佛典跨語文獻的詞彙庫及索引建立之輔助方法研究

*釋法源
**李家名
***黃乾綱 教授

*中華佛學研究所佛學資訊學程
**中華佛學研究所網資室
***台灣大學工程科學與海洋工程學系

摘要

在佛典文獻的研究上，不論是漢語或是跨語的文獻，經常需要建立詞彙庫或索引；這是一項重要卻很耗費時間的整理工作，以往只能靠人力慢慢去做。如何運用資訊技術來加速佛典文獻詞彙與索引的工作，就是本文所要加以研究的。

本研究係運用「詞彙抽取（term extraction）」技術，對於佛典古代典籍與近代文獻作自動的詞彙抽取，來建立漢語的詞彙庫。另外；在佛典跨語詞彙抽取方面，則是利用平行語料庫（parallel corpus）及統計法，自動建立跨語詞彙庫。自動建立的詞彙，導入專家的驗證機制後，即為完成的漢語及跨詞詞彙庫；詞彙的抽取皆由電腦程式自動執行，人工的部份只需作詞彙的驗證與確認。建立完成的詞彙庫，尚可運用資訊檢索技術（information retrieval）來自動產生超連結（hyperlink）索引（index）用語索引（concordance），如此可以節省大量的時間與人力。

以上的詞彙庫與索引技術，還可以整合發展為「網路服務」。運用網際網路提供：線上佛學多語(Multi-Lingual)專有詞彙之檢索及抽取服務。將提供使用者在線上進行中文詞彙的檢索及抽取。線上佛學詞彙關聯性分析服務，將結合統計分析以及資料探勘演算法的工具，讓使用者可透過此介面在線上，研究單語(Mono-Lingual)詞彙，或多語詞彙之間的關連性。線上佛學多語專有詞彙之出處比對服務，可以在線上針對所收集的佛典，提供研究學者查詢不同語文佛學詞彙的出處，並可透過介面進行比對，可幫助有意研究譯本之間用詞差異之學者。

關鍵字：抽詞、詞彙庫、資訊檢索、平行語料庫、跨語檢索

1. 前言

在傳統的佛典文獻的研究上，不論是漢語或是跨語的文獻，經常需要建立詞彙庫或索引；這是一項重要卻很耗費時間的整理工作，以往大都只能靠人力慢慢的彙集與整理。本研究係運用「詞彙抽取（term extraction）」技術，來加速佛典文獻詞彙與索引的整理工作，提供作為研究學者的輔助工具。

本研究所發展的文獻處理輔助工具，並不是要取代人力，而是可以用於節省人力與時間的輔助用途。文獻研究學者可以運用此項工具，將文獻資料迅速加以整理並產生初步的對照索引（index）或用語索引（concordance），再由研究學者的專業經驗去篩選出重要的索引資料。或是將現有的索引資料，運用此項輔助工具來做校對驗證，以確保資料的精確性。

建立完成的詞彙庫及索引資料，還可以自動產生超連結（hyperlink），來連結全文（full text）或多語(Multi-Lingual)文獻資料，如此可以大幅增進文獻研究的效益。另外；可以在線上針對所收集的佛典文獻，提供研究學者查詢不同語文佛學詞彙的出處，並可透過介面進行比對，可幫助有意研究譯本之間用詞差異之學者。

2. 研究方法與步驟

本研究係運用「詞彙抽取（term extraction）」技術，對於佛典古代典籍與近代文獻作自動的中文詞彙抽取，來建立漢語的詞彙庫。另外；在佛典跨語詞彙抽取方面，則是利用平行語料庫（parallel corpus）及統計法，自動建立跨語詞彙庫。本研究除了發展漢語佛典專有名詞的抽取技術之外，更希望藉由這個機會建立跨語佛典的平行語料庫。跨語佛典的平行語料庫建立，是一項很有價值的工作。整理後的詞彙資料，不僅可用於佛教典籍的研究，更可用於當時的各地區語言學的研究。

