現。

統計分析表明，「結合強度」的引入能有效的排除那些結合鬆散的字組；而作爲一個考察標準，在唐宋詩語料庫中，當字串的出現頻度大於 20[1] 且其結合強度大於 1 時，在超過 90% 的概率的意義上可以確定該字串是一個詞。當詞頻高於 14 時，如果以結合強度作爲選詞標準，得到的抽詞效果遠優於互信息統計的結果。由於這一標準與語料庫大小基本上沒有關係，語料庫的規模越大，利用這一標準的效益就越高一些。

〈表 2〉列出了按結合強度排序的部分字串的統計情況。注意到其中有些詞的互信息值比較低。

[1] 這裡串頻大於 20 以及下文提到的詞頻高於 14 是一個同語料庫大小相關的閾值。在這裡可以簡單理解爲一個經驗值。

字串	字串頻度	前字	前字頻度	後字	後字頻度	共現度	結合強度	互信息
洞庭	431	洞	1768	庭	2699	2.4474747	6.010199	-2.76105231
平生	590	平	3929	生	9687	1.3466639	6.007944	**-4.52347747**
洛陽	473	洛	1667	陽	5624	2.2656053	6.005762	-3.34340359
參差	429	參	870	差	641	7.0456570	6.005336	-.618984145
煙霞	417	煙	5306	霞	1742	1.9183422	6.004254	-3.45521132
芙蓉	424	芙	493	蓉	468	10.683971	5.937182	.251836301
楊柳	445	楊	1798	柳	3338	2.3222110	5.936898	-2.9584022
別離	480	別	6557	離	3578	1.2801800	5.924366	**-4.24598123**
蒼蒼	378	蒼	2113	蒼	2113	2.1234169	5.872586	-2.82575304
鴛鴦	368	鴛	633	鴦	391	8.99526353	5.84438439	0.03998522
殷勤	345	殷	692	勤	703	5.7810700	5.84354441	-.700318095
蕭條	344	蕭	2069	條	1218	2.62066150	5.8406416	-2.34806415

〈表 2〉按結合強度遞減排序的部分字串的統計情況

2.1.3 字串的使用頻度與多維統計模型的構建

現在我們有了三種並不完全獨立而又的確各具特色的統計標準「頻度」、「共現度」(互信息)、「結合強度」來考察一個二字組是否是詞。那些同時具備三種較好的特性的二字組在 95% 的概率意義上是詞，三種特性都不好的幾乎肯定不是詞。在實際應用中是採用多維度數值擬合的方法建立了一個多維統計抽詞模型，經過最終人工校對，在查準率爲 56% 的情況下，查全率達到了 89.3%，取得了較爲理想的效果。〈表 3〉顯示了通過該統計模型提取的按綜合指數遞減排序的部分詞的實例。

2.2 唐宋詩多字詞的標準的確立及唐宋詩詞典的人工標注

統計分析表明，在唐宋詩的語料中除了一些常規的雙聲疊韻詞、專有名詞外，已經出現了大量現代意義上的多字詞如：故鄉、梳洗、寂寞、天真等。但相對而言，另外一些詞如：青山、秋水、白髮、歸雁等，按照一般的觀點可能不被認爲是詞，但卻表現出了綜合指數高的統計特徵。進一步分析表明，這類詞匯在古詩詞中往往表現有特定的隱喻義，從字面義與隱喻義的分離來看確實具有了詞的基本特徵。因此，在構建古詩詞詞表的過程中，字面義與隱喻義的分離是作爲是否

成詞的一個重要標準來考慮的。因此，在系統中，青山、白雲、明月、落花、遠遊、苦吟都是作爲一個詞收錄的。

這樣做也會帶來另一方面的副作用，那就是人爲的掩蓋了明月、秋月與月光之間的天然聯繫。當然，通過建立一部語義詞典也許可以較好地解決這個問題，目前系統在實現的時候只選擇了對複合詞的內部結構進行標注的方法，當對複合詞的內部結構進行標注時，使用了北大計算語言學研究所的《現代漢語語法信息詞典》（俞士汶等 1998）及《現代漢語語素庫》（俞士汶等 1999）的成果。

