行政院國家科學委員會補助專題研究計畫 ■ 成 果 報 告
□期中進度報告

中文詞彙及跨語詞彙抽取技術在佛典數位典藏上之研發與應用

The Development and Application of the Chinese and Cross-Lingual Term Extraction for Buddhist Digital Archives

計畫類別：■ 個別型計畫　□ 整合型計畫
計畫編號：NSC 95－2422－H－002－018－
執行期間：95 年 3 月 1 日至 96 年 2 月 28 日

計畫主持人：黃乾綱
共同主持人：歐陽彥正、釋惠敏（郭敏芳）
計畫參與人員：釋法源(唐國銘)、李家名

成果報告類型(依經費核定清單規定繳交)：□精簡報告　■完整報告
本成果報告包括以下應繳交之附件：
□赴國外出差或研習心得報告一份
□赴大陸地區出差或研習心得報告一份
□出席國際學術會議心得報告及發表之論文各一份
□國際合作研究計畫國外研究報告書一份

處理方式：除產學合作研究計畫、提升產業技術及人才培育研究計畫、列管計畫及下列情形者外，得立即公開查詢
□涉及專利或其他智慧財產權，□一年□二年後可公開查詢

執行單位：國立台灣大學工程科學及海洋工程學系
中　華　民　國　九十六年　五月　二十八日

計畫中文摘要：

本研究計劃的目的在長期目標在研究中文詞彙及跨語詞彙抽取技術在佛學典籍上之研究與應用。以中華佛學研究所的佛學典藏為基礎，我們長遠的目標是要建立一個能夠支援佛學研究學者進行佛學典籍的研究及教學的重要平台。由於詞彙的研究是邁向知識架構建立的必經之路，因而在 95 年的計畫中，是以佛典中文詞彙及跨語詞彙抽取技術的研究與應用作為主要的研究對象。

本計劃中，開發了以 Suffix Array 為基礎的統計抽詞方法，能夠提供比 PAT-tree 基礎的統計抽詞方法更佳的空間與時間效率。我們並且分析統計抽辭應用於佛典的古典文獻以及現代文獻時所呈現的現象。並將各項統計數據呈現以提供進一步的研究。最後並且建置了網站服務，並且提供程式碼以及各項報告的下載。

本計劃以資訊技術及工具，便利大規模佛學典藏中專有辭彙之研究。藉由詞彙抽取技術的發展，內容開發計畫也同時可以得到便利的參考資料，也能協助研究者在線上進行中文或跨語詞彙的關聯性分析，因而能夠加速對於佛學研究知識架構的建立。且本計劃的研究對象，是以佛典的內容為主，而且其中又包含了韻文及非韻文。其研究成果，有極大的可能性可以延伸至其他古典文、史、哲資料集。如果本計劃的長期目標－建立「佛學典籍研究平台」，能夠完成，則可穩固我國在東亞哲學思想研究的地位，提升我國文哲研究的國際能見度。

中文關鍵詞：統計抽辭、跨語詞彙，平行語料庫、PAT-tree、Suffix Array

計畫英文摘要：

The object of this research project is to study The Development and Application of the Chinese and Cross-Lingual Term Extraction for Buddhist Digital Archives. Our research based on the Buddhist Digital Archives of Chung Hwa Institute of Buddhist Studies. In the long term, we hope to build a Buddhist Archive Research Platform. As the study of term extraction is the basis for the establishment of Buddhist Knowledge Hierarchy, therefore, we target our research direction in term extraction.

In this project, we developed a Suffix Array based Term Extraction Algorithm, which is more space and time efficient than PAT-tree based Term Extraction Algorithm. And we also provide the analysis of term extraction method applied in Buddhist Archives, including classical literatures and modern literatures. These analyses can be further studied for classical Chinese literature research. Finally, we build a web server, which provide tools and reports about this project. All the developed tools are free for download and modification.

In this project, we apply the information technology and tools to help the study of terminology research in Buddhist Digital Archive. Via the development of terminology extraction technology, the content creation project can get the convenient reference material and tools. It also supports the Buddhist researcher to do quantitative analysis about the relationships of Chinese or Cross-Lingual Buddhist terms. Therefore, the establishment of Buddhist Knowledge Hierarchy can be accelerated. As the research issues are about classic articles, even the rhyme articles, the results of this research plan might be able to extend to other classic art, history, and philosophy collections. When the Buddhist Archive Research Platform is established, it can increase the visibility of Taiwan in the research area of East thoughts.

英文關鍵詞：Statistical Term Extraction, Cross-language Terminology, Parallel Corpus, PAT-tree, Suffix Array

數位典藏國家型計畫技術研發分項

中文詞彙及跨語詞彙抽取技術在佛典數位典藏上之研發與應用

The Development and Application of the Chinese and Cross-Lingual Term Extraction for Buddhist Digital Archives

壹、 研究計畫之背景及目的

佛教思想及文化，不僅僅是民間的宗教信仰而已，千年來更影響了南亞、東亞及東南亞等國家在政治、經濟、文化、語言以及生活等各方面的發展。因而研究佛教思想，在哲學、民族學、社會學、甚至文學等各面向的社會科學領域來說，都是重要的課題。因此收集並整理佛學思想的各種典籍文獻，自然也是必要的工作。

此間，法鼓山中華佛學研究所在佛學典藏內容上，已持續多年進行內容開發。其發展的計畫包含了國科會「 數位博物館計畫：玄奘西域行 (NSC89-2750-H-119-001, 89/08/01~90/04/31) 」、國科會國家數位典藏「台灣佛教數位博物館：蓬萊淨土遊 (NSC91-2422-H-119-0302, 90/04/01~91/03/31) 」、蔣經國國際學術交流基金會「台灣佛教文獻數位資料庫的建構與研究 (DB001-D-00 89/07/01 ~91/06/30)」與「《法華經》多種語文版本數位資料庫的建構與研究 (DB002-D-02 92/07/01~94/06/30) 」等計畫。多年來並參與製作「漢文電子佛典資料庫」、「台大佛學網路資料庫」,「佛教文獻數位資料庫」及「台灣佛教數位博物館」等涵蓋古今的佛典數位化計畫經驗。今年並依數位典藏國家型計劃之規範，再次提出「佛典數位典藏內容開發之研究與建構－經錄與經文內容的標記作業與知識架構」計畫。以中華佛學研究所的發展方向來看，其終極目標為建立完整的佛學研究知識架構。

如同過去千年來中國佛教典籍歷經各朝各代的收集、整理及闡釋的過程，中華佛學研究所今日所進行的工作，也正是再一次將東亞佛教典籍進行大規模的整理。只是，這一次的整理，所運用的不只是歷代以來皆倚靠的大量時間及人力，更積極運用現代資訊科學技術進行佛教典籍的收集、整理及闡釋。雖然建立完整的佛學研究知識架構，對於佛教典籍的傳承有非常重要的意義，但究竟知識架構的成形必須要靠眾多從事宗教與哲學思想研究的學者經過長時間的研究來達成。因此為了讓宗教與哲學思想研究學者能夠有一個方便建立知識架構的研究環境，我們提出一個佛學典籍研究平台的構想。

所謂佛學典籍研究平台，其長期目標在建立一個能夠支援佛學研究學者進行佛學典籍的研究及教學的主要平台。就細部來說，過去中華佛學研究所已經逐步的建立佛學典籍的資料庫，使用者可以透過所提供的界面，查詢並檢閱佛學典籍的內容。現在則更進一步希望提供統計分析(Statistical Analysis)，資訊檢索及抽取(Information Retrieval and Extraction)，文件自動分類與分群技術(Document Classification and Clustering)，資料探勘(Data Mining)，機器學習(Machine Learning)及知識管理(Knowledge Management)等資訊技術及工具，將其與佛學典籍資料庫結合，提供佛學研究學者更便利的研究環境。

此研究平台表面上只是單純的資料庫加上分析工具的平台，看起來並無特出之處。但是深入探討其重要性，對於佛學研究實在有其新的開創性。其開創性在於提供佛學研究者，進行過去所無法研究的，或非常耗時耗力的議題。一直以來佛學知識的研究方式，皆是以培養研究者扎實的基本功為基礎。亦即研究者必須熟讀佛學的經文及典籍，倚靠記憶及文字整理來進行研究，其研究的範圍皆是以單一經典，或是少數經典為主。如果要研究跨越大量經典的單一議題，則非常的困難。搜尋的功能雖然簡化了跨經典的查詢，

但是使用者進行跨經典的分析及歸納時則仍然費時耗力。因而此平台的建立是極具開創意義的。

為了達成建立佛學典籍研究平台的長期目標，本研究計畫今年提出的研究方向就在建立這個平台中的其中部份技術，亦即本研究計畫的主題－「中文詞彙及跨語詞彙抽取技術在佛學典籍上之研究與應用」。

以佛學典籍為目標，研究「中文詞彙及跨語詞彙抽取技術」的重要性如下：

1. 對於佛學典籍研究平台發展的重要性：
   詞彙抽取技術，是進一步運用其他分析工具－如跨語檢索及文件分類－之基礎。尤其是，當我們要進行語意相似性的分類而非形式相似性分類，特別需要先能夠有建立詞彙的資訊技術。此外，此技術對於進一步發展 Topic Map，和 Semantic Web 等知識架構與管理技術，並進行時空等多維資訊分析及呈現時，是重要的基礎。
2. 對於佛學內容開發計畫的重要性：
   詞彙抽取技術，能夠協助統整詞彙來源參考，並建立一致的詞彙表。由於佛教經典傳至中國時，可能經由不同的路徑，典型如經由西藏或是經由南洋兩條不同的路徑傳入中國。此外同一典籍在不同朝代的釋譯及傳抄，也進一步加深了各版本漢文典籍之間的差異甚至產生謬誤。因而，佛學內容開發計畫的工作人員，在建立經錄和電子全文時，常需要比對持有版本的差異。
3. 對於佛學研究者的重要性：
   統整佛學辭典，補足未收錄的詞彙。如前所述，本計劃所發展之技術可以協助佛學內容開發計畫作為統整詞彙來源的參考。這個統整詞彙的整理結果，將可開放佛學研究者進行深入的討論與研究。可進一步深入的議題包含，漢語詞彙出現的頻率，可用以驗證漢文典籍釋譯的年代，而跨與詞彙的整理，則有助於對照不同傳譯過程的經文優劣與否。

基於上述背景、動機與目標，本計畫結合了台灣大學工程科學所及資訊工程所在資訊技術上的專長，以及中華佛學研究所在佛學典籍上的專業能力，在中華佛學研究所既有的佛學典藏基礎上，為邁向建立佛學典籍研究平台的長遠目標進行一年期的計畫，期以研發佛典詞彙抽取技術及工具，並提供當前及未來佛典內容開發計畫中所需要的詞彙基礎。

貳、 國內外有關本計畫之研究情況、重要參考文獻之評述

本計劃的相關研究及重要參考文獻，將分兩部份討論。第一部份是關於佛典典藏的相關研究。第二部份是關於漢語詞彙及跨語詞彙抽取技術的相關研究。

一、關於佛典典藏的相關研究

近年來佛典典藏的研究，國際間以台灣的發展最為突出。其中，尤其以中華佛學研究所的研究成果最為豐碩。更由於本計劃是以中華佛學研究所的典藏為基礎，我們先介

紹中華佛學研究所過去的研究成果及現有典藏，再檢視國內外其他研究機構所進行的佛典典藏的相關研究。

1. 中華佛學研究所過去的研究成果以及現有典藏

中華佛研所 e-Lib 以豐富並符合國際標準的數位資源為基礎。進一步結合資訊檢索技術及文獻計量分析，配合傳統圖書館概念，以建構佛學知識研究的服務為目標。其主要成果包括：佛學學報全文檢索服務、學報引用資料庫、宗教比較資源整理、各國語言佛學論文蒐集等。重要的典藏條列如下:

a. 漢語文獻內容

有關漢語數位內容的計畫，有「中華電子佛典協會資料庫」、「台灣佛教文獻數位資料庫」、「佛典數位典藏內容開發之研究與建構」等計畫，其中以 CBETA 大正藏的字數約一億字為最多，並持續建構中。

CBETA(Chinese Buddhist Electronic Text Association)電子佛典資料庫是以《大正新脩大藏經》第一卷至第八十五卷、《卍新纂續藏經》第一卷至第九十卷為底本。已發行 CBETA 電子佛典系列光碟片，除包括原《大正藏》1 至 55 冊及 85 冊之外，又新增《卍續藏》禪宗部 9 冊，資料持續建構中，為免費提供電子佛典資料庫以供各界作非營利性使用。

「台灣佛教文獻數位資料庫」整合當代科技與人文科學以建構與研究，期能建置一涵蓋明清、日據時代、戰後臺灣至今的大型台灣佛教資料庫。臺灣佛教雖延綿三百餘年，為臺灣歷史與文化的重要環節之一，此漢文數位資料庫是國際間唯一的台灣佛學資料庫，亦將成為國內外漢學或佛學學者之入口網站。內容包含台灣佛教的相關研究資料，如期刊論文全文、書籍和期刊論文目錄、訪談紀錄、文件、圖片等。

「佛典數位典藏內容開發之研究與建構」計畫係以 CBETA 現有的數位典藏為基礎，及歷代佛典經錄、法寶總目錄、《法寶義林--大正大藏經總索引》等 CBETA 所建置或未列入的作業規劃等文獻資料為基礎，配合當代資訊科技，與 TEI Markup 等標準規範，建構佛典知識管理系統，提綱挈領掌握浩瀚佛典整體內容。內容包括古今中外完整的經錄、提供檢索等資訊功能、連結至各項相關內容、經文知識內容的管理等。

b. 多語文獻內容

在多語數位內容方面，共計有「緬甸聖典寫本簡明目錄」、「佛學數位圖書館暨博物館」、「漢文電子佛典製作與應用之研究——以《瑜伽師地論》為主」、「《法華經》多種語文版本數位資料庫的建構與研究計畫」等專案計畫，語文內容涵蓋中文、英文、巴利文、緬甸文、梵文、藏文等佛教經、律、論文獻。

