基於語料庫的明清小說人名、稱謂分類及組合分析*

熊丹[1]，陸勤[1]，羅鳳珠[2]，石定栩[3]，趙天成[1]
香港理工大學 [1]電子計算學系 [3]中文及雙語學系
[1]{csdxiong,csluqin,cstschiu}@comp.polyu.edu.hk [3]ctdshi@polyu.edu.hk
[2]台灣元智大學中國語文學系
gefjulo@saturn.yzu.edu.tw

摘　要：人物稱謂作為一種重要的語言單位，不僅承擔著信息載體的功能，更是表達思想、抒發情感的有效方式，因此在文學作品中會以各種形式大量出現。在自然語言處理及其應用領域，人物稱謂是命名實體識別的一個主要部分，是語言處理的關鍵因素之一。基於一個以人工標註爲主而計算機為輔建構的四本中國古典文學名著的語料庫，本文從命名實體識別和資訊提取的角度出發，將語料庫中的人名和稱謂分為單一型和複合型兩大類，根據複合型的內部組成元素和組合方式，將其進一步分為固定式、同位式、附屬嵌套式、靈活嵌套式，并結合數據對各種類型進行深入分析，探索其組合規律，此項工作是一項對古典文學中有關人名、稱謂所做的基於語料的、較有系統性的研究工作。

关键词：人名和稱謂，明清小說，語料庫，命名實體

A Corpus-Based Study of Categorization and Composition of Names and Addressing Terms in the Novels of the Ming and Qing Dynasties

Dan Xiong[1], Qin Lu[1], Fengju Lo[2], Dingxu Shi[3], Tin-shing Chiu[1]
[1]Department of Computing, [3]Department of Chinese & Bilingual Studies, The Hong Kong Polytechnic University
[1]{csdxiong, csluqin, cstschiu}@comp.polyu.edu.hk [3]ctdshi@polyu.edu.hk
[2]Department of Chinese Linguistics & Literature, Yuan Ze University, Taiwan
gefjulo@saturn.yzu.edu.tw

Abstract: As an important linguistic category, addressing terms not only carry particular information, but also express feelings and emotions. They are therefore widely used in literature works. In natural language processing (NLP) and its applications, addressing term is one of the key elements in named entity recognition, which can affect the overall performance of an NLP system. Based on the analysis of a manually annotated corpus of four Chinese classical novels in the Ming and Qing dynasties, this paper presents a classification system for personal names and addressing terms from the perspective of named entity recognition and information extraction in NLP. Personal names and addressing terms are categorized into simple type and compound type and the compound-type is further categorized into four sub-types, namely, fixed expressions, appositive constructions, embedding constructions, and other embedding constructions.

Keywords: names and addressing terms, novels in the Ming and Qing dynasties, corpus, named entity

1　前言

稱謂是人與人交際過程中必不可少的語言單位，在語言中承載重要的資訊，是有效理解和處理自然語言的關鍵因素之一。不同的歷史時期有不同的稱謂習慣，是當時歷史文化以及人際關係複雜程度的一種反映。在文學作品中，稱謂不僅能夠傳遞資訊，更能幫助刻畫人物形象，使語言顯得生動，因此在古典文學作品中不僅應用廣泛，而且種類紛繁、形式多樣。本文利用一個以人工標註爲主而計算機為輔建構的四本中國明清古典文學名著[1]的語料庫[1]，從命名實體識別和資訊提取的角度，經過對數據的完整統計和歸納，分析明清小說人名和稱謂的組成元素和組合方式，在此基礎上進行綜合分類、並探索複合型稱謂內部成分的組合規則。

*本文承九十八年度蔣經國國際學術交流基金會“歷代語言知識庫建置計畫”（RG013-D-09）部分資助。
[1] 包括《紅樓夢》、《三國演義》、《水滸傳》、《金瓶梅》。

2 研究概況

稱謂自古以來就引起了人們廣泛的研究興趣。《爾雅• 釋親》[2]通常被認為是中國最早的一部系統介紹親屬稱謂的著作，書中分類介紹了基於血緣和姻親關係的親屬類稱謂。此後的很多朝代都出現過稱謂語的研究文獻，清代梁章鉅所撰的《稱謂錄》[3]是古代稱謂研究的一部集大成之作，對基於血緣關係、社會地位、身份、官職等多種因素形成的稱謂進行了系統性的彙編和溯源。當代學術界對稱謂的研究也層出不窮（如[4]所做的綜述），針對明清文學中人物稱謂的研究同樣成果豐碩，其角度涵蓋語言學、社會學、文化、歷史、翻譯等多個領域，研究的細微性各不相同，有對人物稱謂的綜合性研究，也有對某類書籍中的稱謂、某個人或某類人的特定稱謂等進行的深入研究。