字串	字串頻度	前字	前字頻度	後字	後字頻度	互信息	結合強度	綜合指數
蜘蛛	30	蜘	41	蛛	80	1.856679	3.4011973	1.6855152
渭水	149	渭	534	水	11666	-4.08980	4.8722692	1.6848631
驄馬	87	驄	208	馬	5491	-2.93141	4.3649868	1.6707346
芍藥	46	芍	60	藥	1427	-.977952	3.5162348	1.6701811
桃花	384	桃	1786	花	12131	-4.38953	3.6997160	1.6693602
梨花	127	梨	345	花	12131	-3.85179	4.0438845	1.6681254
山川	349	山	19186	川	2145	-5.12667	5.2375762	1.6665757
故園	272	故	3652	園	2296	-3.78507	5.4444924	1.6605610
咸陽	127	咸	422	陽	5624	-3.28453	4.6951441	1.6551203
枯槁	43	枯	795	槁	60	-.460406	3.7612001	1.6494450
青山	652	青	6669	山	19186	-5.63603	5.5898346	1.6461514
鶗鴂	20	鶗	27	鴂	23	3.115485	2.9957322	1.6427110
獼猴	25	獼	29	猴	47	2.552512	2.9760316	1.6424351

〈表 3〉按綜合指數遞減排序的部分二字組的實例

在目前建立的詞表中，包含了 43164 條多字詞，7277 條單字詞。在通過校對後的詞表中，不單包含了詞性、結構的信息，而且還附帶了插入率的統計信息，這爲今後進行詞語自動切分創造了條件。

〈表 4〉顯示了部分詞條的標注結果。表中 rm、dm 分別表示人名、地名。pn 表示一個偏正結構的名詞。

詞條	頻度	詞性	插入率	綜合指數
箜篌	36	n	0	2.7775933
縱橫	209	n	.033492822	2.7592618
千載	289	mq	.069204152	2.7490443
相思	726	v	.037190082	2.7396797
平生	590	n	.030508474	2.7300329
擾擾	87	a	0	2.7296483
朝廷	102	n	.0196078431	2.7280782
嫦娥	50	rm	0	2.7179832
巫峽	151	dm	.0596026622	2.7032908
嫋嫋	84	a	.0238095238	2.6786934
明月	881	pn	.0953461750	2.6740700

〈表 4〉部分詞條的標注信息

3. 唐宋詩語料的自動切分與詞匯分析

由於古代詩的語體與音律節奏的限制，古詩語料的自動切分相對較爲簡單。應用基於詞典的最長匹配演算法再輔以基於結合強度的切分消歧策略，其抽檢正確率高於 99%，基本上能達到實用要求。

下面重點就古詩語料的詞匯分析加以介紹。

3.1 詞匯的共現與對仗

同一個詞匯在不同的時代、不同的話語群落往往會具有不同的意義。尤其是作爲古漢語來講，脫離了當時的語境，要想對詞匯的意義達到精確的把握是比較困難的。

從另一個角度來看，詞匯的用法又反映了詞匯的意義。如果能夠瞭解一個古漢語的詞當時使用的上下文語境，這對正確理解這個詞的意義能夠起到啓示的作用。基於這種考慮，系統對全唐詩、部分宋詩語料的詞匯進行了共現及對仗分析，全唐詩部分分析結果如〈表 5〉所示。[2]

[2] 表中 5、6 的詞條對仗關係完全由計算機根據詞匯在詩歌中的位置關係自動統計而得，其中有些並不是完全是嚴格意義上對仗詞。當然可以進一步通過平仄關係進行過濾。考慮到本研究的主要目的還是發掘詞匯之間語義上的相關關係，即使不是嚴格意義上的對仗詞，如果在對仗句的相關位置反復出現（在語料比較小的情況下包括出現一次）也被認爲是對下一步研究有用的線索予以保留。