在「緬甸聖典寫本簡明目錄」專案中，共整理二百四十五函貝葉寫本，包含巴利、緬甸等二種語文的上座部三藏文獻及其疏、鈔以及單行的佛學論著或作品。內容含律藏 127 部，經藏 66 部，論藏 67 部， 以及 49 部語法書和辭典。典藏除佛教律、經、論之外，尚有故事集、詩集、佛學論著、錫蘭佛教史、未見他錄的巴緬逐詞對照譯本以及

多達四十九部的語法書和辭典涵蓋各層面及歷史、語法等重要學科。

「佛學數位圖書館暨博物館」專案計畫，係由中華佛學研究所與台灣大學合作建構與發展。該資料庫自 1994 年起至今已經完成的佛學書目資料約計十萬筆，建置中英文學術論文全文資料庫有 4000 篇左右，以及「佛學網路梵文、巴利文、藏文佛典語言教學」、「佛學網站資源」(Buddhist Internet Resource)、「佛教工具」等資料庫提供學界檢索應用。

「漢文電子佛典製作與應用之研究——以《瑜伽師地論》為主」將《瑜伽師地論》及其綱要書、異譯本（三經二論）、諸注釋書、梵文原典、藏譯本等佛典電子化，並整合參考書目、解題及工具書等，設計使用介面及參照等功能。於工具列中依序提供「解題」、「科判」、「梵漢藏全文檢索」、「辭典」、「參考書目」、「藏經查詢」，以及「Goto」、「引用複製」與「全文檢索」等功能。

「《法華經》多種語文版本數位資料庫的建構與研究計畫」計畫係以漢譯《法華經》為主軸，佐以其他各種版本的《法華經》寫本，進行蒐集、研究，並製作成數位資料。進而，建構梵文寫本與漢譯本及其它譯本的比對，並建立一大型的資料庫，藉由網路資源，廣為流傳於國際間，以作為現代研究教學之參考與應用。內容包括含發現於尼泊爾—西藏(Nepal-Tibet), 中亞(Central Asia), 以及在北巴基斯坦的基爾基特(Gilgit)的梵文法華經寫本或殘卷。並盡力收集分別館藏於尼泊爾寫本檔案館、大谷大學圖書館、大英圖書館斯坦因收藏、劍橋大學圖書館等處的各種寫本，並以科判比對及參考文獻等方式呈現。

c. 多媒體內容

多媒體內容之專案計畫，包括「數位博物館玄奘西域行」、「台灣佛教數位博物館：蓬萊淨土遊」等，內容包含影像圖片、影音動畫、地理資訊及互動學習等。

「數位博物館玄奘西域行」計畫由國立台灣大學哲學研究所主持，共含「文獻、圖像、史地資料之組織與研究」、「互動式資訊視覺化設計與研究」、「數位博物館中知識庫系統之研發」三個子計畫，分別由中華佛學研究所、國立台北藝術大學科技藝術研究中心、國立台灣大學資訊工程研究執行。內容包括大唐西域記、西遊記、絲路之旅、文物藝術、互動式學習區、文獻檢索、相關網站及辭典檢索等。

「台灣佛教蓬萊淨土遊」計畫，建構台灣三百年多來的佛教資料庫，並結合網路、多媒體、視覺設計、腳本設計、內容管理及檢索技術等，建立一國際化，且屬於各年齡層適用的台灣佛教數位博物館。含台灣佛教寺院介紹、文教組織介紹、佛教人物介紹、寺院文化地圖、台灣佛教文物、台灣佛教尋旅互動式學習，以及寺院虛擬巡禮讓使用者身歷其境。

2. 其他國內佛典資料的相關研究及狀況

a. 中華電子佛典協會 － CBETA

如前所述，CBETA 是目前漢文大藏經製作最成功，流通及使用率最高的單位。其主要的經驗與優勢即在於數位全文的標準化。符合 XML 及 TEI 標準的建構主軸，成為目前

數位佛典製作的主要範例。

b. 台大佛學研究中心 – BDLM

台大佛學研究中心的 “佛學數位圖書館暨博物館”，是搜羅最廣泛、資料量最多、使用率最高的重要佛學入口網站。其目標為結合佛學數位資料與現代科技，建立系統化的佛學資料庫，並透過網際網路做最有效率的傳播與分享。

其最重要成果包括：12 萬筆的全文書目檢索系統及 4 千多篇的中、英、日全文檔案，和語言類線上工具的搜集整理。大體上而言，由於台大佛學研究中心，中華電子佛典協會和中華佛學研究所早期是共同合作進行 CBeta 的典籍數位化工作，因此這三個單位的收藏大致相同。

c. 香光尼眾佛學院 – Gaya

香光佛學院圖書館一直有系統的進行佛學圖書資源數位化的工作。並以 “工具書的數位化” 和 “研究適合佛教數位資源的檢索語言” 為主要的發展及研究目標。其重要成果包括：佛學書目、類書、年表、佛教名錄、作者權威資料庫等。

d. 其他重要佛學資源

1. 福嚴佛學院－印順導師全集電子及網路版
2. 佛光山－佛光山電子大藏經

3. 國外在佛典典藏方面的研究

a. Electronic Buddhist Text Intitiative (EBTI)

EBTI 成立於 1993 年，是早期建構中文數位佛典的概念與資源聚集的重要集會。中華電子佛典協會 (Cbeta)也是從當時開始漸漸成型。目前的 EBTI 雖不再活躍，不過早期參予過 EBTI 的個人及單位仍都持續的在佛學(典)數位化領域中努力，並發揮深遠的影響力。

b. 佛教語言相關之研究

佛學的跨語言特性，使得語文類的研究與資料蒐集成為非常重要的課題。在加上數位的技術和網際網路的發展，佛學語文資源的 “質”、 “量”和 “檢索” 等各方面，在近幾年都有長足的進展。國際上主要佛學類的語言資源與研究包括：巴利文 (Pali)、梵文 (Sansk)及藏文 (Tibet)。

現今越來越便利的語言資源仍有共通的幾個問題需要持續努力：

1. 原文校訂、標讀仍不夠完善
2. 許多原典及翻譯仍待收錄
3. 版本資訊不夠明確

4. 部分資源仍無檢索功能
5. 資料的擷取及引用的功能待加強

總的來說，從 EBTI 時期開始，就有大量的資訊人員與數位技術，投入佛典數位化的工作，並持續到今。現今台灣的三個主要佛學資源研究中心即「台大佛學研究中心」、「香光尼眾佛學院」與「中華佛學研究所」。其中台大著重在廣泛的搜集所有的數位資源，以達成佛學數位博物館為目標。香光主力是從圖書館專業上，將佛經書目做大量的數位編目整理，進而做主題分類及檢索資料庫的服務與研究。中華佛研所主要是努力將數位資源標準化，並從底層索引技術到上層知識架構中，結合資訊檢索的研究作使用者介面的開發與服務。

因此，目前絕大部分佛學資源的工作，仍是以累積資料為主。真正有涉及到對所有資料做研究開發或應用的工作，都仍是屬於早期的階段。香光的主題分類研究以傳統圖書館角度出發，以人力為工作進行的主要來源；台大的主題詞資料庫檢索，其範圍與資料量仍需持續增加。以上均尚未真正應用資訊技術作為佛學研究的基本方法。

二、關於中文詞彙及跨語詞彙抽取技術的相關研究

1. 中文詞彙抽取技術的相關研究

談到中文抽辭必定會先考慮到斷詞(分詞)的問題。關於斷詞的技術，中央研究院資訊所陳克健博士所領導的中文詞知識庫小組，已經完成了「具有新詞辨識能力的中文斷詞系統」[1]，並將此技術提供給數位典藏國家型計畫的核心技術－「包含未知詞的斷詞標記系統」[2]。此一系統是結合文法，並運用近十萬辭的辭庫，並加入了同所簡立峰博士所發展的 PAT-tree-based 統計斷詞技術，及同所之馬偉雲先生所發展的未知詞辨識演算法。相關的文獻，可參考「具有新詞辨識能力的中文斷詞系統」網頁中的論文發表結果，就不在此贅述。

在使用幾段佛典的散文，及韻文對「包含未知詞的斷詞標記系統」之展示系統作測試後，驗證中研院資訊所的「中文斷詞系統」確實能夠對佛典的散文，及韻文做相當不錯的斷詞結果，參閱表格一。因此計畫中如果需要斷詞的應用，可以放心的使用此斷詞技術。當然結果不是百分之百正確的，審視斷詞發生錯誤的地方，似乎容易發生在判斷古代譯名為未知詞的情況下。初步推測，應該是起因於佛典的譯名與現代譯名的用字頻率不同，這個差異也可能來自不同地區的譯音造成的問題。如果要進一步改善「中文斷詞系統」所產生的結果，匯入佛學領域的辭彙似乎是最佳的方法，但這似乎成了互為因果的狀況。

表格一：妙法蓮華經散文片段，在中央研究院資訊所的「中文斷詞系統」展示系統上執行的結果。

斷詞前	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，

[1] 「具有新詞辨識能力的中文斷詞系統」網頁, http://godel.iis.sinica.edu.tw/CKIP/wordsegment.htm
[2] 「中文斷詞系統」網頁, http://ckipsvr.iis.sinica.edu.tw/

	逮得己利， 盡諸有結， 心得自在。
斷詞後	如是我聞(VH) ，(COMMACATEGORY) 一時(Nd) 佛住王舍城耆(Na) 闍(FW) 崛(b) 山中(Nb) ，(COMMACATEGORY) 與(P) 大(VH) 比丘(Na) 眾(Na) 萬二千(Neu) 人(Na) 俱(D) ，(COMMACATEGORY) 皆(D) 是(SHI) 阿羅漢(Nb) ，(COMMACATEGORY) 諸(Nes) 漏(VH) 已(D) 盡(D) ，(COMMACATEGORY) 無復(VJ) 煩惱(VH) ，(COMMACATEGORY) 逮得(VC) 己(Nh) 利(VH) ，(COMMACATEGORY) 盡(VJ) 諸(Nes) 有(V_2) 結(VC) ，(COMMACATEGORY) 心得(Na) 自在(VH) 。(PERIODCATEGORY)

中央研究院資訊所的「中文斷詞系統」所產生的結果，雖然相當不錯，但是斷詞的角度仍是從語言研究的目的出發，對於一般學術的研究來說，斷詞並不是最重要目的。然而，中央研究院資訊所的「中文斷詞系統」排除了片語條件，於是古典文獻斷詞後所產生的結果可能過於瑣碎，反而需要再進一步的組合。相反的，如果真要深究古代典籍的字句意含，相對於現代文獻來說單字詞特別的多，從研究者的角度來看，似乎每個單字詞可能都有其意義，那麼做不做斷詞又沒有太大的意義。因而我們仍然需要運用最基本的抽辭技術，畢竟本計劃進行抽辭的目標在提供後續「佛典文獻研究平台」的發展，因此朝向知識架構及思想研究的目標，實高於語言研究的需求。

其他國內外學術單位，亦進行過不少相關的中文詞彙斷詞、抽辭或是詞彙辨識的研究。例如台大資訊工程學研究所陳信希教授所領導的的自然語言處理實驗室，也有關於「中文斷詞及人名、組織名辨識系統」的成果展示。但是經過測試結果，並不適合於佛典文獻的應用。

總的來說，詞彙抽取的方法，可大致分為「文法斷詞」、「辭典抽辭」及「統計抽辭」等三種類型。這三種方法，其實都已經運用於中央研究院資訊所的「中文斷詞系統」。而且中研院資訊所的「中文斷詞系統」已參加第一屆由 ACL SIGHAN 舉辦之中文分詞比賽，並在繁體中文的分組比在中獲得第一名。故不在細數其他的相關研究。本計劃將連絡中研院資訊所，了解是否可利用此「中文斷詞系統」，並加上自行開發的抽辭工具，以達成計畫目標。

2. 跨語詞彙抽取技術的相關研究

數位典藏國家型研究計畫之技術分項的核心技術，提供了網路跨語言資訊檢索系統[3]。這個系統主要是利用 Google 搜尋的結果，加上簡易的詞彙分析，利用單純的統計判斷，即時的從網路上抓取可能的跨語詞彙。不過目前只提供中、日、韓、英等語文之間的詞彙查詢。

不過該技術的方式，我們將運用在漢梵的平行語料庫上。理論上用於平行語料庫上，應該可以獲得比網頁這類非平行語料庫更佳的結果。

[3] 網路跨語言資訊檢索系統。http://livetrans.iis.sinica.edu.tw/lt.html

參、 研究方法及進行步驟

一、本計畫採用之研究方法及進行步驟

由於本計劃包含了中文詞彙抽取與跨語詞彙抽取兩項重要的工作項目，因此所採用的研究方法，也將分兩大部分來討論：

- 第一部分是關於佛典中文詞彙抽取技術之研究。此部分說明本計劃將深入討論的兩種不同作法，包含「辭典抽辭法」，及「統計抽辭法」之研究。
- 第二部份是關於佛典跨語詞彙抽取技術之研究。此部分說明本計劃將深入討論的三項不同的研究議題，包含「利用多個雙語字典的產生新跨語詞典」，以及「利用平行語料庫以統計方式產生跨語詞典」。

雖然本計劃的目的，旨在數位典藏資訊技術的研發，但是其方法的發展並不以純資訊技術為限。在詞彙抽取的技術上，尤其需要佛典專業研究者的參與和協助。且本計劃注重實際的產出結果，因此不論是自動化的技術，或是有人工參與的半自動化技術，只要能夠發揮實質的效益，皆會列入技術研發的範圍之列。因此在後續說明研究方法時，也會考慮加入有人工參與的研究方法。

