Fundamentals of computer-assisted language comparison

Tiago Tresoldi, Mei-Shin Wu, and Nathanael E. Schweikhard

簡介

在歐洲, 歷史語言學創建於 19 世紀, 通過比較不同語言的語料, 進而歸納出語言之間的親屬關係以及語系的演化過程。今日廣為人知的印歐語系即是語言學家藉由一組組的詞彙緩步分析建構而出的第一個語系。由於比照法成功地建造出印歐語系的架構, 語言學家將此方法推廣至世界各地, 進而構築了大大小小族繁不及備載的語系。如此龐大的工程憑藉著語言學家們不斷地收集語料與歸納統整, 已對少數特定語系有清晰的認識。然而, 三百多年之後的今日, 我們對於多數的語系仍然只有模糊的概念, 甚至某些語系的命名恰當與否, 歷史語言學界依舊爭論不休。目前存在的語料提供了龐大且零散的資訊, 僅憑藉著語言學家們做歸納統整與分析, 可想而知研究進度的緩慢是可以預期的。

在過去的數十年間, 許多詞彙語料逐漸電子化, 此舉正因應了跨領域研究的趨勢。電腦可取代重複且枯燥無味的資料前處理, 語音序列比對, 初步的同源詞偵測, 最後找出不同語言之間的聲音對應關係。語言學家可適度地介入各個處理階段做調整並且判讀結果。由於每一個步驟都有不同的挑戰, 舉例而言, 資料的呈現型態就有不同的格式, 不同資料使用的詞彙可能是同義而非相同的詞彙, 音變與意變可能造成同源詞的偵測精準度下降, 電腦不可能完全取代人力。高效率 (Efficiency) 以及一致性 (Consistency) 是資訊處理的特徵, 人腦貴在可適應性 (Flexibility) 以及增進分析正確性 (Accuracy), 兩個領域相輔相成即是我們的計劃核心價值–電腦輔助比照法（computer-assisted language comparison）。

本次演講的主題將電腦輔助比照法分成三個部分。Tiago Tresoldi 博士, 將講解我們已研發的演算法及工具 (Softwares and methods), 博士生吳梅忻會示範如何將已知的工具互相結合成為一個半自動的流程 (Interface and workflows), 博士生 Nathanael E. Schweikhard 會近一步闡述文本註釋以及實例應用 (Data annotation and modeling)。