在自然語言處理及其應用領域，稱謂是一種重要的命名實體，但目前的漢語資訊處理中，未對稱謂及稱謂和人名組合的名稱進行細分。國內較通用的現代漢語語料庫加工規範[5]中，將人名作為一類單獨的實體（標識符為“/nr”），漢族姓、名分開標註，如“張/nrf 仁偉/nrg”。但對於姓名後附加職務或稱謂的名稱，將職務、稱謂只作為普通名詞（標識符為“/n”），例如“李/nrf 主席/n”、“劉/nrf 阿姨/n”。台灣中央研究院建立的近代漢語（唐以後）標記語料庫[6][7]，對詞類進行了標註。該語料庫也包括了《紅樓夢》等明清文學，並將姓名劃歸專有名詞（標識符為“Nb”），稱謂也只劃歸為普通名詞（標識符為“Na”），例如“政(Nb)老爹(Na)”。

由於小說中的稱謂出現頻繁、靈活多變、且相比其他文體更能反映相應時代的文化和語言特色，因此，在建立明清小說標註語料庫時，將其作為一類主要的命名實體進行標註。這種分類需要從資訊處理的角度對稱謂進行分析，不同於一般的文學和歷史分析中對稱謂的研究，同時要考慮如何在資訊平臺上為文學和歷史研究提供方便。本文主要從稱謂的內部元素和組合方式入手，對人名和稱謂進行整合分類，探索複合型稱謂的組合規律，並將總結的規則實際應用到四本名著的命名實體標註中。

3 人名、稱謂的分類

3.1 稱謂的界定

長期以來，關於稱謂的概念、範疇、甚至稱謂和稱呼之間的關係一直存在多種看法、沒有定論（如[4]介紹的幾種主流觀點）。本文對稱謂的定義不予深入探究，而是採用一個廣義的概念，既包含人與人之間言語交際中所使用的直接稱呼，也包含提及某人時使用的指稱性名稱。從這一意義來看，本文將作為人物身份標識和角色定位的官職、爵銜都納入稱謂之列。另外，本研究展示的數據僅包含特指某一人物、並根據上下文語境能判斷其所指人物的稱謂，例如《紅樓夢》中的“姑娘”，如果能夠判斷其所指的對象，則加稱謂標註，而“一/個/姑娘/領著/他”、“姑娘/們”等非特指，則當普通名詞處理，不加稱謂標註。

3.2 分類的目的

在對四本名著的研究過程中發現，小說中對同一個人物的稱謂隨著該人物身份、地位、所處的場合、交流的對象、甚至當時作者想要體現的感情色彩而不斷變化。稱謂可單獨使用，也可通過不同形式靈活組合而成，例如，可以從姓名、字號中截取一部分再加上頭銜組合而成。這種複合型稱謂各元素之間的組合關係相當複雜，需要對這些稱謂進行系統性的分類、並對其組合關係進行分析，既要確保古典文學標註語料庫的建設過程中，採用統一原則進行分詞和標註，又要兼顧靈活性而有利於後續研究，如建立文本內及不同文本間相關命名實體的關聯，基於命名實體為人物屬性建立檔案等。

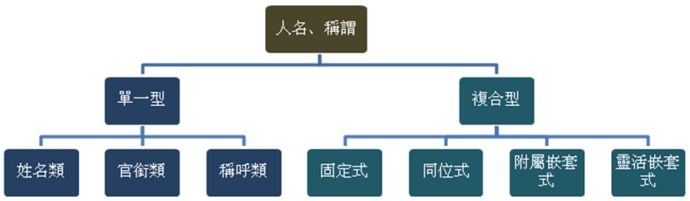
需要進一步說明的是，稱謂可能由不同的分詞單位組合而成，其分詞原則與小說正文的分詞原則相同[1]，以保持全文分詞的一致性、提高系統訓練的效率和計算機輔助的有效性。人名、稱謂的標註主要沿用北京大學詞性標註系統[5]，對於該系統中沒有的類型則新增標識符。如果稱謂中包含其他類型的命名實體，如地名、機構名，則以嵌套方式保留其獨立標註。下文通過實例闡述人名、稱謂的組成元素及其組合方式，本文所有實例均取自於已標註的四本名著。