2.1 中文抽詞

統計抽詞的方法長久以來都是中文抽詞的主要方式之一[1-4]。本文的抽詞技術主要參考簡立峰 [1999] 年所做大量新聞資源抽詞工作中的演算法[1]。如圖 1 所示，主要的辨別依據是如果一個字串（Lexical Pattern）的左右兩邊出現字的種類越多，該處應被視為斷開點的特性就越強，也就是越應該切斷。相反的，如果某個字的下一個字的可能性永遠都只有一種，那這兩個字就必定不可切斷。

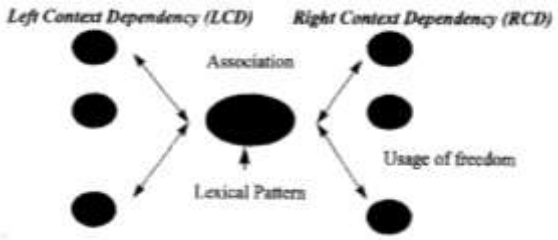

圖 1. 中文抽詞方法概念圖

以上述的標準，進一步產生下列參數與演算法，來進行抽詞篩選的依據：

參數：fx, fy, fz, lRl, fxB

fx: 每個字串(例：abcde)出現的頻率
fy: 字串去掉最後一個字後(例：abcd)，出現的頻率
fz: 字串去掉第一個字後(例：bcde)，出現的頻率
lRl: 字串兩邊出現單字的種類
fxB: 字串兩邊出現最多次的單字的次數

公式：設定篩選值：n1, n2, n3，並抽取符合以下條件者。

a. lRl > n1
b. fxB / fx < n3
c. AE > n2　　*AE = fx / fz+fy-fx

其中公式 a 與公式 b 擇一個執行，再拿其結果執行公式 c。(a or b) and c
本研究的中文抽詞使用大量統計的方式計算結果，所使用的語料庫，包含三個檔案群：

◎《大正藏》全文
使用 Cbeta《大正藏》全文檔，228mb 約一億單字
◎《卍續藏》全文
使用 Cbeta《卍續藏》全文檔，98mb 約 4 千萬字
◎佛學學報全文
華岡、中華佛學學報、中華佛學研究、佛學研究中心學報，共 57mb 約 1700 萬字

以文獻種類來分，《大正藏》與《卍續藏》屬於佛學原典，而學報則屬於當代文獻資源。再經過上述的抽詞演算程式處理後，建立成約有 27 萬個詞條的詞彙庫。

2.2 跨語抽詞

由於佛教經典是透過不同的傳抄、釋譯而流傳；所以，今日可收集到的佛教經典涵蓋中文、巴利文、緬甸文、梵文、藏文等亞洲各國語言。佛典典藏的研究者，常常需要比對不同語文的經典，因此不光是翻譯的問題，就是單純跨語查詢都變的困難。本研究者進行的跨語研究，係採用平行語料庫以統計方法來做佛典跨語詞彙的抽取。

平行語料庫(Parallel Corpus)係指收錄的多語文獻中，跨語對應的文獻。其對應的方式，可能是以 “段落” 為對應的單位，甚至是以 “句” 為對應的單位。我們以《妙法蓮華經》說明佛典經文中，目前收錄的梵、漢、英三語經文中取同一段落，並以段落比對的方式呈現，詳如表 1。

表 1.《妙法蓮華經》散文部份梵、漢、英對照範例

《漢文》
如是我聞，一時佛住王舍城耆闍崛山中， 與大比丘眾萬二千人俱， 皆是阿羅漢，諸漏已盡，無復煩惱，逮得己利，盡諸有結，心得自在。

《梵文》
ekasmin samaye bhagavan rajagṛhe viharati sma gṛdhakuṭe parvate/
mahata bhikṣusaṃghena sardhaṃ dvadasabhir bhikṣusataiḥ/
sarvair arhadbhiḥ kṣiṇasravair niḥklesair vasobhutaiḥ suvimuktacittaiḥ
suvimuktaprajñair rajaneyair mahanagaiḥ kṛtakṛtyaiḥ kṛtakaraṇiyair
apahṛtabharair anuprapta-svakarthaiḥ pariksiṇabhavasaṃyojanaiḥ
samyagajñasuvimuktacittaḥ sarvacetovasitaparamaparamitapraptair
abhijñanabhijñatair mahasravakaiḥ /