詞條	同現詞條	同現次數
草色	青青	10
長安	少年	13
長聲	千載	17
朝朝	暮暮	16
朝雲	暮雨	20
成式	張希	16
城南	嗚咽	17
城下	石頭	9
池塘	春草	9
遲遲	春日	10

詞條	對仗詞條	對仗次數
戰士	武皇	17
楊柳	芙蓉	19
天地	日月	10
嗚咽	千載	17
渭水	善琴	16
萬壑	千峰	8
天子	將軍	19
天涯	海內	8
天上	人間	45
桃花	柳葉	11

〈表5〉

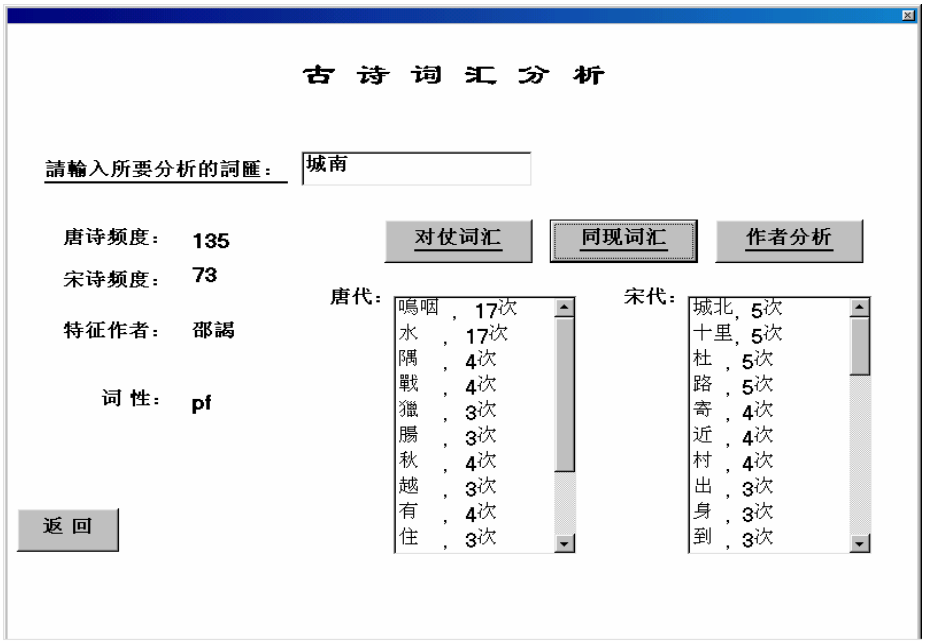
a. 全唐詩中的部分詞匯同現條目　　b. 全唐詩中的部分詞匯對仗條目

觀察發現，詞匯的每一個穩定的對仗、同現關係大多反映了該詞匯的某個特定義項。詞匯的對仗、同現關係的改變，甚至在一定程度上反映了一個詞的語義變遷。〈圖 6〉顯示了「城南」一詞的同現分析結果。

