

國際中國語言學學會第 21 屆年會(IACL 21)的工作坊——歷代語言知識庫的建構與應用
蔣經國國際學術交流基金會《歷代語言知識庫建置計畫》(2009)

唐宋詩詞語言知識庫的建構動機與方法

羅鳳珠

台灣元智大學中國語文學系
gefjulo@saturn.yzu.edu.tw

摘要

要談建構詩詞語言知識庫的動機，不能不先談類書與中國詩歌發展的關係。魏文帝曹丕詔令王象、劉劭等人集五經群書，依「以類相從」方式編列《皇覽》，已被公認為是類書之首。《昭明文選》，以主題分類，堪稱是後代類編詩文集的發端。爾後，類書在隋至初唐達到高峰，延至清代，官、私修類書的編纂，成為各朝各代宣示文教興盛的象徵。唐代歐陽詢有鑑於「《流別》、《文選》，專取其文，《皇覽》、《遍略》，直書其事。文義既殊，尋檢難一。」¹的缺失，改以「撰其事且文……比類相從……故事居其前，文列於後，俾夫覽者易為功，作者資其用。」²兼具「類事」與「類文」功能的體例編纂《藝文類聚》。

文人寫詩，本來是感物緣情而起，到了唐代，卻逐漸起了變化。中國詩歌的發展，在唐代達到高峰，固然有其文體發展的諸多因素，但是與唐代沿襲隋代科舉考試制度，在各科之中又以詩賦所取的進士地位最高，無異於是對唐詩的發展起了推波助瀾的作用。其中講究聲律、對仗、典故等修辭技巧的律詩，在唐代被訂為科考指定的詩體，加上帝王的喜愛與有心推廣，詩歌到了唐代，多了科考應制與應答酬唱的功能，此時事文兼備，既方便學子準備科考，又符合文人雅士唱和之需的類書，便順勢而起。王昌齡《詩格·論文意》一文指出：「凡作詩之人，皆自抄古今詩語精妙之處，名為隨身卷子，以防苦思。作文興若不來，即須看隨身卷子，以發興也。」³王昌齡指的是如白居易《白氏六帖》、杜嗣先《兔園策府》之類的書。由此可見，唐代很多文人加入類書的編纂，或自編小類書以供隨時查檢之用，無形之中形成類書與文學觀念相互滲透的現象，也使唐詩的發展與類書存在著相互依存的微妙關係。

最早論及唐代近體詩發展與類書關係密切的學者是聞一多。聞一多於〈類書與詩〉文中直言：「假如選出五種書，把它們排成下面這樣的次第：《文選注》、《北堂書鈔》、《藝文類聚》、《初學記》、初唐某家的詩集。我們便看出一首初唐詩在構成程序的幾個階段。」⁴又說：「唐初是一個大規模徵集詞藻的時期，我所謂徵集詞藻者，

¹. (唐)歐陽詢編纂，《藝文類聚》，北京：中華書局，1965年11月第1版，〈序〉之第27頁。《流別》是西晉摯虞《文章流別》。《遍略》是南朝梁徐勉《華林遍略》。

². 同註1。

³. 張伯偉編著，《全唐五代詩格彙考》，南京：江蘇古籍出版社，2002年4月第1版，第164頁。

⁴. 聞一多撰，傅璇琮導讀《唐詩雜論》，上海：上海古籍出版社，1998年12月第1版，〈類書與詩〉，

實在不但指類書的纂輯，連詩的製造也是應屬於那個範圍裡的。」⁵賈晉華也認為：「總集和類書的大量修撰，對於初唐詩歌的普及化和技術化也起了一定的促進作用。」⁶除此之外，類書的一般功用，可見於錢謙益為沈自南《藝林彙考》一書所做的序：「書成（指沈自南《藝林彙考》），就正於蒙叟。叟告之曰：子之書，有四便焉。便於好學者一，便於不好學者一，而便於蒙叟者二。」⁷這一段話指出了類書廣泛精選精華知識的特性而對於多數讀者所產生的功用。

類書與科舉考試到了清代都戛然而止，其中因由，不在本文討論範圍，不多贅述。歷代類書編纂的形式、內容、方法、規模，各有不同，但為了方便查閱檢索的目的是一致的。

隨著資訊科技的發達，很多學術領域都引用資訊科技工具作為研究的輔助工具，中國文學領域與資訊科技交叉學科研究的起步比較晚。文學與自然科學不一樣，自然科學比較客觀。善於處理數據的電腦，早已成為自然科學研究時儲存、記憶、計算、分析、歸納的有利工具。相對於自然科學而言，文學以主觀的成分居多，其中尤以詩詞為最。因此，結合文學與圖書資訊學知識，建立一個接近人類知識體系的詩詞語言知識庫，期使文學之語義概念分類能提高圖書資訊檢索及讀者服務的品質與效率。文學研究者也能利用所建構的語言知識庫，開創新的文學研究方法與研究議題，兼能提供語言學、資訊科學研究使用，這是本研究的主要動機。