1. 佛典中文詞彙抽取技術之研究方法

在前述的「中文詞彙抽取技術的相關研究」一節中，我們提到了「文法斷辭」，「辭典抽辭」及「統計抽辭」等三種中文詞彙抽取方式。中研院資訊所的「中文斷詞系統」已經結合了上述三種方法，達到不錯的效果。因此本計畫中斷無再行發展相關斷詞工具的必要性。但是「辭典抽辭」及「統計抽辭」兩種方法，實作並不困難，不需要如中研院系統那樣大量的詞彙庫，也不用人工整理複雜的文法規則，因此仍有應用討論的價值。

a. 辭典抽辭方法

辭典抽辭方式的基礎在於必須先有可用的常用辭典，或是專業辭典。以佛學典藏來說，目前主要的中文佛學辭典，有《丁福保佛學大辭典》和《佛光大辭典》。

《丁福保佛學大辭典》電子檔，由「佛教電腦資訊庫功德會」所提供，以 HTMLHelp 版的形式開發，是由維習安 (Christian Wittern) 依中華電子佛典協會所製的《CBETA 電子佛典》而做的，有些缺字圖檔是取自日本「今昔文字鏡」。內含約 31,266 條詞語。《丁福保佛學大辭典》的範例可參考表格二。

《佛光大辭典》光碟版(v2.0)係由佛光文化事業有限公司所製作發行。由佛光山的大眾，共同研討、修正下，自行開發設計完成。內含約 22,650 條目，7,000,000 萬字的釋文。《佛光大辭典》的範例可參考表格三。

表格二：《丁福保佛學大辭典》的範例辭條

辭條	說明

辭條	說明
法智	（術語）智度論所說十一智之一，觀見欲界苦集滅道四諦法之無漏智也。是初知法，故名法智。又知現在之法，故名現智。大乘義章十五曰：「言法智者，亦名現智。自體名法，初知法故，名為法智。以知現法，故名現智。」[口@又]（人名）四明山延慶寺知禮，宋太宗賜法智大師之號。見佛祖統紀五十。
醫喻經	（經名）佛說醫喻經，宋施護譯，一卷。以醫有四法譬佛法有四諦法。
五鈍使	（術語）謂本惑十使中貪瞋痴慢疑之五使。貪瞋痴慢之四使，都為迷執於世間事物而起之惑，其性分鈍者，故謂為鈍使。疑使為就四諦真理而起之惑，以猶豫不決為自性，其性分亦非銳者，故類從貪等四使而為鈍使。
二種之因果	（名數）分四諦為二種之因果：一、世間因果，苦諦為果，集諦為因。二、出世間因果。滅諦為果，道諦為因。
七使	（名數）一、欲愛，欲界之貪欲也。二、恚，瞋恚也。三、有愛，色界無色界之貪欲也。四、慢，慢煩惱也。五、無明，痴惑也。六、見，五邪見也。七、疑，疑四諦之理也。見輔行六。
世尊	（術語）梵語曰路迦那他 Lokana^tha，譯為世尊，或婆迦婆 Bhagavat 譯為世尊。佛之尊號。以佛具萬德世所尊重故也。又，於世獨尊也。阿含經及成實論以之為佛號中之第十，以具上之九號，故曰世尊。涅槃經及智度論置之於十號之外。智度論二曰：「路迦那他，秦言世尊。」淨影大經疏曰：「佛具眾德為世欽仰，故號世尊，若論胡音樓伽陀伽此云世尊也。」探玄記九曰：「以佛具三德六義，於世獨尊，故名世尊，即梵名婆伽婆。」佛說十號經曰：「天人凡聖世出世間咸皆尊重，故曰世尊。」成實論一曰：「如是九種功德具足，於三世十方世界中尊，故名世尊。」

表格三：《佛光大辭典》的範例辭條

辭條	說明
增上心學	梵語 adhicitta-ciksa，巴利語 adhicitta-sikkha。即「三學」中具有增上勢力之定學，為能增進「心」之學，故稱增上心學。修定能收攝散亂，令心專注於一對象；進而遠離欲望與邪惡，趨向見性悟道。俱舍論卷二十八（大二九・一四五中）：「有餘師說，即心一境相續轉時，名三摩地；契經說此為增上心學故，心清淨最勝，即四靜慮故。」〔集異門足論卷五〕_p5964
三火	<一>指貪、瞋、癡。又作三毒、三垢。即：(一)貪（梵 raga）火，渴取一切順境。(二)瞋（梵 dvesa）火，對所遇逆境引起忿怒。(三)癡（梵 moha）火，心智懵懂，不明事理，顛倒妄取，起諸邪行。大寶積經卷九十六（大一一・五四二下）有「我見諸眾生，三火所熱惱」之句。凡人之煩惱，以貪欲、瞋恚、愚癡荼毒人最劇，故又稱三毒。其微細難察，殘害身心，常使人沈淪於生死輪迴。此三者能害眾生，為惡之根源，故又稱三不善根。〔北本大般涅槃經卷二十九、有部毘奈耶卷三十四、大智度論卷二十一〕 <二>印度婆羅門教及印度教大型祭祀中必置之供奉物。即(一)家主火（梵 Garhapatyagni），為調理供養諸神及婆羅門供物所設之火，火爐為圓形。(二)供養火（梵 Ahavaniyagni），位於家主火之西方，燒供物奉獻諸神，火爐為方形。(三)祖先祭火（梵 Anvaharyapacana），供物奉獻祖先，火爐為半月形；位於家主火之南，故又稱南火（梵 Daksinagni）。_p536
無作	梵語 akarmaka 或 akrtrima。指無因緣之造作。又指心無造作物之念，如無作三昧；或謂不假身口意之動作而自然相續之法，如無表色、無作戒等。維摩詰所說經卷下（大一四・五五四下）：「修學無相無作，不以無相無作為證。」〔維摩詰所說經卷上、無量壽經卷下〕（參閱「無作三昧」5087、「無作戒」5087、「無表色」5097）_p5086

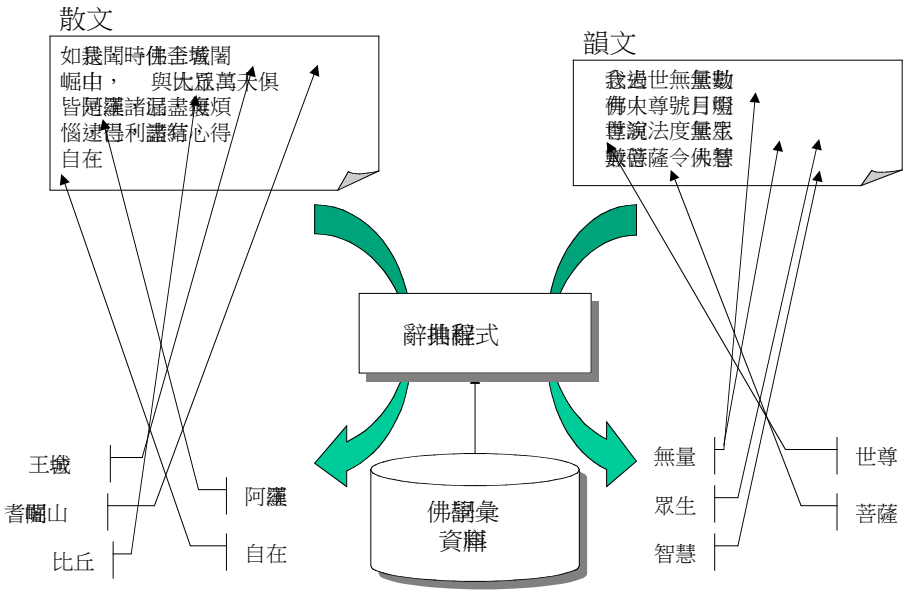
其他常用的中文佛學辭典還有《中華佛教百科全書》、《中國大百科全書(佛教篇)》、《三藏法數》、《英漢對照佛學詞彙》、陳義孝《佛教常見詞彙》等[4]。

辭典抽辭法主要並不需要詞彙的相關內容，只是要利用既有的辭彙，尋找文件中確

[4] 關於佛學辭典的相關連結可參考「香光尼眾佛學院圖書館」的佛教字辭典網頁
http://www.gaya.org.tw/library/b-ip/dictionary.htm

有此詞彙出現之處。利用詞彙斷詞的概念可參考圖表一。如附圖所示，處理規則如下：

1. 佛典散文和韻文經過抽辭工具，可從文章中識別出已知的佛學詞彙。
2. 詞彙抽取過程中如果發現有多個詞彙可以對照，則以長辭為主。如 “羅漢” 與 “阿羅漢” 都可抽取時，以長辭 “阿羅漢” 為主。
3. 如果抽取選擇的辭彙重疊，例如“今天天氣很好”，可能抽出“今天”、“天天”、“天氣”三個詞，但若選擇“天天”，則損失兩個辭彙，依照抽取最多詞彙的原則，放棄抽取“天天”這個辭彙。



圖表一：辭典抽辭法之示意圖。佛典散文和韻文經過抽辭工具，可從文章中識別出已知的佛學詞彙。

上述抽辭法則，基本上就是 Shallow Parsing。亦即在抽詞的過程中，並沒有考慮文法結構的問題，而只是以抽出長辭及抽出最多辭為優先。

抽辭完畢之後，剩餘的文字片段可進行排序整理，如果發現重複達一次數時，可另外篩選出來，提供給佛典專業研究人員做進一步判斷是否要收錄的考慮。

此方法最大的好處在於方式實作簡單，只要收集相當數量的詞庫後，可以立即開始。在現代文獻抽詞的研究上，早有成功的例子。但是在佛典古典文獻上，成功率有多少尚不可知。對於韻文的特殊情況，辭典抽辭方式是否有效，則需要進行大規模的實驗才能確定。此外，是否除了佛學辭典也需要納入其他一般性辭典？又有那些一般性辭典應該納入？也是需要研究的課題。

b. 統計抽辭方法

中文統計抽辭相對於辭典抽辭方式的最大好處，就在於不需要有辭典的存在。統計抽辭的基本條件是需要有足夠的語料庫，則抽辭的過程在以統計原則了解前後單字之前的相依性。中文統計抽辭最主要的研究，是中研院簡立峰博士 1999 在 IP&M 期刊上所發

表的以 PAT-tree 為基礎的統計抽辭方法。

這個抽辭方法先判斷單字是否與前字(或後字)應該相連。如果該單字 c 出現的頻率為 f_c，c 之前(後)可能出現字的種類數為 L ，而其中單字 p 出現最多次，且次數為 f_p。若 $L < t_1$ 或是 $f_p/f_c > t_2$，則 c 字應該與 p 字連在一起。

此外，如果發現兩個長字串 a，b 相連，則需判斷是否有可能為複合詞彙。若共同出現的機率為 P_u 各別出現的機率是 P_a，P_b。若 $P_u/(P_a+P_b-P_u)$的比值須大於 t_3。

簡立峰博士原始論文的 t_1，t_2 值及 t_3 是針對短篇新聞抽辭所作之研究。馬偉雲先生則運用此方法在未知詞的抽取上。至於這些參數值，在佛典文獻上應該定為多少，是需要進一步研究的問題。此外，我們也希望能夠改善此一模型，加入 Machine Learning 的方式，來決定參數的設定。

2. 佛典跨語詞彙抽取技術之研究方法

在這個計畫中，我們不僅希望發展佛典跨語詞彙的抽取技術，更希望藉由這個機會建立佛典的平行語料庫。光是建立這個平行語料庫，就已經是很有價值的工作。因為，這將是非常稀有的古典典籍平行語料庫。這個平行語料庫，不只有散文的對照，更有韻文的對照。整理後的資料，不僅可用於佛教典籍的研究，更可用於當時的地方語言學的研究。

由於佛教經典是透過藏傳及南傳兩條不同的路徑，透過不同的傳抄、釋譯，因而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。如何提供工具協助研究者進行跨語研究，本計劃將先著重於跨語詞彙抽取的問題上。我們的研究方法描述如下：

a. 利用多個雙語字典的產生新跨語詞典

目前可以看到的多語佛教辭典為《中日韓佛教名詞辭典》。本辭典又稱為 CJK(Chinese, Japanese and Korean)佛教名詞辭典，係由日本 Toyo Gakuen 大學的 Charles Muller 等學者於 1986 年開始編輯，約蒐集 30 萬個佛教名詞，內容亦包含英文說明的資料。《中日韓佛教名詞辭典》的詞條範例如下。

表格四：《中日韓佛教名詞辭典》的詞條範例

辭條	說明
菩提	[py] pútí [wg] p'u-t'i [ko] 보제 pori [ja] ボダイ bodai Ⅲ A transliteration of the Sanskrit/Pali term *bodhi*, meaning wisdom, enlightenment or awakening. (1) The wisdom of the true awakening of the Buddha. Enlightenment. The function of correct wisdom. The situation of the disappearance of ignorance due to the functioning of awakened wisdom. (2) The wisdom of perceiving the reality-nature. (3) Sublime enlightenment. The expression of enlightened wisdom. (4) An abbreviation of 菩提道場, (*bodhi-manda*). The place where the Buddha attained his enlightenment. **[Dictionary References]** Naka1221d Iwa735 **[Credit]** cmuller(entry) cwittern(py)
四攝法	[py] sì shèfǎ [wg] ssu-she-fa [ko] 사섭법 sisŏppŏp [ja] シショウホウ shishōhō Ⅲ (*catuh-saṃgraha-vastu*); The 'four methods of winning (people) over.' Also written 四攝

