論文提要

梳理文本的時間順序對真實性研究和詮釋至關重要。對一些西方語言來說，統計語言模型（*SLMs*）已被證明有助於將文本劃歸到一個歷史節點，即一個時段。這是通過語法、語音和詞匯以及歷時文體風格之變化進行的。中國書面資料因偽造、部分流傳史複雜和作者身份不明等因素，再加上某些文本體裁在語言上頗多取法古代範文，殊難確定產生年代。此外，漢文幾乎反映不出實實在在發生過的語音變化。本論文試圖通過研究比較計算機輔助文本測定的新老方法的應用來探討這一問題。

論文是以古漢語史為出發點的。官方正史研究的結果表明，文本即使具有同質的文體風格，其語法和詞匯的變化也能為確定文本產生的時間提供線索。新術語和新概念的出現為新構詞匯創造了條件，而新構詞匯被證明是文本產生時間的一個關鍵性指標。

本論文側重統計語言模型，對計算機輔助和文字學方法的文本斷代研究現狀作了綜述。接下來討論了一些基本性的挑戰：諸如古漢語文本的分節問題、缺乏歷時漢語語料庫，以及識別命名實體（*NER*）和時間概念的困難。為了使詞匯變化能更適用於斷定時間，還新建了一個歷時漢語詞匯庫。這是以《漢語大詞典》數字版中的詞條例證（*loci classici*）為基礎的。

論文的主要部分致力於開發、測試和比較古漢語文本的年代測定方法，使得統計語言模型適用於中文 ***n*** 元語法和純文本語料庫。作為一種創新探索，我們介紹了用詞庫生成文本的新詞條圖。把有具體時間的命名實體和時間概念納入考量，可以進一步提高確定時間的準確性。為了將時間作為連續變量，文本中單詞詞匯化的平均年份也被作為一個實驗性的斷定年代的指標進行測試。結果發現，統計語言模型可以成功地確定古漢語文本的產生時間。事實證明，新詞條圖更抗拒書面語言的風格僵化，也較少需要專門訓練，此外還可用於協助文字學工作。所評估的三種方法都可以通過在本研究框架下開發的在線工具 *VisualTime*（https://visualtime.schalmey.de）進行測試。