3.3 人名、稱謂的總體分類

由於社會結構、文化背景的差異，在不同的時代、地域、以及社會群體中，稱謂具有明顯的特徵。而小說爲了凸顯其藝術效果，使用的稱謂更是變化多樣。例如《金瓶梅》中的蔡京，雖然不是小說主要人物，卻使用了多種指稱方式。既有直接用單姓“蔡”和姓名“蔡京”進行指稱

的，也有用官職指代的，如“左丞相”、“大學士”、“吏部尚書”、“太師”等。下屬、僕役稱呼他時會用“老爺”、“蔡老先生”、“蔡太師”、“太師爺”、“老太師”、“太師老爺”、“蔡太師老爺”等，而內相們私下談論時則貶稱為“老賊”。另外，本文採用的語料文本雖然是明清時期創作的小說，但其故事所處的時代背景、社會環境都不盡相同，不同程度地折射出秦漢、遼宋、明清等多個時期的文化形態和社會風貌，而且故事人物的社會角色千差萬別，因此語料中出現的稱謂非常豐富。

基於對四本名著人名和稱謂的綜合分析，本文從其組成元素和組合方式的角度進行了綜合分類，總體上分為**單一型**和**複合型**兩大類，顧名思義，前者由人名、稱謂本身獨立承擔指稱功能，後者由多個成分疊加或嵌套組合而成。複合型稱謂的內部組合方式非常靈活，有的是由多個獨立使用的單一型人名、稱謂疊加而成，例如“[令郎/先生]/na2”，其中“令郎”和“先生”都可以用來作為獨立的稱謂；有的是截取人名的一部分、再和稱謂合併而成，例如“[鳳/nr2 姐姐]/na1”；還有的是由人名附加修飾、描述語組合而成，例如“[周瑞/nr3 家/的]/na1”，其中“家/的”不能獨立作為稱謂，一般附加於人名後組合成複合稱謂。經過對語料中的人名、稱謂進行歸納分析，本文從其內部構成及其組合方式入手分為以下大類。



圖一：人名、稱謂的總體分類

4 單一型人名、稱謂細分

如圖一所示，單一型人名、稱謂可分為三大類，本小節對其進行了細分，並用實例說明。爲了便於理解，取自語料中的實例均保留語料庫中的標識符號和標註形式，下表還列出北京大學現代漢語語料庫[5]採用的相應標識符，以便參照。

表一：單一型人名、稱謂細分

一級子類	二級子類	本語料標識符	北大標識符	定義及說明	實例
姓名類	姓	/nr1	/nrf	用於特指某人物的姓，包括單姓、複姓、多音節姓。	林/nr1 史/nr1 二/人
	名	/nr2	/nrg	用於特指某人物的本名。	黛玉/nr2
	姓+名	/nr3	/nrf /nrg	用於特指某人物的姓+名。	林黛玉/nr3
	字	/nr4	/nrg	用於特指某人物的字，通常為雙音節，有少量為單音節。	士隱/nr4
	姓+字	/nr5	/nrf /nrg	用於特指某人物的姓+字。	甄士隱/nr5
	別名	/nr6	/nr	用於特指某人物的、所有除本名之外的名稱，包括別號、謚號、帝王廟號、不能確定姓和名的外族名等。	顰兒/nr6 宋徽宗/nr6#[2] 金環三結/nr6（外族名）
官銜類	官職	/nu1	/n（普通名詞）	具有特指意義的官職名。	太師/nu1
	爵位、封號	/nu2		具有特指意義的爵位、封號名，包括帝王根據血緣親疏、功勞等授予的爵銜、尊號，含對皇室、宗室女子的封號。	郡王/nu2 貴妃/nu2