《英譯》
Thus have I heard. Once upon a time the Lord was staying at Râgagriha, on the Gridhrakûta mountain, with a numerous assemblage of monks, twelve hundred monks, all of them Arhats, stainless, free from depravity, self-controlled, thoroughly emancipated in thought and knowledge, of noble breed, great elephants, having done their task, done their duty, acquitted their charge, reached the goal; in whom the ties which bound them to existence were wholly destroyed, whose minds were thoroughly emancipated by perfect knowledge, who had reached the utmost perfection in subduing all their thoughts; who were possessed of the transcendent faculties; eminent disciples,

利用平行語料庫以統計方式產生跨語詞典的步驟如下：
1. 先利用程式建立梵、漢、英的平行語料庫。
2. 利用漢語詞彙找出所有的包含該詞彙的漢文佛典語句。
3. 取出對應的梵文佛典語句。
4. 統計所有語句中的梵文佛典詞彙。依計算次數的高低排序。
5. 去除在整個梵文佛典中出現機率高的辭彙。
6. 剩餘的高頻詞彙即有可能是對應詞彙。

利用上述方法可先找出最可能對應的辭彙，再利用成對比較的方式來尋找其他可對應的辭彙。

3. 索引資料處理

本研究所發展的文獻處理工具，可以用來提昇佛典索引整理工作的效率。運用此項輔助工具可以節省研究學者人力與時間，將文獻資料迅速整理並產生初步的對照索引（index）或用語索引（concordance），研究

學者再以專業經驗去篩選出重要的索引資料。或是將人力已整理好的索引資料，運用此項工具來做校對驗證，以確保資料的精確性。

3.1 對照索引

有關對照索引（index）的建立，係運用前述中文抽詞所建立的 27 萬餘個漢語詞彙（term）的佛典詞彙庫；並利用平行語料庫產生跨語詞彙（cross-language term），來做比對而產生文獻研究所需的對照索引資料。包含詞彙、跨語詞彙、對應文脈（context）及出處頁數（page），其中出處頁數，可自動建立超連結（hyperlink），與 Cbeta 電子佛典做全文連結，方便使用者查詢原文。

若以鳩摩羅什大師所譯《妙法蓮華經》[1]的索引建立為例，只要將經文電子檔交給電腦程式處理，即可自動抽詞並與佛典詞彙庫比對，而取出專有的中文詞彙及梵語詞彙，如【三昧】samadhim、【妙法】saddharma、【陀羅尼】dharani 等。專有詞條所對應的前後文脈（context）及出處頁數（page）則運用資訊檢索技術，亦由電腦程式自動產生，並自動建立與 Cbeta 電子佛典的全文超連結（hyperlink）。詳如表 2 所示：