辭條	說明
	事. The four methods that bodhisattvas employ to approach and save people. They are: (1) *bushi* 布施, giving the gift of Dharma or something that people like; (2) *aiyu* 愛語 using kind words; (3) *lixing* 利行 acting for the purpose of benefit to them; (4) *tongshi* 同事 physically working together with them. **[Dictionary References]** Naka524c ZGD438c JE288b/319 Yo666 ZGo9-P206 ZGo17-P61 FKS1853 DFB **[Credit]** cmuller(entry) cwittern(py)
天龍八部	[py] tiān lóng bā bù [wg] t'ien-lung pa-pu [ko] 천룡발부 ch'ŏnyongp'albu [ja] テンリュウハチブ tenryūhachibu Ⅲ The eight groups of transmundane beings that are usually present at Mahāyāna sutra convocations: *deva* 天, *nāga* 龍, *yakṣa* 夜叉, *gandharva* 乾闥婆, *asura* 阿修羅, *garuda* 迦樓羅, *kimnara* 緊那羅 and *mahoraga* 摩睺羅伽. All of these are considered to be protectors of the buddhadharma. **[Dictionary References]** Naka985c **[Credit]** cmuller(entry) cwittern(py)

由於佛教經典是透過藏傳及南傳兩條不同的路徑到中國，因此在進行佛典典藏內容開發時，常常需要比對梵文或巴利文的原點。但是梵文或巴利文目前都沒有與漢文對照的辭典，僅有與英文對照的辭典。因此我們希望藉由梵英及漢英辭典兩個雙語詞典，彙整出梵漢/漢梵的詞彙對照表。

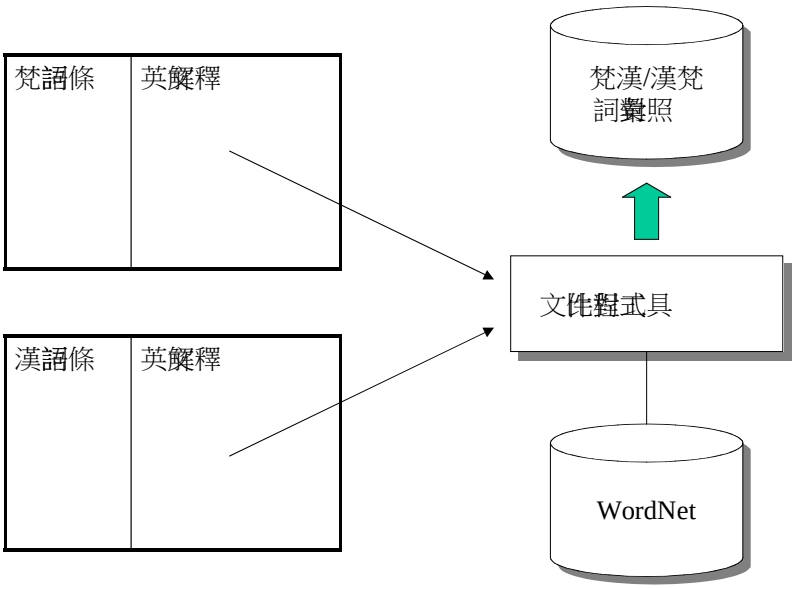
目前佛光大辭典中有少部份漢梵的詞彙對照。而佛學研究者閱讀梵文時，主要是使用梵英辭典－《MW 梵英字典》。本辭典係由 MONIER WILLIAMS 所編輯，牛律大學出版社發行，主要依以英文來解釋梵文的字義，按照梵文字母順序排列。《MW 梵英字典》的詞條範例如下。

表格五：《MW 梵英字典》的詞條範例

辭條	說明
a-jita	mfn. not conquered , unsubdued , unsurpassed , invincible , irresistible ; m. a particular antidote ; a kind of venomous rat ; N. of Vishn2u ; S3iva ; one of the Saptarshis of the fourteenth Manvantara ; Maitreya or a future Buddha ; the second of the Arhats or saints of the present (Jaina) Avasarpin2i , a descendant of Ikshva1ku ; the attendant of Suvidhi (who is the ninth of those Arhats) ; (%{As}) m. pl. a class of deified beings in the first Manvantara.
tulana	n. lifting Mr2icch. ix , 20 ; weighing , rating , iii , 20 ; N. of a high number Buddh. L. ; (%{A}) f. rating ib. ; equalness with (instr. or in comp.) Prasannar. ii , 16.
kuze-zaya	mfn. lying in Kus3a grass MBh. xiii , 1698 ; m. a kind of tree (Pterospermum Acerifolium) L. ; the Indian crane L. ; N. of a mountain in Kus3a-dvi1pa VP. ; (%{am}) n. `" lying in water "' , a water-lily MBh. R. &c. ; [once (%{A}) f. Hariv. 8428] ; %{-kara} m. `" having rays like waterlilies "' , the sun W. ; %{-bhU} m. N. of Brahma1 Ba1lar. ; %{-maya} mf(%{I})n. consisting of water-lilies R. vii , 36 , 10 ; %{-locanA} f. a lotus-eyed woman Bha1m. ; %{kuzezayA7kSa} mfn. lotus-eyed Ragh. xviii , 3 Ra1jat.
sa-sattva	mf(%{A})n. possessing energy or vigour MBh. ; containing living creatures or animals Mn. Ragh. ; (%{A}) f. containing an embryo "' , a pregnant woman Ragh. iii , 9.
hara	mf(%{A} , rarely %{I})n. (only ifc. ; fr. 1. %{hR}) bearing , wearing , taking , conveying , bringing (see %{kavaca-} , %{vArttA-h-}) , taking away , carrying off , removing , destroying (see %{bala-} , %{zakti-h-}) ; receiving , obtaining (see %{aMza-h-}) ; ravishing , captivating (see %{mano-h-}) ; m. `" Seizer "' , `" Destroyer "'N. of S3iva A1s3vGr2. Mn. MBh. &c. ; of a Da1nava MBh. Hariv. ; of a monkey R. ; of various authors &c. Cat. ; (in arithm.) a divisor Col. ; the denominator of a fraction , division ib. ; a stallion (?) L. ; an ass L.

我們會嘗試以英文解釋，當作梵語詞條及漢語詞條的特徵向量，以文件比對的方式找出可以對應的梵漢詞彙。其示意圖如圖表二。

由於不同跨語辭典中的英文解釋可能用詞差異極大，目前並沒有把握能夠有效的利用此方式，解決跨語辭典的問題。但梵漢/漢梵的詞彙對照，對於中華佛研所的內容開發，會產生實質的幫助，因此仍然值得研究此方法的可行性。



圖表二：以文件比對的方式建立梵漢詞彙對照之示意圖

b. 利用平行語料庫以統計方式產生跨語詞典

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，並將散文和韻文的形式個別呈現。

表格六：《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 bhagavān rājagṛhe viharati sma gṛdhakūṭe parvate/ mahatā bhikṣusaṃghena sārdhaṃ dvādaśabhir bhikṣuśataiḥ/ sarvair arhadbhiḥ kṣīṇāsravair niḥkleśair vaśobhūtaiḥ suvimuktacittaiḥ suvimuktaprajñair rājāneyair mahānāgaiḥ kṛtakṛtyaiḥ kṛtakaraṇīyair apahṛtabhārair anuprāpta-svakārthaiḥ parikṣīṇabhavasaṃyojanaiḥ samyagājñāsuvimuktacittaḥ sarvacetovaśitāparamapāramitāprāptair abhijñānābhijñātair mahāśrāvakaiḥ / 《英譯》 Thus have I heard. Once upon a time the Lord was staying at Râgagriha, on the Gridhrakûta (1)mountain, with a numerous assemblage of monks, twelve hundred monks, all of them Arhats, stainless, free from depravity, self-controlled(2), thoroughly emancipated in thought and knowledge, of noble breed, (like unto) 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(3); eminent disciples,

表格七：《妙法蓮華經》韻文部份梵、漢、英對照範例

《漢文》
我念過去世　　無量無數劫
有佛人中尊　　號日月燈明
世尊演說法　　度無量眾生
無數億菩薩　　令入佛智慧
《梵文》
atha khalu mañjuśrīḥ kumārabhūta etamevārthaṃ bhūyāsyā mātrayā pradarśayamāṇas tasyāṃ vel-yām imā gāthā abhāṣata //
atīnamadhvānamanusmarāmi acittiye aparimitasmi kalpe /
yadā jino āsi prajāna uttamaścandrakasya sūryasya pradīpa nāma //57//
saddharmadeśeti prajāna nāyako vineti sattvāna anantakoṭyaḥ /
samādapeti bahubodhisattvān acintiyānuttami buddhajñāne //58//
《英譯》
And on that occasion, in order to treat the subjectmore copiously, Mañgusrî, the prince royal, utteredthe following stanzas:
57. I remember a past period, inconceivable, illimited kalpas ago, when the highest of beings,
the Gina of the name of Kandrasûryapradîpa, wasin existence.
58. He preached the true law, he, the leader of creatures; he educated an infinite number of kotis of beings, and roused inconceivably many Bodhi-sattvas to acquiring supreme Buddha-knowledge.

利用平行語料庫以統計方式產生跨語詞典的研究方法如下：

1. 先利用程式建立梵、漢、英的平行語料庫。目前的初步觀察。其平行對照的情況，應該能以 “句” 為對應的單位。
2. 利用漢語詞彙 C 找出所有的包含 C 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 Bi-partite Graph 中找尋 Max Matching Pair 的方式來尋找其他可對應的辭彙。

3. 研究進行步驟

針對上述說明的各種研究方法，我們將進行的步驟分成數個階段。

第一階段是整理資料。在整理資料的階段中，主要是人工的參與資料整理，以及協助整理程式的開發。其工作項目如下：

1. 整理平行語料庫。佛研所雖然有多國語文的經典，但是並未建立為平行語料庫。在這項工作中，我們將開發工具程式，建立平行語料庫。並以佛典專業的人力，對無法建立平行語料庫的部份，以人工方式檢查。整理好的平行語料庫，將利用目前團隊內已開發的 Suffix Array 及 Signature File 的索引工具建立索引檔。
2. 整理中文佛學辭典，及跨語辭典。開發工具程式，匯整各辭典之詞彙，並將收集之詞彙鍵入關聯式資料庫。同時，也建立 Suffix Array 索引檔。

第二階段是進行中文詞彙的抽辭研究。將以中華佛學研究所既有的中文電子佛典為語料庫，針對「佛典中文詞彙抽取技術之研究方法」一節中，所描述的兩種方法進行深入探討。

1. 利用第一階段所整理中文佛學詞典，進行「辭典抽辭法」的研究。將中華佛學研究所的佛學文獻，區分近代文獻，與古代典籍兩大類，分別進行詞彙抽辭方法的研究。針對現代文獻可利用一般中文辭典協助進行抽辭。但由於近代文獻對於佛學研究的議題來說，仍以佛學專有詞彙，或是宗教哲學研究專有詞彙為主。預期一般詞彙的主要功能是用於協助決定適當的抽辭邊界。至於古代典籍，利用詞彙抽辭方法所能得到的正確性，將是重要的研究議題。
2. 研究「統計抽辭法」對於近代文獻與古代典籍的作用差異，以及運用統計抽辭法在佛典文獻時，應該設定的參數條件。
3. 可以進一步引入機器學習的演算法，如 Neural Network 或是 SVM 方式，讓系統自動學習統計法的最佳參數為何。
4. 整理相關的程式，完成中文佛典的抽辭工具。

第三階段是進行跨語詞彙的抽辭研究。則是針對中華佛學研究所，現有的跨語資料，針對「佛典跨語詞彙抽取技術之研究方法」一節中所描述的兩項研究議題，深入探討。

1. 研究如何利用多個雙語詞典，建立統整的跨語辭典。例如，將梵英的佛學辭典，和漢英的佛學辭典匯整建立梵漢英的跨語辭典，使得梵漢語詞彙有直接對照的基礎。此項工作並非完全倚靠人力完成，而是開發相關的軟體工具，以程式進行初步自動比對，或至少以程式工具協助佛典典籍研究人員快速建立跨語辭典。
2. 研究如何利用平行語料庫及統計法，自動建立跨語辭典。在第一階段的平行語料庫的整理結果，以及第二階段中文詞彙抽取的研究成果，都將進一步應用於跨語辭典的自動建立機制中。此工作項目中，需要研究統計法分析運用於跨語詞彙關聯性抽取的可行性，以及抽取時需要的參數值為合。
3. 整理相關的程式，完成多語佛典的跨語抽辭工具。

第四階段是網路服務的規劃與設計。此階段的任務在將前三階段研究的成果，結合到目前中華佛學研究所的線上佛學典藏資料庫中，以提供相關的服務。工作項目包含：

1. 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。
2. 線上佛學詞彙關聯性分析服務。本項工作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。如何將分析工具適當的結合，以及將詞彙關聯性的定性定量結果做有效呈現，是本項工作的重點。
3. 線上佛學多語專有詞彙之出處比對服務。使用者透過此服務，可以在線上針對所收集的佛典，查詢不同語文佛學詞彙的出處，並可透過介面進行比對。此服務可幫助，有意研究譯本之間用詞差異之學者。

上述的四階段工作，雖有先後之順序，但並非逐段執行。因此計劃執行過程中，會

依照執行的情況進入下一階段，因而可能會有前後階段工作的重疊。實際執行進程的規劃，請參考下一節的計畫甘梯圖。

4. 計畫甘梯圖(Gantt Chart)

表格八：本計劃執行甘梯圖

月次 工作項目	第1月	第2月	第3月	第4月	第5月	第6月	第7月	第8月	第9月	第10月	第11月	第12月	備註
第一階段「整理資料」													
整理中文佛學辭典，及跨語辭典													
整理平行語料庫													
第二階段「中文詞彙抽辭研究」													
辭典抽辭方法研究													
統計抽辭方法研究													
對統計抽辭法做參數最佳化													
整理中文佛學辭彙抽取工具程式													
第三階段「跨語詞彙抽辭研究」													
跨語辭典半自動統整之研究													
平行語料庫及統計法之研究													
整理跨語佛學辭彙抽取工具程式													
第四階段「網路服務規劃」													
中文專有詞彙之檢索及抽取服務													
詞彙關聯性分析服務													
跨語專有詞彙之出處比對服務													
計畫進度報告													
期中進度報告													
期末結案書面報告													
結案程式及網站彙整													