[2] “#”表示歷史人物，爲了便於後續研究，標註語料時通過使用特殊標識符號對小說引用的歷史人物進行了區分，以二十四史為依據。

一級子類	二級子類	本語料標識符	北大標識符	定義及說明	實例
稱呼類		/na2	/n	人與人交往中基於血緣關係、社會地位、身份、宗教等各種因素對某一特定人物的稱呼，既包括當面交流時直接稱呼對方所使用的名稱，也包括提及他人時的間接指稱，不含自稱。	• 老祖宗/na2（通常當面交流時使用） • 祖母/na2（向他人提及時使用）

5 複合型稱謂細分

單一型稱謂可獨立用於指稱，也可作為複合型稱謂中的單元成分。對各類複合型稱謂採用統一的標註系統：使用“[ ]”總括，如內部成分的類別與複合稱謂的類型相異，則保留其獨立標識符。基於內部成分的組合關係，複合型稱謂可分為四大類，下文通過實例進行描述。

5.1 固定式組合

這一類型是由多個成分組合而成的較固定的名稱，其內部成分一般不分開使用、或分開後僅作為簡稱使用，例如：

- 以美號賜封的爵位和封號：帝王封爵時賜予的美號和爵銜組合而成的名稱具有特指性、較固定，例如“[北靜/郡王]/nu2”，“[順平/侯]/nu2”。因爲“北靜”和“郡王”作爲單一成分均為爵位，與其複合稱謂一致，而無需再加獨立標識符。
- “名號＋將軍”組合而成的武將官職：對有軍功者授予“將軍”官銜時會冠以名號，例如“[奮威/將軍]/nu1”，“[冠軍/將軍]/nu1”。

5.2 同位式組合

這一類型由多個存在同位關係的成分堆疊而成，其內部成分為同一類型，可分開後獨立使用，例如“[父親/大人]/na2”，“[都太尉/統制]/nu1”。

5.3 附屬嵌套式組合

這一類型由兩個存在附屬、支配或依存關係的成分組合而成，其內部成分一般為不同類型、但具備依存關係，其中一個成分一般不能獨立用作稱謂。主要包括：

- 封地＋爵位、封號：如果封爵時賜予了封地，爵位、封號名用作稱謂時通常會附帶封地名，例如“[烏程/ns2# 侯]/nu2”。“烏程”為地名，因此保留其地名標識符（/ns），而“侯”則無需重複爵位標識符(/nu2)，系統可默認識別。
- 管轄地＋官職：人物的官職經常和其管轄地連用，爲了不使這一信息丟失，將其作為一個複合型命名實體，例如“[揚州/ns2# 刺史]/nu1”[3]。
- 機構＋官職：小說中提到官職時，往往還會採用“機構+官職”這一組合形式，爲了保持兩者之間的關聯，便於後續的信息提取，將這兩個命名實體作為一個複合型命名實體，例如“[吏部/nt 尚書]/nu1”。

5.4 靈活嵌套式組合

這一類型包括所有其他由兩個或兩個以上的成分靈活嵌套組合而成的複合型稱謂，其內部成分可以是單一型人名、稱謂，也可以是以上幾種複合型稱謂。無論其內部成分多麼複雜，都可逐層剖析成單一型人名、稱謂後使用統一的標註規則進行處理。從其內部組合方式劃分，靈活嵌套式組合可進一步分為七類，在下表通過實例說明。單一型“稱呼類”實體的標識符為“/na2”，對靈活嵌套式組合的複合型稱謂使用“[ ]”總括，并加“/na1”作為總標識符。如這一組合的內部成分為單一型“稱呼類”實體，無需再加“/na2”標識，系統可默認識別，例如，“[蔡/nr1 老爺]/na1#”中的“老爺”是一個單一型稱呼，無需再加獨立標識符。如內部成分為其他類型實體，則需保留其獨立標識符，例如“[蔡/nr1 太師/nu1]/na1#”。