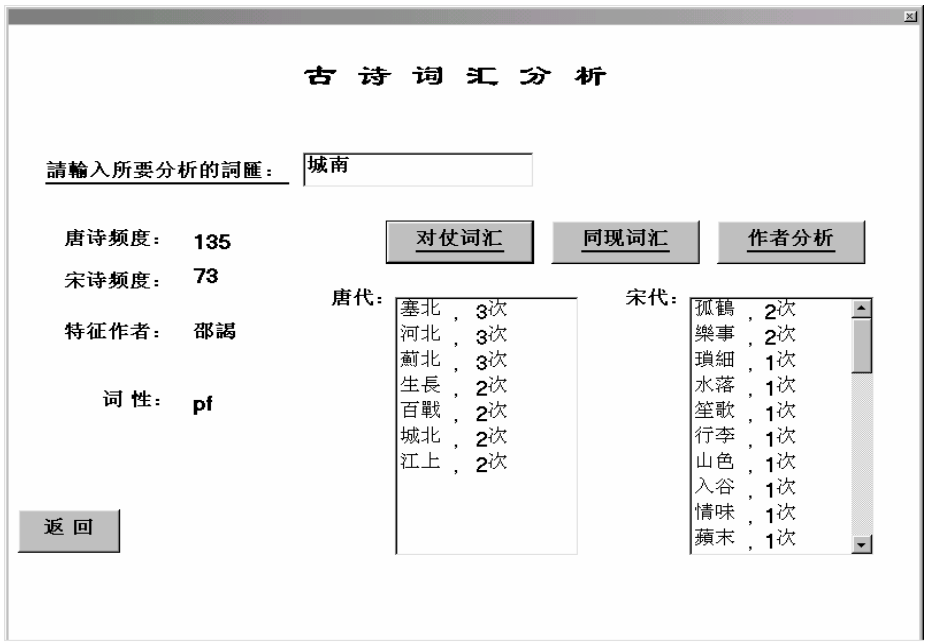
可以看出，「城南」一詞在唐代有悲傷與戰爭的意象，而到了宋代則一般用來表示郊外，有閒適、快樂的意象。有了這種認識，就不難對下面詩歌的意境進行正確地理解與把握了：

雜曲歌辭　遠別離二首（選一）
令狐楚（唐）
玳 織 鴛 鴦 履，金 裝 翡 翠 簪。
畏 人 相 問 著，不 擬 到 城 南。

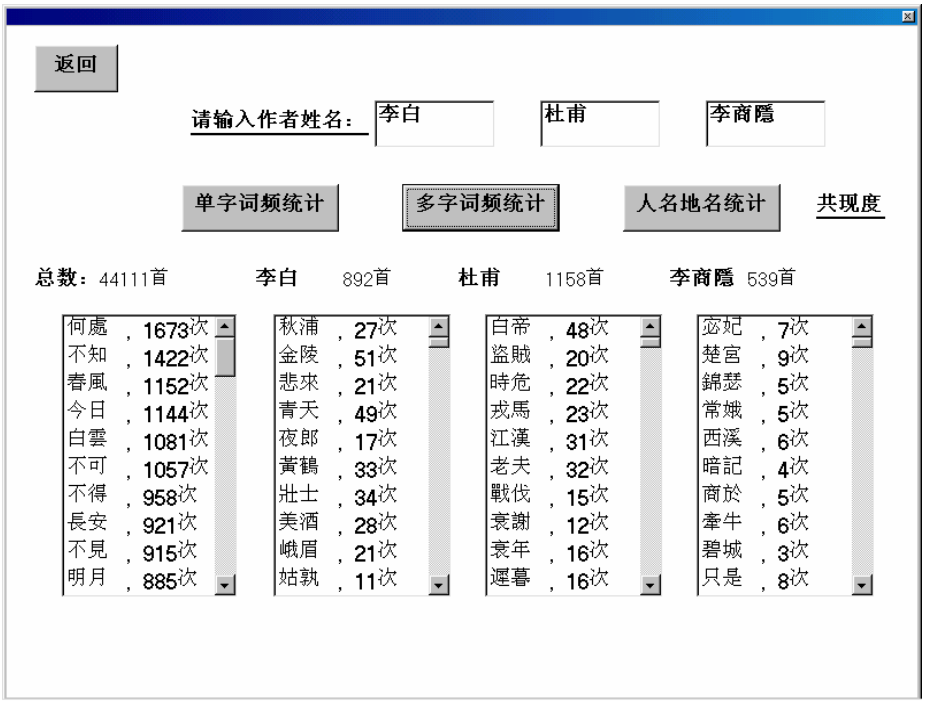
梨　花
陸游（宋）
嘉陵江色嫩如藍，鳳集山光照馬銜。
楊柳梨花迎客處，至今時夢到城南。