表 2.《妙法蓮華經》索引資料範例

Term / Cross-language	context	Page
【三昧】samadhim		
	昧。淨光明三昧。淨藏三昧。不共三昧。日旋三昧...	第 0055 頁 b 欄 02 行
	等亦欲勤修行之。行此三昧。乃能見是菩薩色相大...	第 0055 頁 c 欄 02 行
	經。淨眼菩薩。於法華三昧。久已通達。淨藏菩薩...	第 0060 頁 b 欄 05 行
	于座。身不動搖。而入三昧。以三昧力。於耆闍崛...	第 0055 頁 b 欄 18 行
	喪功於本無。控心轡於三昧。則忘期於二地。經流...	第 0062 頁 b 欄 27 行
	後。得一切淨功德莊嚴三昧。即昇虛空。高七多羅...	第 0060 頁 b 欄 29 行
	。佛力無所畏。解脫諸三昧。及佛諸餘法。無能測...	第 0005 頁 c 欄 17 行
【妙法】saddharma		
	那笈多後所翻者。同名妙法。三經重沓。文旨互陳...	第 0001 頁 b 欄 22 行
	。來白於大王。我有微妙法。世間所希有。若能修...	第 0034 頁 c 欄 15 行
	。常懃求智慧。說種種妙法。其心無所畏。我於伽...	第 0041 頁 b 欄 22 行
	師利。可與相見。論說妙法。可還本土。爾時文殊...	第 0035 頁 a 欄 22 行
	。經千萬億劫。說無漏妙法。度無量眾生。後當入...	第 0039 頁 c 欄 14 行
	十中劫。廣為眾生說於妙法。恒河沙眾生得阿羅漢...	第 0035 頁 a 欄 05 行
【陀羅尼】dharani		
	邊菩薩。得百千萬億旋陀羅尼。三千大千世界微塵等...	第 0062 頁 a 欄 26 行
	方便陀羅尼。得如是等陀羅尼。世尊。若後世後五百...	第 0061 頁 b 欄 09 行
	致。轉不退法輪。得諸陀羅尼。即從座起。至於佛前...	第 0036 頁 b 欄 10 行
	。擁護此法師故。說是陀羅尼。即說呪曰。阿梨那梨...	第 0059 頁 a 欄 08 行
	以見我故。即得三昧及陀羅尼。名為旋陀羅尼百千萬...	第 0061 頁 b 欄 07 行
	菩薩得解一切眾生語言陀羅尼。多寶如來於寶塔中讚...	第 0055 頁 a 欄 04 行

研究學者可以運用此項輔助工作快速的建立索引資料，再以專業經驗將索引資料做篩選。或是將已整理好的索引資料，運用此項工具來做校對驗證，以確保資料的精確性。建立完成的索引資料，除可作為書面附錄外；亦可方便作為電子出版的索引連結。出處頁數（page）可自動產

[1]大正新脩大藏經 第九十冊 No. 0262《No. 262 妙法蓮華經》

生超連結（hyperlink），與 Cbeta 電子佛典做全文連結，便於原文的快速查詢與對照。

3.2 語用索引

語用索引（concordance）的建立，係同時運用中文抽詞與資訊檢索等技術，自動產生文獻的語用索引資料，並提供使用者選擇排序方式，如前文排序（左方排序 left-sorted）、後文排序（右方排序 right-sorted）等。若以《妙法蓮華經》的詞彙「一乘」語用索引建立為例，只要輸入所需詞彙，電腦程式則列出「一乘」的語用索引，由使用者選擇排序方式，即可進一步提供研究學者作佛典文獻的詞彙研究運用。詳如圖 2 所示：