詩詞語言知識庫屬於歷代語言知識庫建置計畫的一部份，建置的目的，是在歷代類書的基礎上，結合資訊科技的優勢，截長補短，建立一個包含「詞彙語義知識」、「詞彙語法知識」、「詞彙語義概念分類體系」的詩詞語言知識庫，預期達到的目標有六種：一、建構唐宋詩詞語言知識庫；二、修正歷代詩詞類書與工具書分類的不足；三、滿足文學研究的需要；四、建立詩詞的共時語言與歷史語言知識庫；五、提供語言學與圖書資訊科學研究的需要；六、建立以語義概念類別檢索的功能，提升文學文本資訊檢索的品質。

本研究所建構的「唐宋詩詞語義知識庫」與語義網、語義概念辭典的差別有四項：一、所處理的語言性質不同：「唐宋詩詞語義知識庫」所處理的語言是特定文體古典韻文的文本語言，韻文語言屬於人工語言⁸；語義網、語義概念辭典所處理的是現代漢語的自然語言。二、所處理的語言範圍不同：「唐宋詩詞語義知識庫」所處理

第 4 頁。

⁵. 聞一多撰，傅璇琮導讀《唐詩雜論》，上海：上海古籍出版社，1998 年 12 月第 1 版，〈類書與詩〉，第 7 頁。

⁶. 賈晉華著，《唐代集會總集與詩人群研究》，北京：北京大學出版社，2001 年 6 月第 1 版，第 478 頁。

⁷. (清) 錢謙益著、(清) 錢曾箋注、錢仲聯標校，《牧齋有學集》，上海：上海古籍出版社，1996 年 9 月第 1 版，第 691-693 頁。

⁸. 王力著，《王力詩論》，南寧市：廣西人民社，1988 年出版。《王力詩論》之〈略論語言形式之美·詩的語言〉，引葉聖陶給王力的書信：「詩之句型，大別為二。一為平常的句型，與散文及口頭語言大致不異。一為特殊句型，散文絕不能如是寫，口頭亦絕無此說法，可謂純出於人工」。

的語言範圍是既有的、固定的，可以盡收，不會增減；語義網、語義概念辭典所處理的語言是沒有範圍的，會隨著時空不斷增減變化。三、語義概念的類別是有限的：唐宋詩詞是過往既存的文獻，唐宋詩詞語言知識庫所收錄的語言僅限於唐宋詩詞文本使用的詞彙，因此所建構的語義類別是有限的，不能概括人類文明的所有知識體系。四、語義標記與分類的困難度不一樣：詩詞語言的語義可以區分為一般用語、特殊用語、情境用語三種⁹，特殊用語與文學修辭技巧結合，屬於語用的層次，情境用語需與全篇章共現的詞彙之語義相結合，才能顯示準確的語義，屬於語用及語境的層次，因此語義標記的困難度比自然語言的現代漢語高。

以文本建置語言知識庫的方法，首先要對文本進行詞彙切分，其次再以詞彙為單位進行語義、詞性標記，再其次是進行語義概念分類。詩詞文體與語體文最大的不同是詩詞受到嚴格的格律限制，因此有固定的句法，詞彙的切分比語體文容易，但也正因為受到格律的限制，使得詩詞語言成爲一種「特異的語言形式」¹⁰，是一種「純出於人工」¹¹的語言，使得詩詞語言的語義標記格外的困難。

本研究所採用的詩詞語言知識庫建置方法，以內容分析法爲主。論述的內容包含：一、詩詞文本之詞彙切分方法；二、依據文本內容特性分類的語義概念架構；三、語義標記方法設計：人工標記與自動標記；四、語義概念分類架構與層級關係；五、語義標記的範例；六、詞性標記的類別。

⁹ 羅鳳珠，〈植基於中國詩詞語言特性所建構之語義概念分類體系研究〉，圖書與資訊學刊，3:3=78，2011年8月，頁63-86。

¹⁰ 袁行霈著，《中國詩歌藝術研究》，臺北市：五南圖書公司，1989年出版。第2、3-24、25-56、57-72頁。

¹¹ 同註8。