〈圖 6a〉「城南」一詞的同現詞匯



〈圖 6b〉「城南」一詞的對仗情況



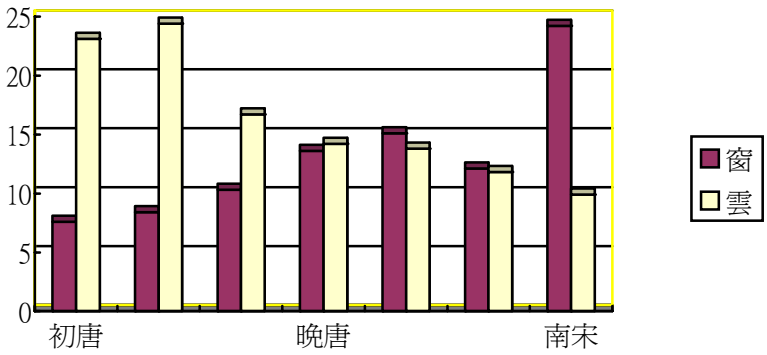
〈圖 7〉不同作者的特徵詞匯對比情況

3.2 詞匯與作者的相關分析

作者的用詞特徵是作者風格分析的一個重要方面，在以往的研究中曾經對唐詩作者的用字進行過統計分析，建立了字一級的索引（欒貴明 1997）。在自動分詞的基礎上，可以進一步完成詞匯級的統計分析。共現度（參考公式 (2)）的引入使得作者詞匯的特徵分析能夠更加直接地反映作者的用詞特徵。〈圖 7〉顯示了系統提供的作者用詞對比統計的功能。由於系統在辭典中標注了相應的人名、地名信息，所以相應的人名、地名引用情況也就很容易能統計出來。這些功能爲進行作者風格研究提供有益的參考。

與作者詞匯特徵相對應的就是詞匯作者特徵，這兩個概念在數學上可以很方便的相互轉換，但在實際應用上卻有完全不同的含義。特定詞匯被不同作者的引用情況不僅反映了作者個人的差異而且在一定程度上反映了詞匯在不同歷史時期的使用情況。特定詞匯在不同歷史時期的使用往往又折射出特定時期的人文及歷史情況。〈圖 8〉顯示了「窗」和「雲」在不同歷史時期的引用分布情況。從總體而言，隨著時代的發展，在詩歌中「窗」使用的越來越多，而「雲」卻出現的越

來越少，是不是因爲詩人在室內呆的時間越長，到外面看雲的機會就越少了呢？在這一領域的工作還有待進一步發掘與探索。



〈圖 8〉「窗」、「雲」在唐宋詩中引用的時代變遷分布[3]