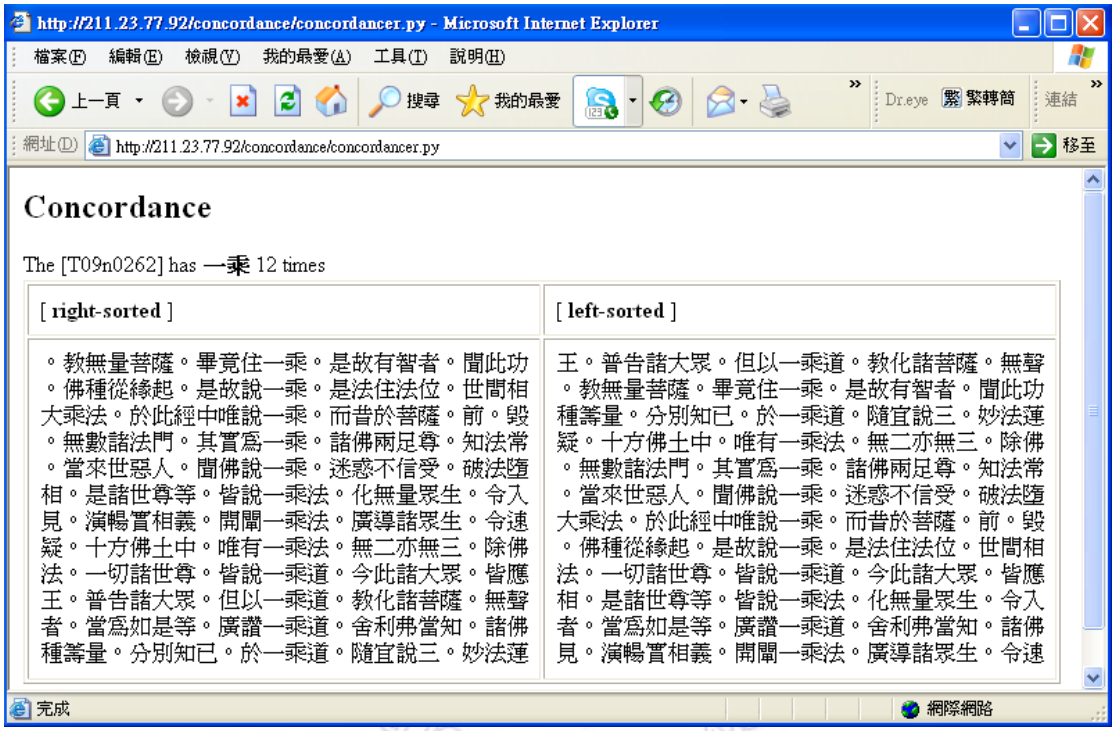

圖 2. 語用索引呈現介面

4. 詞彙研究運用

在文獻詞彙的研究方面，係提供線上（on-line）的文獻比對服務，研究學者查詢不同語言的佛學詞彙之出處，並可透過電腦介面進行比對，其呈現方式係將語用索引（concordance）進一步發展為比對介面，可幫助有意研究譯本之間用詞差異之學者。如以《法華經》的譯本用語研究為例，可以比對西晉月氏國竺法護大師所譯《正法華經》[2]、後秦龜茲國鳩

[2] 大正新脩大藏經 第九十冊 No. 0263 《No. 263 正法華經》

摩羅什大師所譯《妙法蓮華經》[3]、隋代崛多笈多大師所譯《添品妙法蓮華經》[4]等三個漢文譯本，並將梵文本做對照研究。

以上三個《法華經》的漢文譯本，其年代以西晉時竺法護大師所譯的《正法華經》為最早，其次是後秦鳩摩羅什大師所譯的《妙法蓮華經》，較晚的譯本則是隋朝崛多笈多大師所譯的《添品妙法蓮華經》，三個譯本若對照梵文本來看，其文脈頗為一致；不過，在翻譯所使用的詞彙上，竺法護大師的《正法華經》與鳩摩羅什大師的《妙法蓮華經》用詞差異甚大，而崛多笈多大師所譯的《添品妙法蓮華經》，其用詞大致上沿用鳩摩羅什大師的詞彙。

以下舉一個《妙法蓮華經》偈頌中「聞佛說一乘」[5]的用詞為例，其漢語譯文為：

> 當來世惡人，**聞佛說一乘**，迷惑不信受，破法墮惡道
> 有慚愧清淨，志求佛道者，當為如是等，廣讚一乘道

此偈頌與崛多笈多大師所譯《添品妙法蓮華經》的偈頌[6]譯文是一樣的，用「佛」做為一乘教的能說者。從年代的先後，及其他語用索引來看，可以推測崛多笈多大師應該是沿用鳩摩羅什大師的譯本詞彙，詳如圖 3。