肆、 結果與討論

一、統計法文獻抽詞程式的開發

本計劃是研究詞彙抽取技術在佛典上的應用，在文獻討論中，我們提到中央研究院資訊所的「中文斷詞系統」所產生的結果，雖然相當不錯，但是斷詞的角度仍是從語言研究的目的出發，對於一般學術的研究來說，中央研究院資訊所的「中文斷詞系統」排除了片語條件，於是古典文獻斷詞後所產生的結果可能過於瑣碎，反而需要再進一步的組合。因此本計劃的重點之一就是研究統計抽辭在佛典文獻上的應用。

我們以簡立峰博士的所發展的 PAT-tree-based 統計斷詞技術為基礎開發文獻抽辭的程式。簡博士的 PAT-tree 抽辭技術在發展上已經有一段時間，分析所得到的抽辭程式，我們發現有下列幾項問題：

1. Overhead 高：由於整個技術底層的資料結構建構在 PAT-tree 的基礎上，因此針對 Nbytes 的原始資料，其建立 PAT-tree 所需要的記憶空間非常的大。如果為了節省計算 pattern 次數的時間，使用擴展資料結構，並且使用正向和反向兩個 PAT-tree，則整體所需要的空間大約是 10~20 倍（$10N \sim 20N$ bytes）左右的額外空間。從這點來看，PAT-tree-based 統計斷詞技術雖然在抽詞功能上已經達到，但是非常不合適於大量文獻的抽辭工作。

2. 抽辭程序慢：此外由於 PAT-tree-based 抽辭程式在建立好 PAT-tree 索引後，抽辭方式是逐步檢查雙字詞、三字詞、…，一直到 k 字詞（$k \cong 10$），因此抽辭程序非常緩慢。計算的複雜度為 $O(Nk)$。若資料大量增加時，抽取大量詞彙所需要的時間會太長。

3. PAT-tree 索引檔只能增加語料不能刪除：PAT-tree Index 在建立的時候最大的問題是當使用部分語料庫逕行抽辭的時候，必須要針對部分的語料庫，重新建立 PAT-tree Index。

4. 多個 PAT-tree 索引檔無法合併：如果針對不重複部分的語料庫，已經各自建立好 PAT-tree index。但卻無法很容易的將這多個 PAT-tree index 合併。換言之，必須要重新建立整個語料庫的 PAT-tree Index。

5. 在記憶體外作業困難：目前的 PAT-tree 統計斷詞技術必須將整個 PAT-tree index 建於記憶體中。如果記憶體不夠，需要在記憶體外使用 PAT-tree 索引，也就是在硬碟上進行運作的話，需要將程式大幅度的重寫。

由於 PAT-tree 所需要的空間很大，因此在本計劃中的黃乾綱博士開發了直接以 Suffix Array，進行抽辭工作的演算法。使用 Suffix Array 做為統計抽辭技術的底層資料結構，其好處有下列幾點：

1. Overhead 低：若原始資料為 N bytes，則建立 Suffix Array Index 所需要的記憶空間，即使包含正向和反向兩個 Suffix Array，整體所需要的空間也只要 2 倍（$2N$ bytes）左右的額外空間。由於 Overhead 低，Suffix Array based index 合適於大量文獻的抽辭工作。

2. 抽辭程序快：此外由於 Suffix Array 抽辭程式在建立好索引後，抽辭方式是以 s 掃過（scan）整個 Suffix Array 的方式來做，因此 k 字詞（$2 \le k \le 100$）的抽取不需要反覆掃描，因此抽辭程序的計算複雜度為 $O(N)$。若資料大量增加時，抽取大量詞彙所需要的時間仍然只需要掃描資料一次。

3. Suffix Array 索引檔可以快速增刪除不需要的語料：Suffix Array 刪除不需的語料部分，不用重新建立 PAT-tree Index，只要從 Suffix Array 中過濾掉資訊即可。

4. 多個 Suffix Array 索引檔可以快速合併：如果不重複部分的語料庫，已經各自建立好 Suffix Array index。如果要合併做抽辭，只需要利用 Merge Sort 做排序，其實間複雜度為 $O(n)$，不需要重新建立 Suffix Array。

5. 方便於記憶體外作業：Suffix Array 的運作方式簡單，所以不一定將整個 Index 建於記憶體中。如果記憶體不夠，可以直接在硬碟上進行運作，也不需要將程式大幅度的重寫。

6. 加速容易：Suffix Array 的概念簡單，使得程式可以很容易的使用其他程式技巧，如 Caching、Tree Index 或是 Hashing function 來加速 Pattern 搜尋的速度，如此可以更快的處理詞邊界判斷的工作。

整個 Suffix Array based 的統計抽辭技巧，最主要的部份就在掃描 Suffix Array 進行抽辭的程序。相關資料及程式碼可以從 http://dev.ddbc.edu.tw/BuddhistTermExtract/ 網站上取得。

二、佛學辭彙庫自動抽辭研究

1. 古典文獻抽辭結果

本計劃中針對古典文獻抽辭所使用的語料庫資源，是電子佛典協會 CBETA 典藏電子佛典，相關的統計資料可以參考表格九。

表格九：古典文獻抽辭語料庫的統計資訊

抽辭資源	CBETA
檔案大小(bytes)	1.2 GB (utf8 files)
UTF-16 中文字碼(bytes)	324,754,728 (utf16 file)
Suffix Array Index (bytes)	567,406,444 (4 bytes for each character)
總中文字數	141,851,611 (Index/4)
總標點符號字數	20,525,753 (no index)

基礎候選詞條數	15,694,556

抽辭的方式為利用簡立峰博士所提出的計算詞邊界判斷的基本原則，第一批篩選出的候選詞條，其條件為該詞條的左右兩邊均出現過的字超過兩種以上。例：

…一一佛世尊…
…第一佛字維衛佛…

上面例子中，「一佛」這個詞條左右均出現兩種以上的字，這樣的辭條便會被篩選出來。將所有 128 個字以內的連續字串的前後可能性都整理過後，總共有 15,694,556 個詞條被篩選出來，這就是本計劃抽辭部分最基礎的候選詞條數。

這樣的基礎詞條數在本計畫取用的文本中仍會有非常少許的詞條被遺漏，例如：

… 陳主譯。見一乘寺藏。眾經目 …
… 陳主譯。見一乘寺藏眾經目錄 …

上述例子中，一乘寺為一寺廟名，但在文本出現狀況太少，其前後出現字的狀況又完全一致，因此不會出現在我們第一步的基礎候選詞條中，這是本計劃使用的大量統計方式的演算法中，暫時無法處理的狀況。因此以下所有的分析比較，一直到最後的詞彙抽取結果，都不會出現「一乘寺」這類比較依據太少的詞彙。

在一千五百多萬個基本候選詞條中，本計劃使用「與佛學辭典詞彙比對」的方式，來分析兩者交、聯集的變化，以便進一步找出基礎候選詞條中，真正有意義的辭彙有哪些，並且決定使用的標準值是多少。

(1) 與 Muller’ s 佛學辭典詞條比對：

【第一項數據分析】

本計劃設定參數 lRl (在詞前為lLl) 為任意長度詞條左右出現的字的數量，因此基礎候選詞條的 lRl 值為左右的 lRl >= 2，計算基本資料與方式如下：

表格十：統計抽詞與 Muller 佛學字辭典的相關統計數據

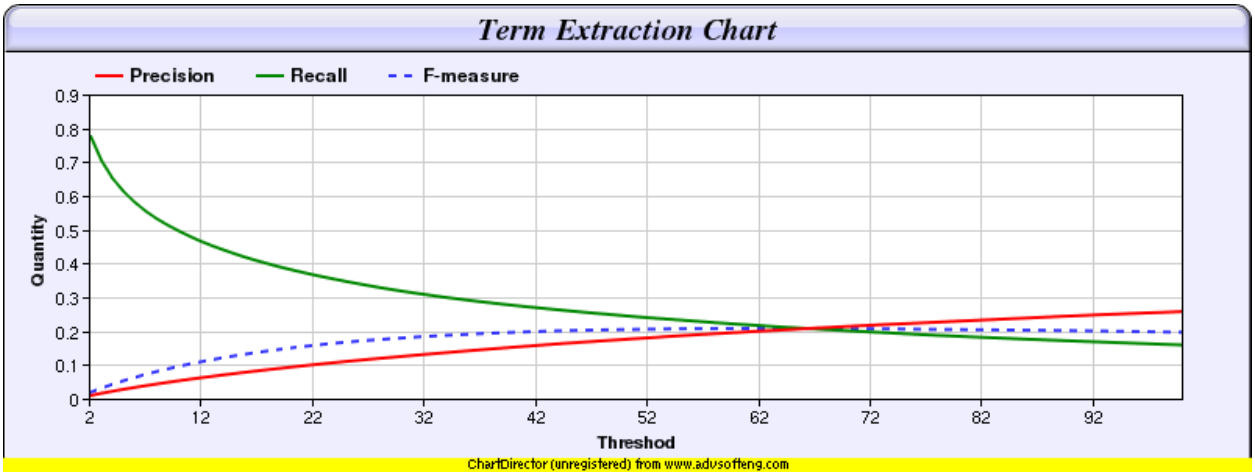
Muller’ s 字典詞條總數	289,838
Muller’ s 在 Cbeta 中出現的詞條數	178,652
基礎候選詞條數	15,694,556
基礎候選詞條與 Muller’ s 字典的交集詞條數	139,486

Precision = 139,486 / 15,694,556 (基礎詞條與 Muller’ s 交集數 / 基礎詞條數)
Recall = 139,486 / 178,652 (基礎詞條與 Muller’ s 交集數 / Muller’ s 詞條數)
F-measure = 2 / (1 / Precision) + (1 / Recall)

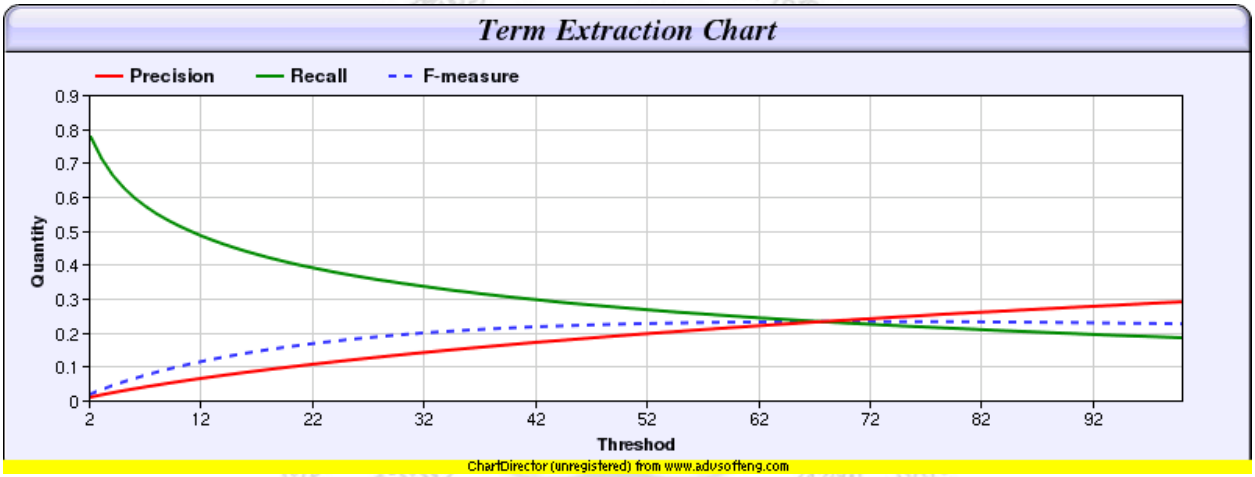
以下為每次 lRl 值改變後，Precision, Recall, F-measure 三個數值的變化圖

Y : Precision, Recall, F-measure

X : |R|數量的變化，僅取 R >= 2 到 R <= 100



圖表三：詞條前 |L| 值對應 Muller 的變化



圖表四：詞條後 |R| 值對應 Muller 的變化

以下為上述左右的 Precision, Recall, F-measure 交會處的參數資料，較多的數據請在本計劃網站中取得: http://dev.ddbc.edu.tw/BuddhistTermExtract。

表格十一：(左側)變化值

X(|L|)	Y(Precision)	Y (Recall)	Y (F-measure)
66	0.206775402199	0.208203658509	0.2074870725
67	0.208646691	0.206384479323	0.207509419839
68	0.210300256211	0.204453350648	0.207335590642

表格十二：(右側)變化值

X(|R|)	Y(Precision)	Y (Recall)	Y (F-measure)
70	0.236575540238	0.227470165461	0.231933521294

71	0.238729441543	0.225953249894	0.232165708584
72	0.240633711429	0.224195642926	0.232124022023

以上述 F-measure 最高點為決定 |R| （和|L|） 值的條件，則從基本候選詞條中，篩出 109,681 個詞條（左側 |L| >= 67 及 右側 |R| >= 71）。此條件下的詞條與 Muller’ s 詞條的關係分析如下：

表格十三：統計抽詞設定|L|>=67 && |R|>=71 與 Muller 佛學字辭典的相關統計數據

基礎候選詞條數	15,694,556
Muller’ s 字典詞條數	178,652
基礎候選詞條中 |L|>=67 && |R|>=71 的詞條數	109,681

表格十四：統計抽詞參數不同的數據

	與 Muller 交集數	未出現於 Muller
|L|>=2 && |R|>=2	139,486 (78%)	39,166
|L|>=67 && |R|>=71	33,166 (18.5%)	145,486