4. 基於詞匯的詩歌相似性分析

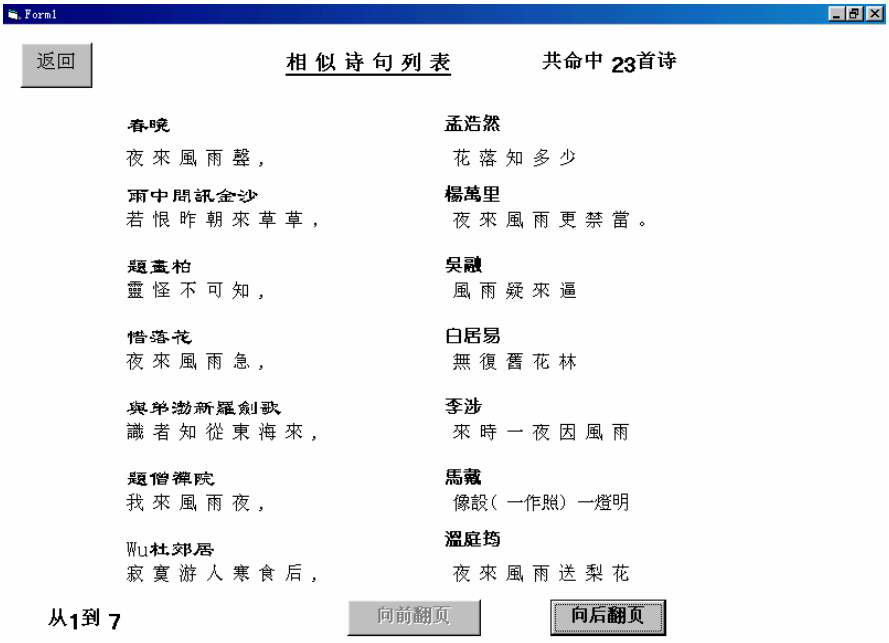
4.1 古詩詞詩句的相似檢索

漢語的基本組成元素是「字」。但在進行詩句的相似性分析的時候，完全按照「字」進行匹配並不能夠取得較好的結果。例如：「白雲留永日」與「白日既雲暮」有較多的字相同，但整句意義上的相似性卻不強。相對而言，如果按詞進行相似性匹配，效果可能就會好一些。〈圖 9〉顯示了對詩句「夜來風雨聲，花落知多少。」的相似性檢索結果。

4.2 詩句的風格的相似檢索

作者的特徵用詞在一定程度上反映了作者的寫作風格。基於作者特徵用詞而進行的作者風格檢索，爲不同作者之間的作品相關分析研究提供了一個新的視點。針對特定詩人（收錄詩需要超過 500 首）的風格相似檢索在效果上要優於單句之間的相似檢索，〈圖 10〉顯示了在杜甫的詩中檢索出的一部分風格類似李白的詩句。

[3] 需要說明的是圖中縱軸所標明的值並不是「雲」、「窗」出現的絕對頻度，而是在該時代的詩歌中出現的頻度與所收錄的詩歌數的比值並經過統計平滑調整得到的量。由於與本文主題不直接相關，詳細步驟從略。



〈圖 9〉詩句「夜來風雨聲，花落知多少。」的部分相似性檢索結果

標題	上句	下句
橫吹曲辭後出塞五首	少年別有贈	含笑看吳鉤
橫吹曲辭前出塞九首	丈夫四方志	安可辭固窮
絕句四首	兩箇黃鸝鳴翠柳	一行白鷺上青天
故右僕射相國張松齡	上君白玉堂	倚君金華省
短歌行贈王郎司直	王郎酒酣拔劍斫地歌莫哀	我能拔爾抑塞磊落之奇才
陪王侍禦同登東山最高頂宴姚通	清江白日落欲盡	復攜美人登絲舟
乾元中寓居同穀縣作歌七首	我行怪此安敢出	拔劍欲斬且復休
飲中八仙歌	飲如長鯨吸百川	銜杯樂聖稱世賢
樓上	天地空搔首	頻抽白玉簪
奉送蜀州柏二別駕將中丞命赴江陵	報與惠連詩不惜	知吾斑鬢總如銀
奉寄李十五秘書二首	行李千金贈	衣冠八尺身
晚晴	江虹明遠飲	峽雨落餘飛
赤甲	笑接郎中評事飲	病從深酌道吾真

〈表 10〉杜甫詩中檢索出的部分風格類似李白的詩句

4.3 意象索引技術的探索

意象索引技術無論是在古漢語還是現代漢語的研究中都是一個極富挑戰性的研究領域。如果能夠讓電腦正確分辨哪一首詩是表達歡樂意象，哪一首詩是表達悲傷意象，並能夠按照意象的「強度」進行排序，這無疑將會對詩人、作品的風格研究、對詩歌的篇章分析乃至句法分析起到有益的參考。

鑒於文本的篇章理解技術目前還遠未達到實用階段，因此本研究依然定位在詞匯一級。在人工選擇了與某個特定意象相關的一些特徵詞匯後（例如：針對「悲傷」意象選擇：悲、苦、愁、淒涼、自憐等），再根據詞匯的共現、聯想網路搜索到與之相關的詞匯（如：蹉跎、蕭然、浮生、西風、殘燈、柳色等）共 304 條，在此基礎上運用神經網路演算法對每一首詩的「悲傷度」進行打分，並據此建立起以「悲傷」爲主題的意象索引。