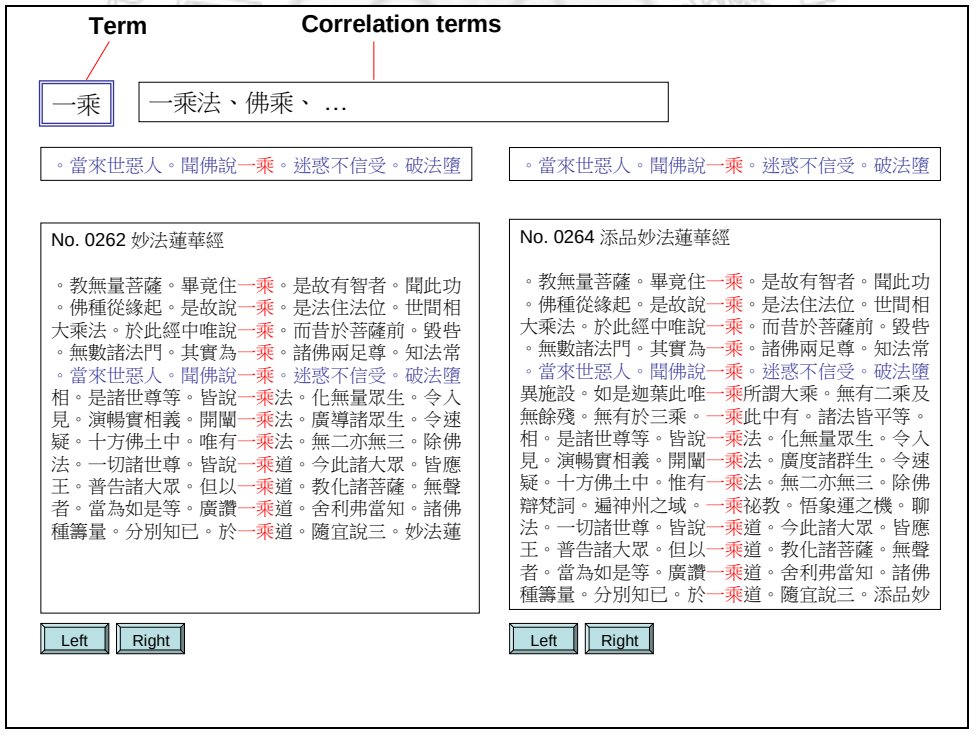

圖 3. 詞彙用法比對呈現介面

[3] 大正新脩大藏經 第九十冊 No. 0262《No. 262 妙法蓮華經》
[4] 大正新脩大藏經 第九十冊 No. 0264《No. 264 添品妙法蓮華經》
[5] 大正新脩大藏經 第九十冊 No. 0262《No. 262 妙法蓮華經》 (卷 1) T09, p0010b
[6] 大正新脩大藏經 第九十冊 No. 0264《No. 264 添品妙法蓮華經》(卷 1) T09, p0143b

然而，竺法護大師的《正法華經》所翻譯的用詞，卻譯為「聽察如來，一乘之教」[7]，其用法有相當的不同，詳列如下：

> 若當來人，而說此法，**聽察如來，一乘之教**，設復覩見，諸最勝名誹謗斯經，便墮地獄，假使有人，慚愧清淨，發心志願，來尊佛道

若參照梵語 sd-kn 本法華經[8]，梵文羅馬轉寫原文：

anāgate'dhvani bhrameyu sattvāḥ sūtraṃ kṣipitvā narakaṃ vrajeyuḥ //142//
lajjī śucī ye ca bhaveyu sattvāḥ saṃprasthitā uttamamagrabodhiṃ /
viśārado bhūtva vademi teṣām ekasya yānasya anantavarṇān //143//

