国际中国语言学学会第 21 届年会(IACL 21)的工作坊--——歷代语言知識库的建构与应用

蒋经国国际学术交流基金会《历代语言知识库建置计划》（2009）

先秦典籍及明清小說語言知識庫的建構方法

The Construction of Chinese Diachronic Knowledge Base for Pre-Qin and Ming-Qing Classics

陸勤[1], 熊丹[1], 羅鳳珠[2], 石定栩[3], 趙天成[1]

[1]香港理工大學電子計算學系 [2]台灣元智大學中國語文學系 [3]香港理工大學中文及雙語學系

[1]{csluqin,csdxiong,cstschiu}@comp.polyu.edu.hk, [2]gefjulo@saturn.yzu.edu.tw, [3]ctdshi@polyu.edu.hk

明清的語體文可視作文言文至現代漢語的過渡語體，其知識庫的建構在設計理念、規範制定、工作流程、質量保證方面，都可作為先秦典籍知識庫建構的參考。但先秦典籍的語言和明清小說存在較大差異，例如，單音節詞佔據更主導的地位、語言更精練、句法更靈活、文言虛詞使用頻繁、詞類活用情況普遍，致使詞義豐富。因此，先秦典籍標註語料庫的建構主要基於先秦漢語的特徵，同時盡可能地沿用和借鑒明清小說語料庫建構規範[1]。

台灣中央研究院的上古漢語標記語料庫[2]也包括了先秦典籍，該語料庫對全文詞類進行了標註，但命名實體沒有進一步細分，例如，將姓氏名號統一劃歸“人獸名（有生）”，如“邾儀父(NB1)[+prop]”，稱謂和普通有生類名詞均劃歸“有生名詞”，如“夫人(NA1)[+attr]”。本項目的工作重點為分詞和命名實體標註，

并根據先秦典籍中命名實體的特點對其進行逐層細分。

先秦典籍分詞的基本原則和明清小說一致，即以語義為出發點，致力於在不造成語義丟失、轉換、引申或歧義的情況下，切分到最小語義單位。由於語義的差異，同一字串的處理方式可能不同，例如“四國”如果為實數，指“四個國家”，則切分；如果為虛數，泛指四方諸侯國，則不切分。另外，還從詞彙和句法結構兩個角度規範了一些針對先秦典籍的輔助分詞原則，例如，動賓結構儘量切分，如“定/社稷”；但如果字串的組合具備引申意義，則不能切分，如“蒙塵”。

在命名實體標註方面，總體上仍繼承明清小說的標註方法，對多層次複合命名實體採用統一的標註系統：以嵌套的方式體現複合命名實體的組合關係，如內部成分的類別與複合命名實體的類型相異，則保留其獨立標識符，如“[邾/ns1 儀父/nr4]/na1”（國名+字組成）。然而，由於先秦典籍中的姓氏名號更加複雜，無法完全照搬明清小說中人名的分類和標註方法，例如，明清小說中完整的人名組合為“姓氏+名”或“姓氏+字”，因此人名無需採用嵌套的方式；而先秦典籍中的人名則出現了更複雜、更靈活的組合，如“[百里/nr1 孟明/nr4 視/nr2]/nr7”由“姓氏+字+名”組合而成，因此也需要以嵌套的方式保留其內部成分的獨立標識符，以便計算機對命名實體的識別和資訊提取。

參考文獻：

1. 李昀燕,熊丹,陸勤,羅鳳珠,石定栩,趙天成.明清章回小說的分詞準則及命名實體標註.第十三屆漢語辭彙語義學研討會(CLSW2012) 論文集,2012:16-21.
2. 魏培泉,譚樸森,劉承慧,黃居仁,孫朝奮.建構一個以共時與歷時語言研究為導向的歷史語料庫. Computational Linguistics and Chinese Language Processing, 1997, 2(1):131-145.