Fundamentals of computer-assisted language comparison

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

簡介

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, 已對幾個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。研究進度的緩慢是可預期的, 畢竟目前存在的語料提供了龐大且零散資訊, 僅憑藉著語言學家們的力量做歸納統整與分析是毫無效率可言的。

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, 音標序列比對,

Abstract

The comparative method, which is the core of historical linguistics, originated in the 19th century. Linguists compared lexical items in different languages to examine the genealogical relationships among languages. Through repetitively analysis various sets of lexical items among target languages, linguists have categorized various large and small language families in the world. Even though, we have gained in-depth knowledge about certain language families after more than 300 years of studies, but many more questions about the rest of the world await linguists to answer. The issue of comparative method, which relies on experts' manual judgements, has reaches its practical limit as the amount of language data piles up through time.

On the other hand, digitalized language data are also accumulated and able to access easily, it opens up the era which experts can focus on applying their knowledge to examine the results, and let computer power to handle the repetitive and boring tasks. In our term, it is called **computer-assisted language**

comparison. Combining qualitative and quantitative approaches is not a completely new concept, it has been widely used in biology and natural language processing. With the help of computing power, the cross-linguistic data can be easily compiled and manipulated from various individual studies, the cognate detections as well as sound correspondences are detected in merely few hours, plus the algorithms can be flexibly connected into workflows and apply onto many linguistic aspects. The experts could evaluate and make changes to the data at any stage so to increase the accuracy.

In today's three talks, we organized as the following:

(A) Softwares and methods (Tiago Tresoldi)
(B) Interface and workflows (Mei-Shin Wu)
(C) Data and Modeling (Nathanael E. Schweikhard)

With these three talks, we introduce the core value, the methods, the tutorial and the applications of our lab and the tools we are developing.