由於缺乏對所收錄的詩歌進行手工分類的基礎，這種分類方法的查全率有待進一步驗證。就已有實驗結果來看，對於「意象」值較高的短詩（八行以內），實際檢索效果較好。

5. 結語

運用計算語言學手段對中國古詩詞進行研究是一個全新的領域，除了上面討論的部分應用外，基於詞及同義詞擴展的檢索、電腦自動對詩、電腦輔助寫詩填詞（羅鳳珠 1999）等工作都有待進一步的開展。相關的研究成果將能夠對古詩詞、古漢語領域的研究提供有益的幫助。從另外一個角度來看，對古漢語的研究也爲現代漢語的研究提供了一個新的視角，有利於從一個新的角度來審視現有的一些概念與問題。

在本項研究中課題組成功地將一些現代計算語言學技術根據古詩詞語言的特點加以改造，取得了一些有益的成果。更主要的是，通過本項研究，系統積累了唐宋詩語料庫及有關中國古詩詞的語言信息知識庫，爲今後進一步的研究奠定了良好的基礎。

總體而言，對古詩詞的分析加工目前還只限於詞匯與詞匯共現一級，一些相關的應用如：詞匯自動切分，相似句檢索技術等都是建立在這個基礎上的。語言本身還有更加高層次的結構成分（如：句法結構，篇章結構等）。因此，僅在詞匯一級進行分析是不夠的，其相關應用的效果自然也受到一定的局限。可以預見，在這一領域的研究工作還有很長的路要走。同時，由於古詩文體天然的特質（文

體簡短，大量使用隱喻義，題材相對單純等），有利於開展古詩詞的篇章分析、意象分析、認知心理等課題的研究，希望這些工作能爲計算語言學的發展、爲典籍數位化研究、爲漢學研究奉獻綿薄。

引用文獻

俞士汶, 朱學鋒, 王惠, 張芸芸. 1998.《現代漢語語法信息詞典詳解》。北京：清華大學出版社。

俞士汶, 朱學鋒, 李峰. 1999.〈現代漢語語素庫的開發及應用〉,《世界漢語教學》1999.2:38-45。

劉岩斌, 俞士汶, 孫欽善. 1997.〈古詩詞研究的電腦支援環境的實現〉,《中文信息學報》1997.1:27-36。

穗志方, 俞士汶, 羅鳳珠. 1998.〈宋代名家詩自動注音研究及系統實現〉,《中文信息學報》1998.2:44-53。

羅鳳珠, 李元萍, 曹偉政. 1999.〈中國古代詩詞格律自動檢索與教學系統〉,《中文信息學報》1999.1:35-42。

欒貴明. 1997.《全唐詩索引》。天津：天津古籍出版社。

[Received 16 March 2003; revised 20 April 2003; accepted 26 May 2003]

俞士汶
北京大學計算語言學研究所
中國 100871 北京市
yusw@pku.edu.cn

胡俊峰
北京大學計算語言學研究所
中國 100871 北京市
Hujf@pku.edu.cn

Word-based Statistical Analysis of Chinese Ancient Poetry

Shiwen Yu Junfeng Hu
Peking University

This paper is concerned with the automatic extraction of multi-character words from a corpus of ancient Chinese poetry and with some applications at word level. A detailed description of the word-extraction algorithm is given and compared with the mutual-information method. The study has been based on a 4.8 million-character corpus of Tang Dynasty poetry and a 1.6 million-character corpus of Song Dynasty poetry. Statistical analysis to date includes collocation, word-to-author analysis information, etc. Further research would include sentence-similarity retrieval and a semantic index.

Key words: automatic word extraction, collocation, sentence similarity retrieval, semantic index, ancient Chinese poetry