約可推測係同一偈頌的不同譯法，其它的詞彙用法，也可以從圖 4 的介面做比較：

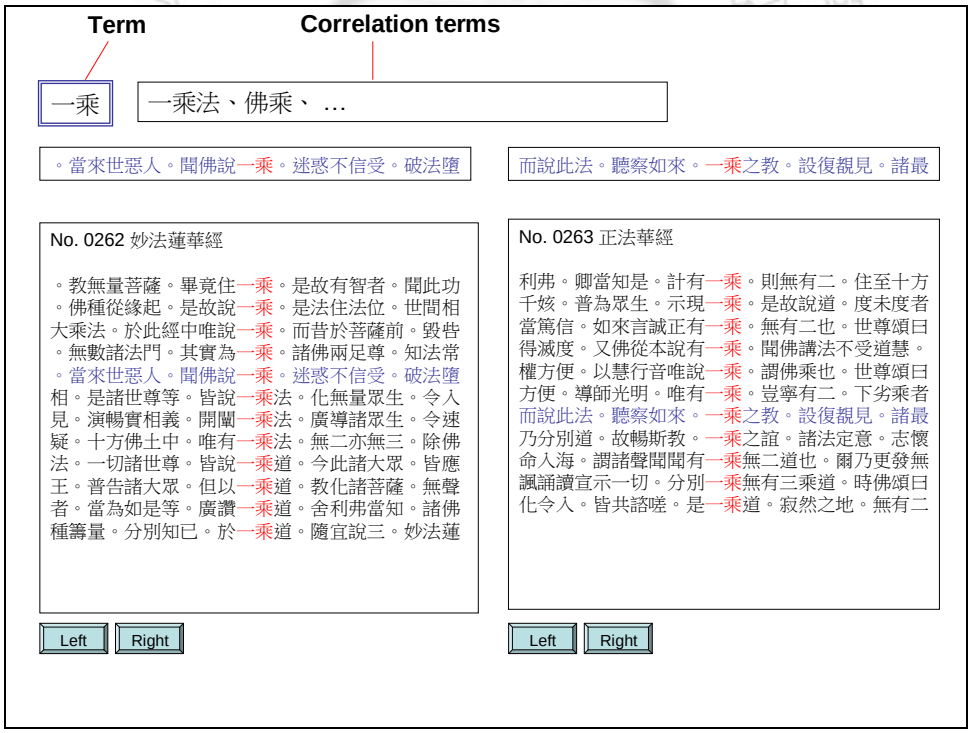

圖 4. 詞彙用法比對呈現介面

[7]大正新脩大藏經 第九十冊 No. 0263 《No. 263 正法華經》(卷 1) T09, p0073a
[8] Prof. H. Kern and Prof. Bunyiu Nanjiao *saddharmapuṇḍarikasūtram*

以上的詞彙研究運用，可以提供研究學者一個方便的使用介面，對有興趣的詞彙找出不同譯本的用法，再進一步去分析詞彙用法的時代性，或演變軌跡，讓研究者對於文獻詞彙的使用，有更深入的了解。

5. 未來發展方向

本研究所建立的詞彙庫與索引技術，未來將以整合發展成「網路服務」為目標。運用網際網路提供：線上佛學多語(Multi-Lingual)專有詞彙之檢索及抽取服務。將提供使用者在線上進行中文詞彙的檢索及抽取。線上佛學詞彙關聯性分析服務，將結合統計分析以及資料探勘演算法的工具，讓使用者可透過此介面在線上，研究單語(Mono-Lingual)詞彙，或多語詞彙之間的關連性。

另外，將發展資訊集成代理人（Information Aggregation Agent）技術。將所抽取的古典詞彙與現代詞彙，進行集成處理，結合時間、分類（人、事、時、地、物）及跨語等關連，提供研究學者更深入的語言或詞彙研究的應用。詳如圖 5 所示：