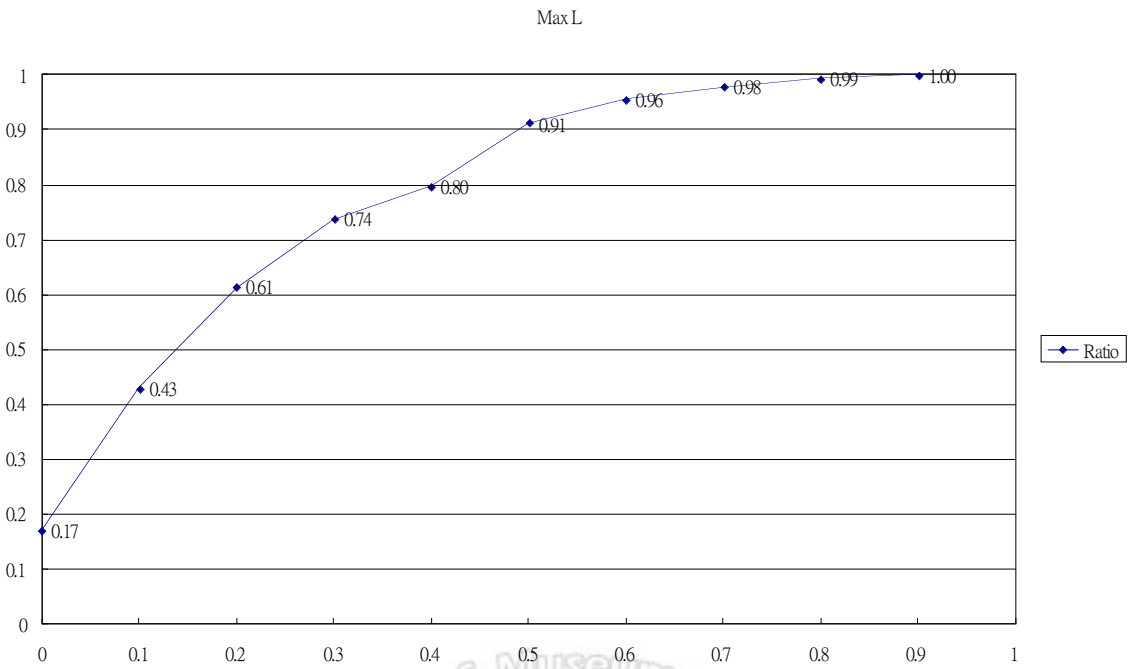
與 Muller 詞條的交集變化在提高 |R| 值後，下滑的幅度符合圖表三、圖表四中的 Recall 的曲線。這表示左右字種類較少的詞條，在數量上的變化幅度較大，且這部份（圖 14,15 曲線交會處左側）的詞彙在現有辭典中的比例相當高（見上表，|R|>=2 與 |R|>=67 兩個結果比較，在 Muller 中出現的量大幅下滑）。但是曲線右側的辭彙仍然很多（十萬以上），且準確性（人工大致辨識後，是有意義的詞彙）也相當高。因此可以確定用此方式篩出的詞條，確實將傳統字辭典不曾出現，但在文本的紀錄與敘述中出現的辭彙大量的篩選出來。

以上述的結果可延伸出兩個主題：

1. 調整演算法，將字辭典以有的辭彙，準確的篩選出來。
2. 分析傳統辭典沒有出現的詞的種類，並做進一步字辭典編纂的補助工具與資料。

【第二項數據分析】

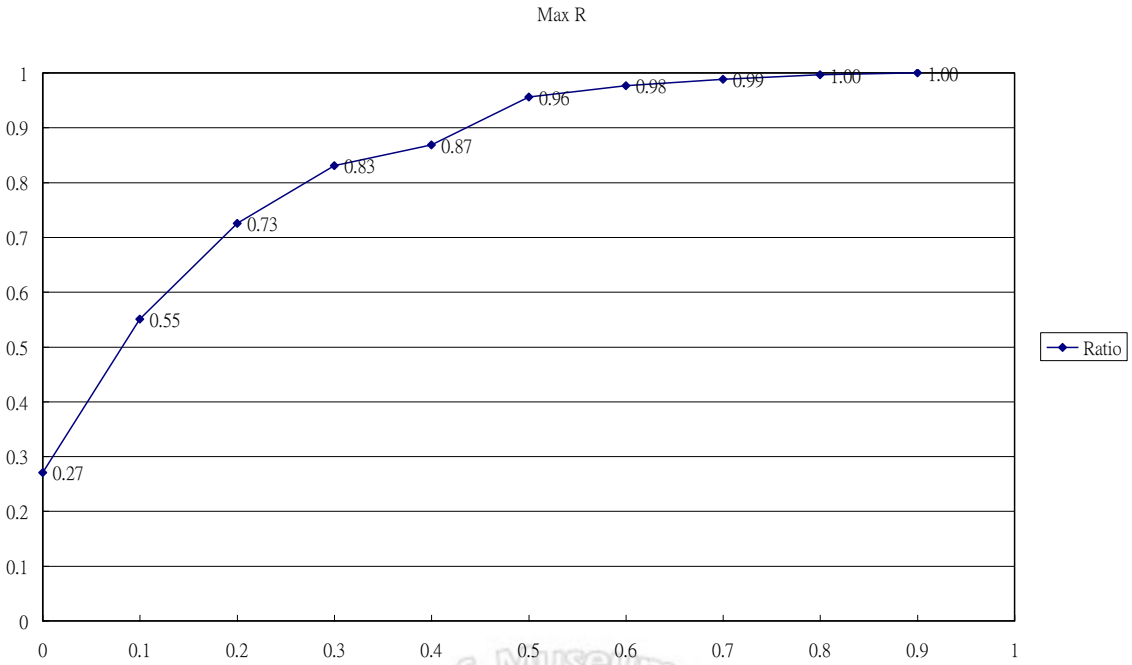
找出候選詞條左右出現字的種類中，數量最多的字，以該數量除以詞條出現的次數（fxb/fx）。fxb/fx 值越高時，表示與後方詞共通出現的強度較高，則此處可能不是辭的邊界。



圖表五：詞條前 fax/fx 值對應 Muller 的變化

表格十五：詞條左側 fax/fx 值對應 Muller 的變化

Max(fax)/fx	Accumulated Count	Ratio
0.0	23778	0.170469
0.1	59928	0.429635
0.2	85649	0.614033
0.3	102937	0.737974
0.4	111307	0.79798
0.5	127174	0.911733
0.6	133371	0.95616
0.7	136371	0.977668
0.8	138562	0.993376
0.9	139486	1



圖表六：詞條後 fxb/fx 值對應 Muller 的變化

表格十六：詞條左側 fax/fx 值對應 Muller 的變化

Max(fxb)/fx	Accumulate Count	Ratio
0	37698	0.270264
0.1	76774	0.550406
0.2	101226	0.725707
0.3	115914	0.831008
0.4	121160	0.868618
0.5	133354	0.956039
0.6	136236	0.9767
0.7	137816	0.988027
0.8	138999	0.996509
0.9	139486	1

2. 當代文獻抽辭結果

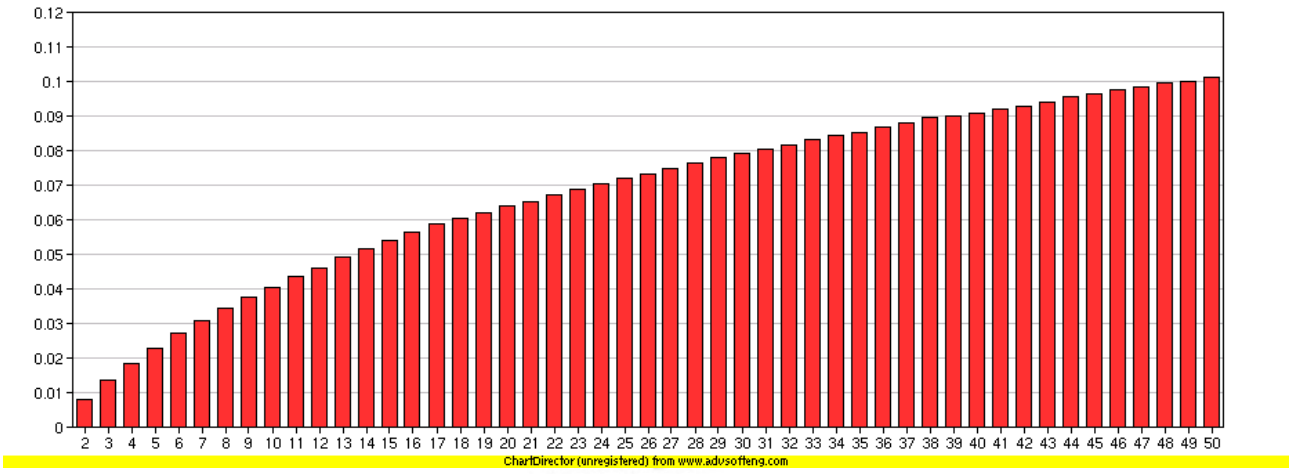
表格十七：當代文獻抽辭語料庫相關統計數據

抽辭資源	781 單篇文獻 （中華佛學學報、華岡學學報、中華佛學研究、台大佛學學報、法鼓全集等）
檔案總 bytes	78 MB (utf8 files)
所有中文字所佔 bytes	19,328,504(utf16 file)
Suffix Array Index bytes	33,851,932(4 bytes for each charactor)
總中文字數	8,462,983
總標點符號字數	1,201,269

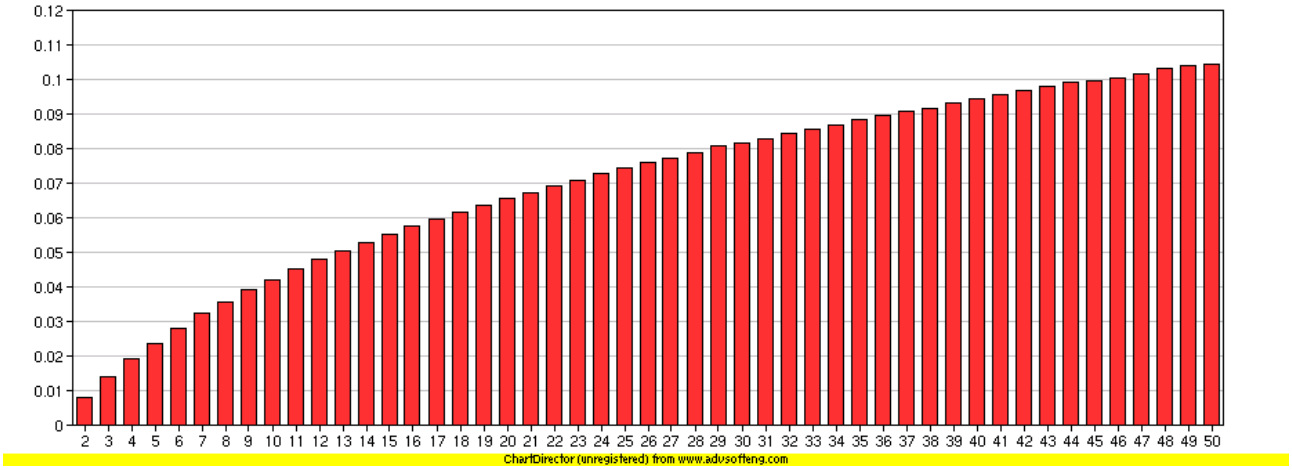
當代文獻理論上應與非佛學辭典做比較，本計畫拿詞條數量比 Muller's 少的佛光大字典做比較，數據僅供參考，在此列出，但並未做進一步的分析與結果描述。

以下為每次 |R| 值改變後，Precision 個數值的變化圖

Y : Precision
X : |R|數量的變化, 僅取 R >= 2 到 R <= 100



圖表七：詞條前 |L| 值對應 Muller 的變化

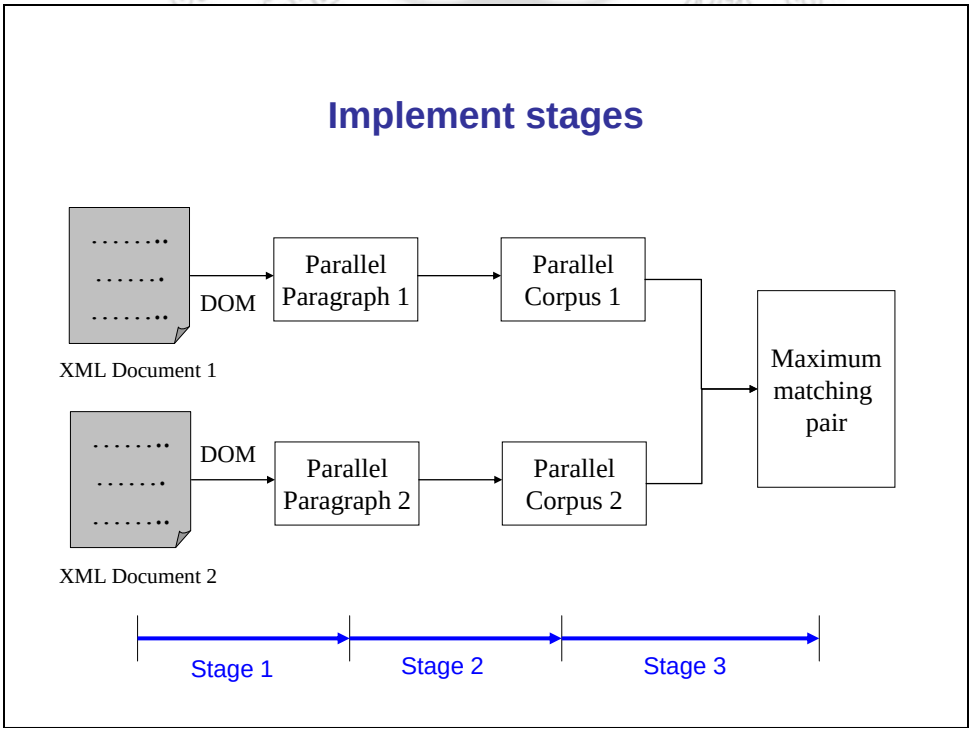


圖表八：詞條後 |RI 值對應 Muller 的變化

更多的數據請在本計劃網站中取得: http://dev.ddbc.edu.tw/BuddhistTermExtract

三、平行語料庫

有關佛教名詞的平行語料庫的詞彙抽取，其執行步驟約略可分為三個階段：第一階段是先將《妙法蓮華經》的梵文、漢文、英文、藏文等具有 TEI（Text Encoding Initiative）標記的 xml（Extensible Markup Language）數位檔案，依其科判的標記（markup tag）作為對應，以 DOM（Document Object Model）程式取出對照的平行段落（parallel paragraph）。第二階段則執行計算每個詞彙的字頻，與可能對應的機率值，來決定平行語料庫（parallel corpus）。第三個階段再利用 Bi-partite graph 中尋找 Max Matching pair 的方式，找出其它可能對應的詞彙，並進一步與跨語辭典做比對。詳如圖表九。

