歷史佛典文獻外來語借詞對辨識系統

王昱鈞 [*]　蔡宗翰 [**]

同源詞（cognate）及外來語的借詞（loanword）對於研究語源及跨地域的文化交流具有十分重要的意義。能夠從古代的歷史文獻中找出可能的同源詞或借詞對，將對於研究歷史及聲韻學、歷史語言學等有相當的助益。然歷史文獻卷秩浩繁，故本論文提出一個自動自文言文歷史文獻中擷取並辨識出可能的借詞對之系統。本系統主要包括詞彙抽取、候選詞過濾、語音相似度比對三大模組。詞彙抽取我們採用後綴數組抽詞方法，並利用規則方法進行候選詞的過濾，抽取出可能的借詞候選詞。在語音相似度比對上，我們利用漢語中古音之韻書《廣韻》配合漢語中古擬音建立出中古音的發音字典，將漢字轉換爲中古音之語音符號串列。而後利用 ALINE 演算法，結合實際的語音特徵計算二個語音符號串列的相似度，找出可能的借詞對。我們以《雜阿含經》與《阿毘達磨大毘婆沙論》二部佛教典籍作爲實驗集，借詞對辨識的 Recall 達到 0.7446，而 Precision 達到 0.6541。透過實驗證明本系統能夠有效地辨識出二部不同時代不同體裁的佛教典籍中的同源的借詞對，包含人名、地名，以及各種佛教的專有名詞等。藉由我們開發的借詞對辨識系統，能從大量的文言文獻中找出詞彙的語音對應關係，亦能深入了解查找出各詞之間的關聯，從而發掘出更多的研究議題的可能性。

關鍵字：借詞、外來語、聲韻學、語音相似度、歷史文獻處理

[*] 國立臺灣大學資訊工程學研究所博士研究生，E-mail：d97023@csie.ntu.edu.tw。
[**] 元智大學資訊工程學系助理教授，E-mail：thtsai@saturn.yzu.edu.tw。

Loanword Pair Identifying System for Classical Chinese Literature

Yu-chun Wang* Tzong-han Tsai**

Cognates and loanwords take very important roles in many research fields such as historical linguistics and the history of culture influence. Identifying cognates and loanwords from historical literature helps researchers analyze the sound change of languages, phonology of ancient language or other fields. Since the amount of historical literature is tremendous, we propose a loanword identifying system which extract and identify possible loanword pairs from classical Chinese literature automatically. The system comprises three modules: term extraction, term filtering, and phonetic similarity measurement. For term extraction, we adopt suffix array method to extract potential terms from the text. Then, the extracted terms are filtered by manually constructed rules to obtain the possible loanwords. Next, these loanwords are compared with their phonetic similarity mutually to identify the two words are actual loanwords from the same foreign word or not. In order to measure the phonetic similarity of the Chinese characters from classical Chinese literature, the middle Chinese rime book "Guangyun" is used to transliterate Chinese characters into phonological sequences. Then, ALINE algorithm is adopted to measure phonetic similarity between these two phonological sequences by the phonetic features to identify the loanword pair. We take the Buddhist scriptures Samyukta Agama and Mahavibhasa as the evaluation set. The recall of the identifying loanword pairs is up to 0.7446 and the precision is up to 0.6541. The evaluation results show that our system can identify the loanword pairs such as person names, location names, names of Hindu Gods and Buddhist named entities from the Buddhist scriptures which are translated into classical Chinese in different periods. With our system, the phonetic relationship of loanwords can be analyzed from mass historical literature automatically to help researchers in phonology or historical linguistics.

Keywords: Phonology, Phonetic Similarity, Historical Literature, Processing, Loanword

* Department of Computer Science and Information Engineering (CSIE), National Taiwan University，E-mail：d97023@csie.ntu.edu.tw.

** Assistant Professor, Department of Computer Science and Engineering, Yuan Ze University, E-mail：thtsai@saturn.yzu.edu.tw.