表二：靈活嵌套式組合細分

[3]指稱關係是受時空限制的，在特定的時間地點，某個官銜和某個人物有一一對應的關係，但時過境遷，擔任這個官職的人物變了，指稱關係會相應變化。

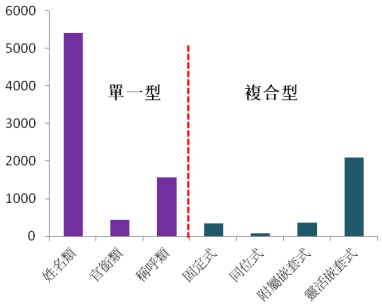
組合方式	實例	說明
姓、名＋稱呼	[蔡/nr1 老爺]/na1# [玄德/nr4# 公]/na1#	由各種形式的姓、名（包括“姓名類”所有二級子類）附加稱呼組成。
官職＋稱呼	[太師/nu1 老爺]/na1#	由官職附加稱呼組成。
爵位、封號＋稱呼	[[臨安/伯]/nu2 老太太]/na1	由爵位、封號附加稱呼組成。
姓、名＋官職	[蔡/nr1 太師/nu1]/na1#	由各種形式的姓、名附加官職組成。
姓、名＋官職＋稱呼	[[蔡/nr1 太師/nu1]/na1# 老爺]/na1#	此例中，“[蔡/nr1 太師/nu1]/na1#”作為一個組成元素，本身已是複合型稱謂。
姓、名＋爵位、封號	[賈/nr1 妃/nu2]/na1	此例是由姓和封號“妃”組合而成的複合型稱謂。
姓、名＋爵位、封號＋稱呼	[[賈/nr1 妃/nu2]/na1 娘娘]/na1	此例是由“姓+封號”後再附加“娘娘”這一稱呼組合而成的複合型稱謂。
說明： 1. 以上各種組合的內部成分先後順序不定，例如“官職＋稱呼”組合，其內部成分的順序也可能是“稱呼+官職”，如“[義士/提轄/nu1]/na1”。 2. 以上組合中，任何一種內部成分的數量不定，例如“姓、名+稱呼”組合中，可能出現多個稱呼，如“[[晁/nr1 頭領]/na1 哥哥]/na1”。		

6 數據分析

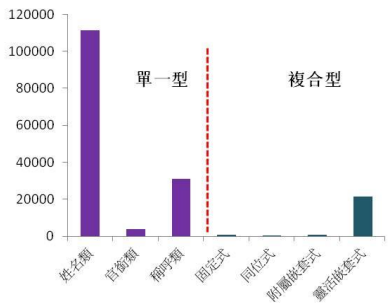
6.1 綜合數據

本文基於四部小說中人名、稱謂的構成方式，將其分為“單一型”和“複合型”兩大類。以下兩個圖形分別從人名、稱謂的類型及其在文中出現的頻率這兩個角度展示各類型分佈。圖二展示的數據對文本中同一人物的相同人名、稱謂只計一次，重點體現類型分佈；圖三則對各類人名、稱謂的次數分別進行了統計，重點體現頻率。需要說明的是，這一數據包括小說中出現的所有人物，含小說塑造人物和小說中引用其他處的人物。

無論從類型還是頻率上看，“單一型”比例都大於“複合型”。這一結果和預期有所不同，古典文學中很多稱謂較長、形式複雜，直覺認為複合型稱謂會更多，但從絕對數量上看結果相反。其中一個主要原因是小說體裁採用第三人稱敘述故事，使用姓名的語境很多，因此姓名類所占比重最大。另外，古代人名的形式多種多樣，除了組成現代人名的姓和名之外，還存在字、別號等其他形式的名稱，可選擇性強，因此使用姓名的頻率很高。在複合型稱謂中，靈活嵌套式組合的數量最多，這一結果和預期一致，因為這種組合能幫助塑造人物形象，增強小說語言的吸引力。



圖二：人名、稱謂類型分佈（個數）



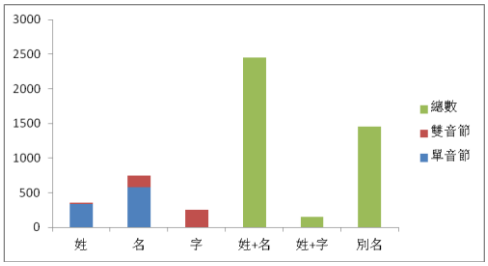
圖三：人名、稱謂頻率分佈（次數）

6.2 姓名類數據分析

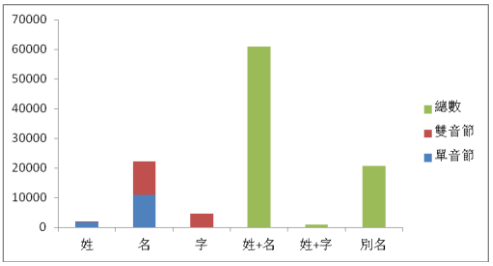
鑒於姓名類在小說中出現的頻率最高，因此有必要對其數據進行進一步分析。圖四展示姓名類中不同類型的數據，對文本中同一人物的相同名稱只計一次，圖五則對其出現的次數分別進行統計。由數據可見，“姓+名”的比例最大，這和現代日常交流中稱謂的使用規律一致。值得注意的是，小說中單音節“名”[4]的比例較大，從類型上看約占“名”的 77%，尤其是《三國演義》中，絕大多數為單音節名，這在一定程度上反映了當時的姓名文化。另外，古典小說中使用別名