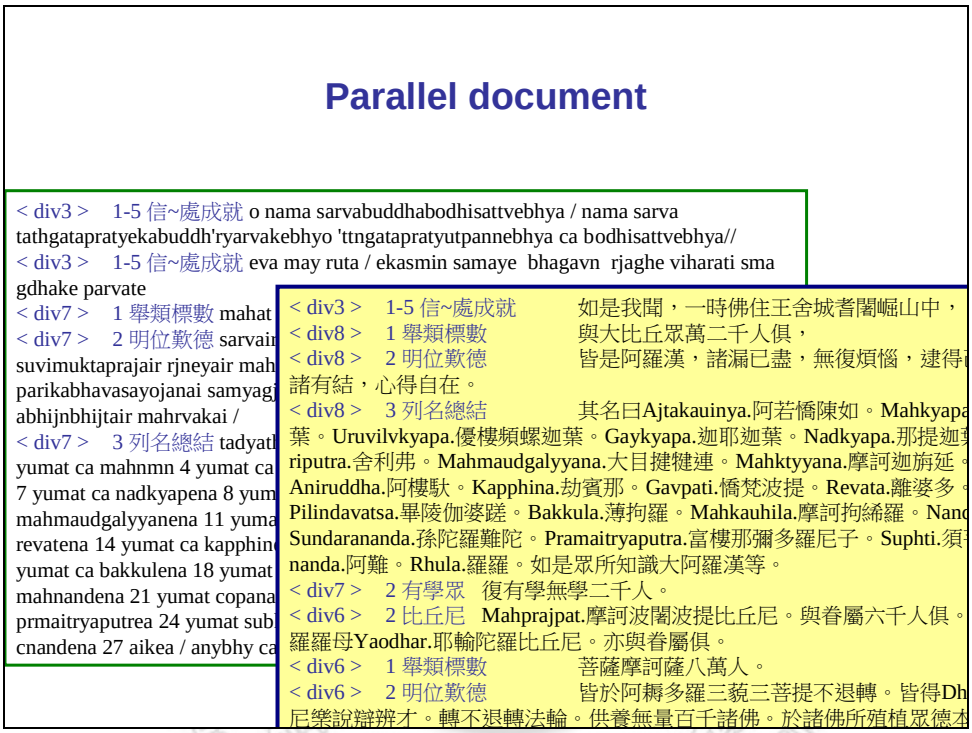


圖表九：執行階段及步驟

有關跨語平行語料庫的研究成果，主要有「平行詞彙的對照」及「詞彙的機率統計」等項，分別說明如下：

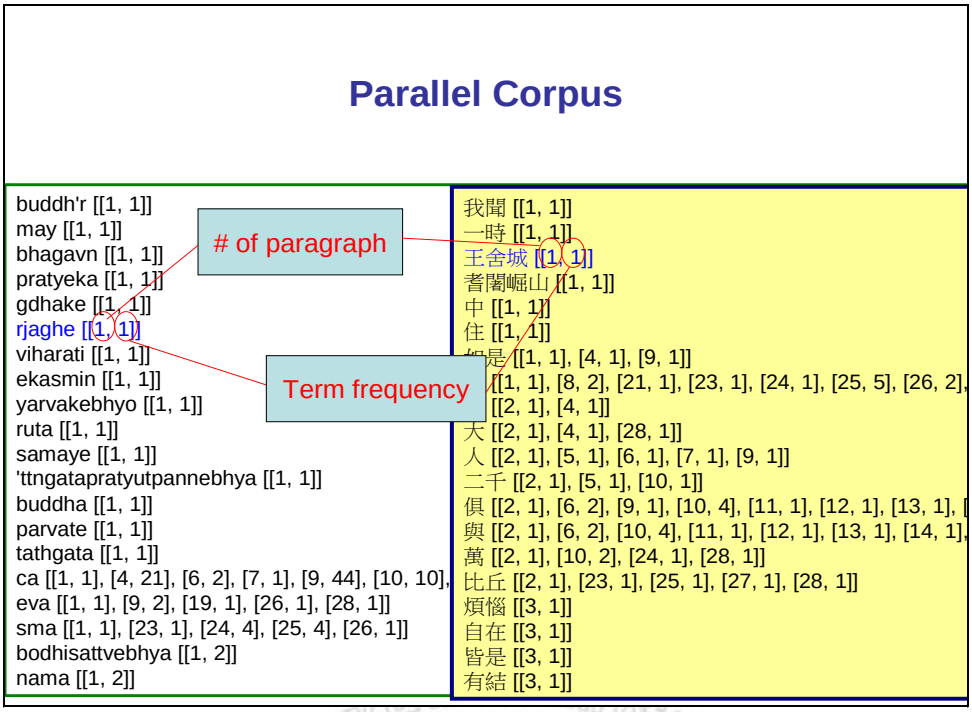
a. 平行詞彙的對照

現階段是以《妙法蓮華經》的梵文及漢文的數位文獻檔案做測試，將二個文獻依其科判的標記（markup tag）作為段落，並以 DOM（Document Object Model）程式取出對照的平行段落（parallel paragraph）。詳如圖表十。



圖表十：佛典文獻平行段落的對照

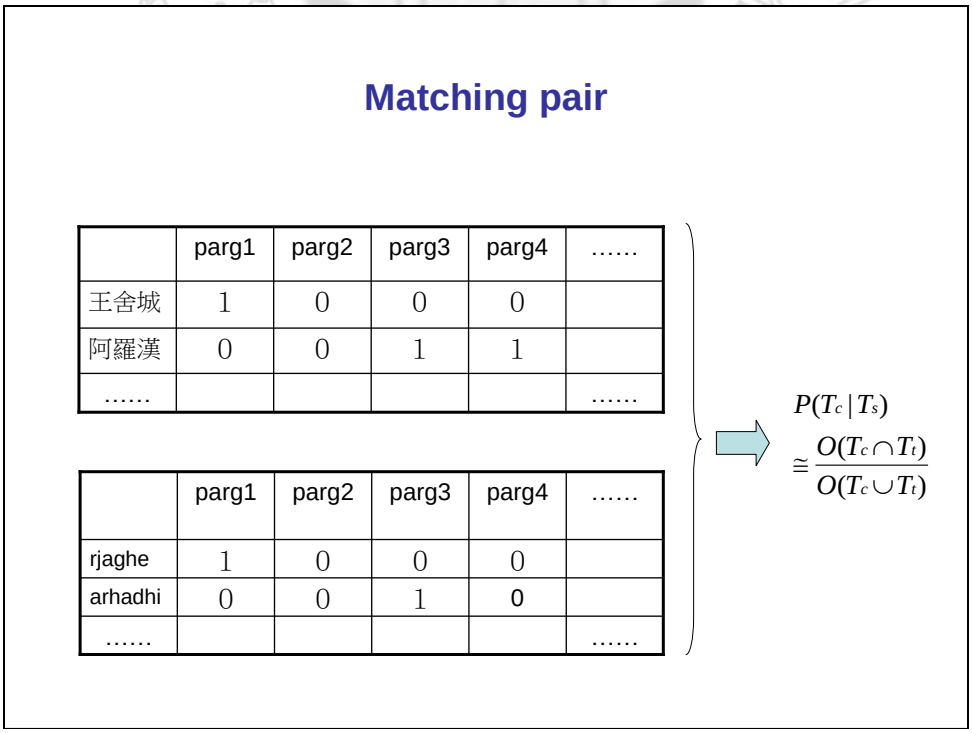
取出對照的平行段落（parallel paragraph）後，再以程式計算每個詞彙的字頻、出現的段落，以及可能對應的機率值，用以決定平行語料庫（parallel corpus）。詳如圖表十一。



圖表十一：平行詞彙的計算

b. 詞彙的機率統計

取出的平行詞彙，再以程式記錄詞彙出現的段落，詞彙若出現在該段落則記錄值為 1，若無出現記錄值為 0，而形成 signature file，詳如圖表十二。再將此 0 與 1 組成的 signature file 數列，套入機率公式計算，即可進一步決定跨語的詞彙對照關係。



圖表十二：平行詞彙的機率統計

四、網路服務

本計劃為了能夠將網頁快速提供給其他研究團隊，從 2006/11/22 起架設了計劃的網站，網址為 http://dev.ddbc.edu.tw/BuddhistTermExtract/

網站上提供下列相關資訊：

1. 古典文獻（Cbeta）抽辭結果
2. 當代文獻（佛學學報）抽辭結果
3. 語用索引及時空地理檢索系統
4. CBETA 語用索引 線上服務 (XML-RPC service)
5. 書面資料及程式

所有程式及資料目前都是開放給需要的人直接下載。

伍、 成果自評

本計劃執行時間為 2006/3/1 至 2007/2/28，執行情況與原計劃的程度大致相符，且 90% 的工作都有達到預期的目標，以下就一年內完成的工作項目及成果作說明。

一、完成之工作項目及計劃直接成果

完成的工作項目，如前面「研究進行步驟」與「計畫甘梯圖(Gantt Chart)」兩節所述。項目包含了「整理資料」、「中文詞彙的抽辭研究」、「跨語詞彙的抽辭研究」和「規劃並提供網路服務」等四階段共 10 項主要工作。

計畫產出的直接成果，敘述如下：

1. **佛學古典文獻平行語料庫**：僅依目前的規劃，我們將針對法華經整理中文及梵文平行與料庫。甚至在整理資料過程中，可能會產生漢、梵、藏、泰的平行語料庫。除了其他宗教的經典外，此平行語料庫的特性在於這是極少有的東亞及南亞古典文獻平行語料庫。而且更重要的是，這個平行語料庫中同時包含了各國**韻文**及**非韻文**的平行語料庫。光就此語料庫而言就已經是重大的產出。這對於語言學、文學、史學、哲學等各方面在時間軸及空間軸的研究上，都是重要的資料來源。而本計劃所研究的成果能夠先行協助學者在佛學古典文獻的詞彙研究議題上，對此平行語料庫做充分利用。
2. **佛學詞彙庫**：本計畫也會產出多語及跨語的佛學詞彙庫。佛典內容開發者，以及佛典研究者，都可以查詢此詞彙資料庫。這個詞彙資料庫，將能夠使佛典思想及語意的研究更趨完整及一致性。
3. **工具程式**：本計畫的重要目的，即是經由計畫的執行產出能夠協助資訊分析及抽取的工具程式。這些工具程式，不論自動或半自動，目的在協助數位內容的開發，即典藏的利用。這些工具程式的研究成果，將近一步納入長期規劃的「佛學典籍研究平台」之中，成為平台系統的一部分。

4. **網路服務**：我們在進行步驟的第四階段進行三項網路服務的規劃。則是將本計劃的研究成果，進一步透過 WWW 的方式，直接提供研究人員使用。透過這個網路服務所得到的反應及回響，也有助於我們具體了解未來整體「佛學典籍研究平台」的發展方向。

二、論文發表

本計劃執行過程已發表過下列相關的會議論文

1. 「中文詞彙與跨語詞彙抽取技術在數位佛典上的研發與應用－階段成果研討報告－」第五屆數位典藏研討會，台北，2006

2. 「Initial Results of Chinese and Cross-lingual Term Extraction for Buddhist Digital Archives」(Poster) Pacific Neighborhood Consortium 2006 Conference, Seoul, Korea, 8/15~8/18, 2006

3. 釋法源,李家名,黃乾綱（2007）。佛典跨語文獻的詞彙庫及索引建立之輔助方法研究 。東京學藝大學，第三屆文學與資訊科技國際研討會，東京。

三、對於學術研究、國家發展及其他應用方面預期之貢獻

本計劃以資訊技術及工具，便利大規模佛學典藏中專有辭彙之研究。由於詞彙的抽取，是建立知識架構的基礎，因此此研究計畫，能夠加速對於佛學研究知識架構的建立。

由於本計劃的研究對象，是以佛典的內容為主，而且其中又包含了韻文及非韻文。其研究成果，有極大的可能性可以延伸至一般性文史哲資料集，例如，中國史料，古典散文，以及古典韻文－詩詞曲等的。此外，並可催生結合不同資料集，進行文、史、哲的跨領域社會科學研究，例如從時間與空間兩軸上研究中國佛學及中國文學之間的相互影響。

如果本計劃的長期目標－建立「佛學典籍研究平台」，能夠完成，則可奠定我國在東亞哲學思想研究的地位，提升我國文哲研究的國際能見度。而且完善的線上服務，能夠吸引國際學者在我們建立的平台上，進行相關的研究。藉由觀察其他人所進行的研究，可以更進一步快速提升我國在東亞哲學思想－特別是佛學思想研究的廣度與深度。