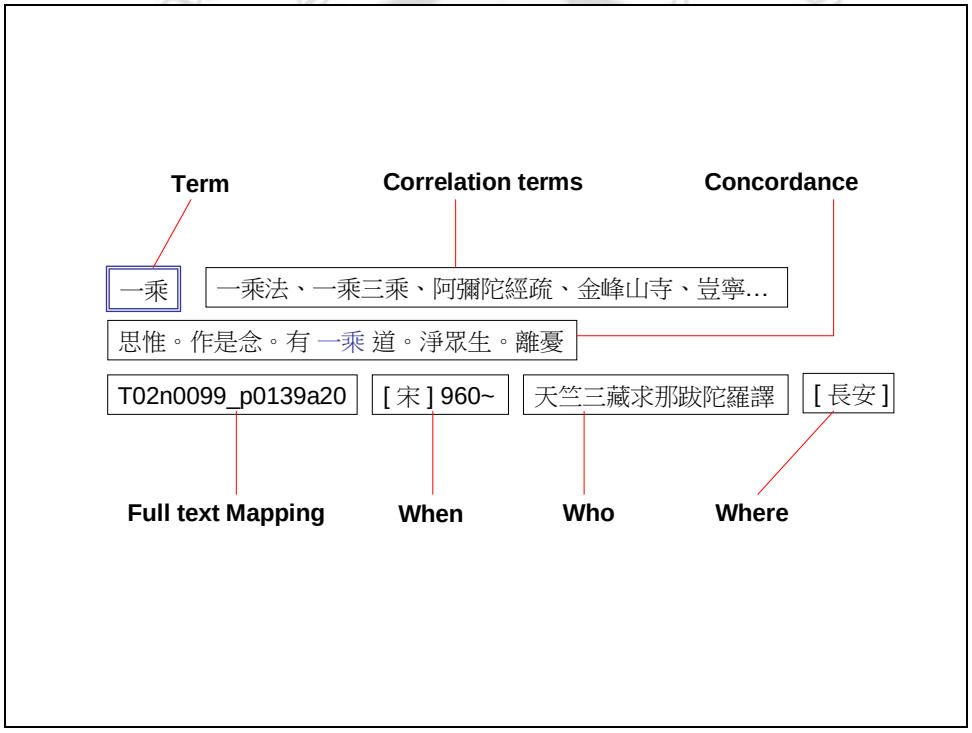

圖 4. 詞彙資訊集成概念

6. 結論

本研究所發展的文獻處理工具，並不是要取代人力，而是可以用於節省人力與時間的輔助用途，幫助研究學者處理耗費時間的文獻整理工

作。研究學者可以運用此項工具，迅速產生初步的對照索引（index）或用語索引（concordance）資料，再以專業經驗去篩選出所需要的索引資料。或是已整理好的索引資料，運用此項工具來做校對驗證，以確保資料的精確性。

資訊技術的在人文領域的發展與運用，藉由輔助工具的建立，應該可以幫助完成繁複的資料處理。再與專家學者的智慧及專業經驗互相搭配，或許是現代人文與資訊科技結合的一種模式，從資料的整理到資訊的整合，說不定可以為人文領域找到一個新的研究方式。

參考文獻

[1] Chien, L.-F. "PAT-tree-based adaptive keyphrase extraction for intelligent Chinese information retrieval" Information Processing and Management 35 (1999) 501-521。

[2] Wu, D. and Xia X., "Large-scale automatic extraction of an English-Chinese translation lexicon" Machine Translation 9:3-4 (1994) 285-313.

[3] Kwong, O.Y. and Tsou, B.K. 2001. "Automatic Corpus-Based Extraction of Chinese Legal Terms." In Proceedings of the 6th Natural Language Processing Pacific Rim Symposium (NLPRS 2001), Tokyo, Japan.

[4] Tsou, B.K., Tsoi, W.F., Lai, T.B.Y., Hu, J. and Chan, S.W.K. 2000. LIVAC, "A Chinese Synchronous Corpus, and Some Applications." In Proceedings of the ICCLC International Conference on Chinese Language Computing, pages 233-238, Chicago.

[5] Kwong, O.Y., Tsou, B.K., Lai, B.Y., Luk, W.P., Cheung, Y.L. and Chik, C.Y. "A Bilingual Corpus in the Legal Domain and its Applications" Workshop on Language Resources in Asia, Proceedings of the Sixth Natural Language Processing Pacific Rim Symposium (2001), 39-46

[6] Kwong, OY, Tsou, BK, Lai, TBY "Alignment and extraction of bilingual legal" Terminology, 2004 - ingentaconnect.com 10:1 (2004), 81–99.

[7] Cheung, L, Lai, T, Luk, R, Kwong, OY, Sin, KK, Tsou, BK, "Some considerations on guidelines for bilingual alignment and terminology extraction" International Conference On Computational Linguistics 18 (2002), 1-5.

[8] Fujii, A., Ishikawa, T., Lee, J.H., "Term Extraction from Korean Corpora via Japanese" CompuTerm 2004: 3rd International Workshop on Computational Terminology (COLING 2004), 71-74

[9] Yang, C., Luk, J., "Automatic generation of English/Chinese thesaurus based on a parallel corpus in laws" Journal of the American Society for Information Science and Technology 54:7 (2003), 671 – 682

[10] Chen Keh-Jiann, Wei-Yun Ma, 2002, "Unknown Word Extraction for Chinese Documents", Proceedings of Coling 2002, pp.169-175.