Advancing explanatory and tonal dialectometry

Sung, H.W.M.

粵語摘要

方言測量學係方言學一個運用電腦同統計學方法去了解語言喺地理上嘅變異嘅分支。本論文利用方言測量學嘅方法去研究粵語同平話嘅方言分區，同時亦借助粵語同平話將方言測量學嘅應用範圍加以擴大。雖然方言測量學可以利用海量嘅數據去做一個比較客觀嘅量化分析，但係到目前為止，方言測量學喺針對研究聲調語言嘅方法上仍然停留喺起步階段，所以需要更多嘅研究。本論文將會探討以下一系列研究問題：1）粵語同平話喺音段上嘅方言分區、2）自動化嘅方言特徵識別、3）粵語喺聲調上嘅方言分區、4）粵方言音段同聲調上嘅分區比較。

粵語同平話嘅數據來自唔同方言調查報告同同音字表。成個數據庫一共有 113 種粵語同平話嘅方言，涵蓋 130 個字嘅讀音（國際音標轉寫）。對於粵語同平話喺音段上嘅方言分區，本文首先利用萊氏距離（Levenshtein distance，又稱編輯距離）嚟計算方言語音距離，然後再用多維縮放（Multidimensional scaling / MDS, 降維技術嘅一種）同埋聚類分析（Cluster analysis）去了解粵語同平話方言嘅內部結構。結果顯示，傳統分類上嘅桂北平話唔屬於粵語嘅方言連續體，但傳統分類上嘅桂南平話就同粵語形成方言連續體。基於呢個結果，本論文視桂北平話為離群點（outliers），以便對粵語方言連續體作更加深入嘅探討。利用剩餘嘅 104 個方言點進行進一步嘅研究後，發現粵方言大致可以分為兩到五個方言群，具體劃分視乎分析層面嘅細微度。

方言計量學嘅分類法成日都因為缺乏對方言分群嘅細節或解釋而受到批評。咁係因為方言測量學嘅分類需要將唔同方言嘅特徵差別量化成唔同方言之間嘅距離，但一般嘅量化過程其實好難再喺距離計算之後提取返具體嘅方言特徵。因此，喺計算粵語音段分類的方言距離之前，本論文採用咗多重序列比對（Multiple Sequence Alignment, MSA）。MSA 能夠將所有方言嘅語音轉寫標記進行排序 (alignment)，有助於識別具有歷史關聯嘅音段，並對其進行分拆。呢個操作令到語音轉寫標記更容

易喺歷時比較做更精準嘅比對，亦有利於後續分析，例如自動化嘅方言語音特徵提取。經 MSA 處理過嘅語料亦可以輕鬆地套用喺常規嘅方言測量學分析流程裡面——直接用嚟計算方言距離，再進行聚類分析同多維尺度分析等，劃分方言群。最後，利用一種自然語言處理中常用嘅關聯度量度方法——點向互資訊（Pointwise Mutual Information），我哋就可以識別出同每個方言群有密切相關嘅方言特徵。以上嘅方法大大增強咗方言計量學對大規模方言數據嘅嘅詮釋力，令方言測量學超越單純嘅方言分區。

另一方面，雖然聲調語言喺世界上幾常見，但係利用方言測量學方法去研究聲調語言嘅研究其實唔多。如果要運用方言測量學去研究聲調語言嘅話（方言聲調測距法, dialect tonometry），其中一個問題就係究竟應該點樣測量聲調距離呢？目前文獻裡面有幾種方法，例如簡單嘅二元相似度對比、基於感知嘅距離計算方法。本論文將萊氏距離應用喺一種改良咗嘅聲調標記法嚟計算方言嘅聲調距離，並探索唔同方言嘅聲調喺空間上嘅變化同關係。呢種改良咗嘅標記法叫 modified Onset — Contour — Offset 標記法（以下簡稱 mOCO），來自 Yang and Castro (2008) 嘅 Onset — Contour — Offset 標記法。同其他方法相比，mOCO 係一種更適合方言計量學嘅聲調標記方法，因為佢能夠區分數據裡面接近 99% 嘅聲調，而且呢個方法亦有助呈現聲調之間嘅漸進式距離，符合人類嘅感知。

將 mOCO 用喺粵語方言嘅聲調方言測量學分析，結果顯示音段同聲調嘅變化模式其實並唔相同；音段變化喺地理上呈現連續性模式，而聲調變異就呈現出更有清楚方言區域嘅模式。進一步的分析表明，並唔係所有方言區域都同時出現音段同聲調層面嘅分區。呢個差異反映方言有趣的複雜情況，本論文就此提出咗一啲有待進一步研究嘅問題。