四、對於參與之工作人員，預期可獲之訓練

本計劃的參與人員除了包含三位計畫主持人之外，主要的研究人力來自中華佛學研究所的佛學資訊組研究生，以及中華佛學研究所圖書館的工作人員。並輔以台大工科所，及台大資訊所的碩士班學生，提供資訊檢索、資料探勘與機器學習等技術的的協助。由於過去台大資訊所與中華佛學研究所的進行計畫合作時，台大資訊所開發完成的技術與系統，往往因為台大研究生的畢業而無法繼續相關的研究。因此，在本計畫的進行過程中，中華佛學研究所的人員扮演更積極重要的角色。目前，計畫主持人黃乾綱博士，與共同主持人歐陽彥正博士，都在中華佛學研究所開設知識管理相關之技術課程。所以除

了計劃內的工作實務經驗，相關的技術知識也同時利用教學的方式散播。

換言之，中華佛學研究所的資訊處理能力從過去單純以資訊技術做儲存、檢索及呈現等主要應用，進一步發展成利用資訊技術進行文獻研究與即時分析的能力。因為本計劃，會大幅將各種資訊技術，如資訊檢索、資料探勘與機器學習等相關演算法及統計分析技巧引入，因而透過參與本計劃，將培養出能夠以計量分析方法 (Quantitative Analysis) 進行文史哲研究的人力，並開創文史哲計量研究的新時代。

台大工科所及台大資訊所的研究人力，目前以資料探勘及機器學習演算法在生物資訊上的運用作為主要的的研究方向。學生參與本計畫，也可以藉由研發的過程了解這些演算法在文件處理上的運用，特別是處理古代典籍時是否有不同的考量，藉此可更熟悉演算法的掌控。

陸、 參考文獻

1. Aming Tu, 2003.11.07, “Taiwan Buddhist Digital Museum and XML Markup, Metadata”, Pacific Neighborhood Consortium Annual Conference and Joint Meetings 2003 (PNC/ECAI, 2003.11.07~09), Bangkok: The Maha Chakri Sirindhorn Anthropology Center, Thailand.
2. Aming Tu, 2003.12.05, 「佛學研究與資訊素養 Buddhist Studies and Information Literacy」, 佛教與當代社會學術研討會 (2003.12.05~07)，香港：香港大學佛學中心。
3. Aming Tu, 2004.05.08,” A New Environment for the Buddhist Digital Text”, Congress of Cultural Atlases: The Human Record, May 7-10, 2004, University of California, Berkeley, USA.
4. Aming Tu, 2004.05.09,” CBETA as a Tripitaka Traslation Tool” at Institute for World Religions, Berkeley , USA.
5. Aming Tu, 2004.05.12, ” Buddhist Digital Text and CHIBS Digital Projects ” , University of the West, LA, USA.
6. Chen Keh-Jiann, Ming-Hong Bai, "Unknown Word Detection for Chinese by a Corpus-based Learning Method", International Journal of Computational linguistics and Chinese Language Processing, 1998, Vol.3, #1, pages 27-44.
7. Chen Keh-Jiann, Wei-Yun Ma, 2002, "Unknown Word Extraction for Chinese Documents", Proceedings of Coling 2002, pp.169-175.
8. Wei-Yun Ma and Keh-Jiann Chen, 2003, "A Bottom-up Merging Algorithm for Chinese Unknown Word Extraction", Proceedings of ACL, Second SIGHAN Workshop on Chinese Language Processing, pp31-38.
9. Wei-Yun Ma and Keh-Jiann Chen, 2003, "Introduction to CKIP Chinese Word Segmentation System for the First International Chinese Word Segmentation Bakeoff", Proceedings of ACL, Second SIGHAN Workshop on Chinese Language Processing, pp168-171.
10. Yu-Fang Tsai and Keh-Jiann Chen, 2003, "Reliable and Cost-Effective Pos-Tagging",

Proceedings of ROCLING XV, pp161-174.
11. Yu-Fang Tsai and Keh-Jiann Chen, 2003, "Context-rule Model for POS Tagging", Proceedings of PACLIC 17, pp146-151.
12. Yu-Fang Tsai and Keh-Jiann Chen, 2004, "Reliable and Cost-Effective Pos-Tagging", International Journal of Computational Linguistics & Chinese Language Processing, Vol. 9 #1, pp83-96.
13. Lee-Feng Chien, "PAT-Tree-based adaptive keyphrase extraction for intelligent Chinese information retrieval", Information Processing and Management, 35(1000), pp501-521
14. Christian Wittern，Editing XML，佛教圖書館館訊，第 24 期，54-59，2000
15. Daphne Koller, Mehran Sahami "Hierarchically classifying documents using very few words." Proceedings of ICML-97, 14th International Conference on Machine Learning, 1997.
16. F. Li and Y. Yang. "A loss function analysis for classification methods in text categorization" The Twentith International Conference on Machine Learning (ICML'03), pp472-479, 2003.
17. J. Zhang and Y. Yang. "Robustness of regularized linear classification methods in text categorization" ACM SIGIR'03, pp 190-197, 2003.
18. R. Ghani, S. Slattery and Yiming Yang. "Hypertext categorization using hyperlink patterns and meta data" The Eighteenth International Conference on Machine Learning (ICML'01), pp 178-185, 2001.
19. R.E. Valdes-Perez and etc, "Demonstration of Hierarchical Document Clustering of Digital Library Retrieval Results". Joint Conference on Digital Libraries (JDCL '01), Roanoke, VA (presented as a demonstration), June 24-28, 2001
20. S. Chakrabarti, B. Dom and P. Indyk "Enhanced hypertext categorization using hyperlinks." Proceedings ACM SIGMOD International Conference on Management of Data (pp.307-318), Seattle, Washington: ACM Press, 1998.
21. S.-L. Chuang and L.-F. Chien, "Automatic query taxonomy generation for information retrieval applications" Online Information Review (OIR), 27(4):243-255, 2003
22. S.-L. Chuang and L.-F. Chien, "Enriching Web taxonomies through subject categorization of query terms from search engine logs" Decision Support System, Special Issue on Web Retrieval and Mining, 30(1):113-127, April 2003.
23. S.-L. Chuang and L.-F. Chien, "Towards automatic generation of query taxonomy: A hierarchical query clustering approach" Proc. the 2002 IEEE International Conference on Data Mining (ICDM), pages 75-82, Maebashi City, Japan, Dec. 9-12, 2002.
24. Steve Pepper and Chief Strategy Officer, "The TAO of Topic Maps, Finding the Way in the Age of Infoglut," http://www.ontopia.net/topicmaps/materials/tao.html
25. V. Kashyap, C. Ramakrishnan and T. C. Rindflesch, "Towards (Semi-)automatic Generation of Bio-medical Ontologies" Poster Proceedings of the AMIA 2003 Annual Symposium, November, 2003, Washington, DC.
26. Y. Yang "A study on thresholding strategies for text categorization" Proceedings of ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'01), pp 137-145, 2001.

27. Y. Yang "An evaluation of statistical approaches to text categorization." Journal of Information Retrieval, Vol 1, No. 1/2, pp 67--88, 1999.
28. Y. Yang and J.P. Pedersen "A Comparative Study on Feature Selection in Text Categorization Proceedings" of the Fourteenth International Conference on Machine Learning (ICML'97), 1997, pp412-420.
29. Y. Yang and Xin Liu "A re-examination of text categorization methods." Proceedings of ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'99, pp 42--49), 1999.
30. Y. Yang, J. Zhang and B. Kisiel. "A scalability analysis of classifiers in text categorization" ACM SIGIR'03, pp 96-103, 2003.
31. Y. Yang, S. Slattery and R. Ghani. "A study of approaches to hypertext categorization" Journal of Intelligent Information Systems, Volume 18, Number 2, March 2002.
32. 王威澤，原生型 XML 資料庫關聯規則之探勘與利用關聯規則探勘壓縮資料庫，朝陽科技大學，資訊管理系，碩士論文，2002
33. 光音天，漢文電子佛經檢索軟體的實作，佛教圖書館館訊，第 14 期，51-54，1998
34. 吳寶原，從實務經驗談電子佛典初步工程之演進，佛教圖書館館訊，第 14 期，25-31，1998
35. 李志強，淺談電子佛典與知識管理，佛教圖書館館訊，第 30 期，12-20，2002
36. 杜正民，2003.03.18，「電子佛典缺字於經文、注疏及校勘的處理」，數位典藏國家型科技計畫－技術研發分項計畫「漢字智慧編碼與應用研討會」(2003.03.17~19)，台北：中央研究院歷史語言研究所。
37. 杜正民，2003.10.18，「佛學與資訊的整合」，兩岸佛教學術研究現況與教育發展研討會(2003.10.18~19)，台北：中華佛學研究所。
38. 杜正民，2003.12.26，「從佛典的傳譯談佛教數位博物館的建構 From the Circulating and Translating of the Budhist Stutras to the Constrution of the Buddhist Digital Museum」，數位典藏國家型科技計畫「從紙本碑拓到數位化博物館」座談會，台北：中央研究院歷史語言研究所。
39. 杜正民，2004.05.01，“ The Digital Silkroad Around the world” ，第五屆印順長老與人間佛教海峽兩岸學術研討會 － 國際絲路座談會，主辦單位：佛教弘誓學院、佛光山文教基金會、法鼓山中華佛學研究所、慈濟大學宗教與文化研究所，花蓮: 慈濟講堂。
40. 杜正民，2004.05.02，「閱讀數位原典——大藏經電子版與讀經新環境」，印順導師百歲嵩壽弘法會，主辦單位：福嚴佛學院等單位，新竹: 玄奘大學。
41. 杜正民，當代國際佛典電子化現況，佛教圖書館館訊，第 15 期，28-39，1998
42. 周伯戡，從傳統佛典到電子佛典，佛教圖書館館訊，第 14 期，14-24，1998
43. 周邦信，標記語言的應用，佛教圖書館館訊，第 24 期，41-53，2000
44. 林光龍；葉建華；歐陽彥正，知識庫在數位化圖書館中的應用，佛教圖書館館訊，第 30 期，21-32，2002
45. 林光龍；歐陽彥正，佛教知識庫的建立：以 Topic Map 建置玄奘西域行為例，佛教圖書館館訊，第 31 期，41-54，2002
46. 林信成，基於 XML 之新一代 Web 技術及其在電子出版之應用，佛教圖書館館訊，第 23 期，18-40，2000

47. 施郁芬，佛教資料數位化的重要性及發展重點，佛教圖書館館訊，第 15 期，40-43，1998
48. 洪一禎，以知識庫為基礎的模糊資訊擷取新方法，國立交通大學，資訊科學系，碩士論文，2002
49. 黃士銘，建置一網際網路資料倉儲系統，中華民國資訊管理學會會報，第 9 期，135-152，2002
50. 黃博聖，XML 文件存取控制系統之建置，華梵大學，資訊管理學系，碩士論文，2002
51. 黃惠株，知識管理前傳：環境建立實例探討，佛教圖書館館訊，第 30 期，33-41，2002
52. 維習安，數位化中文佛教大藏經，佛教圖書館館訊，第 15 期，24-27，1998
53. 鄭恩娥，XML 知識庫之整合，國立臺灣大學，資訊工程學研究所，博士論文，2001
54. 鄭振煌，從佛教傳承談佛教知識體系，佛教圖書館館訊，第 31 期，72-81，2002
55. 謝大寧，佛教與佛教知識，佛教圖書館館訊，第 31 期，82-90，2002
56. 謝水鳳，以調整 FP 樹狀結構為基礎之關聯規則漸進式探勘方法，國立臺灣師範大學，資訊教育研究所，碩士論文，2002
57. 謝清俊，佛教資料電子化的意義，佛教圖書館館訊，第 18 期，11-17，1999
58. 謝清俊，資訊科技及人文社會的影響，1996
59. 謝清俊、陳昭珍、莊德明、周亞民〈電子佛典中處理版本的方法〉，中央研究院資訊所，1994 年 4 月
60. 釋惠敏，大藏經電子化的實作，佛教圖書館館訊，第 18 期，18-26，1999
61. 釋惠敏、杜正民、周邦信，2003.09.01「數位化古籍校勘版本處理技術——以 CBETA 大正藏電子佛典為例」，數位典藏國家型科技計畫 2003 年古籍數位典藏研討會(2003.09.01)，台北：中央研究院歷史語言研究所。
62. 釋惠敏、杜正民、周邦信，2003.12.10「數位化古籍校勘版本處理技術－以 CBETA 大正藏電子佛典為例」，「第一屆文學與資訊科技會議」 (2003.12.09~11)，新竹：清華大學。
63. 釋惠敏、維習安、杜正民、郭麗娟、周邦信，漢文電子佛典製作與運用之研究—以《瑜伽師地論》為例，中華佛學學報，第 14 期，43-53，2001

可供推廣之研發成果資料表

□ 可申請專利　■ 可技術移轉　　日期：__年__月__日

國科會補助計畫	計畫名稱：中文詞彙及跨語詞彙抽取技術在佛典數位典藏上之研發與應用 計畫主持人：黃乾綱 計畫編號：NSC 95－2422－H－002－018－　學門領域：人文
技術/創作名稱	Suffix Array based Term Extraction Algorithm
發明人/創作人	黃乾綱
技術說明	中文：以 Suffix Array 為資料結構，所發展的 Statistical Term Extraction Algorithm，比 PAT-tree based 的 Term Extraction 在時間與空間上更有效率
	英文：A Statistical Term Extraction Algorithms based on Suffix Array data structure, which is more space and time efficient than PAT-tree based Term Extraction Algorithm.
可利用之產業及可開發之產品	可用於專有名詞詞典製作。
技術特點	
推廣及運用的價值	

※　1.每項研發成果請填寫一式二份，一份隨成果報告送繳本會，一份送　貴單位研發成果推廣單位（如技術移轉中心）。

※　2.本項研發成果若尚未申請專利，請勿揭露可申請專利之主要內容。

※　3.本表若不敷使用，請自行影印使用。