[4]包括“名”為雙音節、但用於指稱時僅使用一個音節的情況，例如“迎/nr2 、/探/nr2 、/惜/nr2 三/人”。

的頻率很高，因為古代別號、諡號等名號非常豐富，爲了渲染場景，小說中還大量使用了昵稱和綽號。“字”是中國古代姓名文化中的重要元素，通常為雙音節，單音節的比例僅約 1%，以“字”和“姓+字”作為稱呼的情況較少。



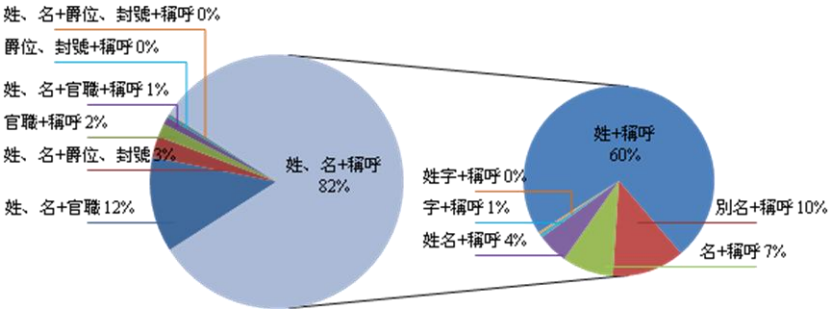
圖四：“姓名類”類型分佈（個數）



圖五：“姓名類”頻率分佈（次數）

6.3 靈活嵌套式組合數據分析

靈活嵌套式組合是複合型稱謂中比重最大的一類，其組合靈活多變、內部成分相對複雜，在語料中長度也很突出，例如“[[梁/nr1 中書/nu1]/na1 相公]/na1”，由一個“姓＋官職”組合而成的複合型稱謂作為內部成分，再附加單一型稱謂“相公”組成多層次的複合型稱謂。圖六顯示了 5.4 所描述的七類靈活嵌套式稱謂在類型上的比例分佈情況（即同一稱謂只計一次），由此可見，比例最大的組合是“姓、名+稱呼”，這在一定程度上也是因歷史上人名形式的多樣性所致。圖上的子餅圖是對“姓、名+稱呼”這一子類所做的進一步分析。結果顯示，“姓+稱呼”的比例最大，因為在對話中使用較多，這也是小說語言的特徵之一。



圖六：靈活嵌套式組合分析

7 結語

鑒於明清小說語料中人名、稱謂的重要性及複合型稱謂組合的靈活性，本文從命名實體識別和資訊提取的角度對其進行逐層分類，其目的在於幫助理解、識別、處理和提取這一類型文學作品中的人名和稱謂。此項工作是基於現有古典文學語料庫對人名、稱謂所做的全面、綜合性分析，對使用計算機技術進行古典文學研究意義重大。在後續的研究中，可以和更早期時代的語料或現代語料進行比較分析、明確不同時代的差別。另一項頗有意義的工作是在本文分類研究的基礎上建立文本內指稱對象的關聯，進一步為文學和歷史的研究提供基礎關聯信息。

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] 李學勤主編,（晉）郭璞注.爾雅注疏.北京:北京大學出版社,1999.

[3] （清）梁章鉅.稱謂錄.北京:中華書局,1996.

[4] 鄭爾寧.近二十年來現代漢語稱謂語研究綜述.語文學刊,2005(2):120-122.

[5] 俞士汶,段慧明,朱學鋒,孫斌,常寶寶.北大語料庫加工規範：切分• 詞性標注• 注音. Journal of Chinese Language and Computing, 2003, 13(2):121-158.

[6] 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.

[7] 中央研究院近代漢語語料庫網址：http://early_mandarin.ling.sinica.edu.tw/