明清章回小說的分詞準則及命名實體標註*

李昀燕[1] 熊丹[2] 陸勤[3] 羅鳳珠[4] 石定栩[5] 趙天成[6]

摘　要: 文本分詞及標註的質量對自然語言處理及與之相關的後續應用有重要的意義。本文表述對明清章回小說制定的分詞準則，其目的是希望能夠對中國文學從古漢語的韻文、文言語體文，到現代漢語的詞彙和語義的變化進行計算機的輔助分析。從文學分析的角度出發，更對明清章回小說的命名實體制定了基本的標註原則，並依此規範對明清章回小說《紅樓夢》進行了計算機輔助的分詞和標註。本文所建立的分詞準則著眼於明清小說與現代漢語對於詞彙應用及語義的同異性，並適度地沿用和借鑑了北京大學和臺灣中央研究院的分詞和詞性分類規範。

关键词：分詞與詞性標註規範，命名實體，明清章回小說，計算機輔助標註，語義分析

The Segmentation Principles and Named Entity Annotations for Fictions in the Ming and Qing Dynasties

Li Wanyin, Xiong Dan, Lu Qin, Lo Fengju, Shi Dingxu, Chiu Tin-shing

Abstract: The quality of text segmentation and annotation plays a significant role in Natural Language Processing especially in downstream applications. This paper presents a set of segmentation principles and named entity annotation targeted for fictions of the Ming and Qing dynasties. The purpose of this work is to lay the foundation for computer aided lexical semantic analysis of classical Chinese literature. Hopefully the result of this work can provide insights into the transition of Chinese literature from its traditional forms such as traditional verses and vernacular styles to modern Chinese. To assist in literature study, an elaborate named entity annotation scheme is also developed. Computer aided segmentation and named entity annotation are conducted on the famous Ming and Qing Chinese novel *Dream of the Red Chamber*. The specification for the segmentation and annotation is produced based on the studies of the morphology and semantics differences as well as similarities between traditional classical Chinese and modern Chinese with reference to segmentation and annotation principles developed at Peking University in Mainland China and Academia Sinica in Taiwan.

Keywords: segmentation and PoS principles, named entities, fictions in the Ming and Qing dynasties, computer aided annotation，semantic analysis

1　前言

隨著資訊科技的發展以及大量文獻的數字化，各種語料庫逐漸建立，大規模真實文本內容計算和理解的要求日益迫切。自然語言處理在語法和語義等方面形成了一些理論體系和計算模型，在機器翻譯、資訊檢索、資訊提取等重要領域也取得了初步成果。

詞彙語義知識庫和詞彙語法知識庫是計算機能否實現文本內容理解的關鍵因素，詞彙語義的計算和理解成為語義分析中最為關鍵的一步。在此需求的推動下，台灣元智大學、北京大學、香港理工大學、日本早稻田大學、韓國首爾市立大學啟動了“歷代語言知識庫建置計畫”。該項目結合以上五所大學已有的歷代文本語言之語料資源、漢語教學資源、技術資源，致力於通過詞彙語義、語法的計算和理解、語義標記與語義概念的分類、語法資訊的建構，建立包含“詞彙語義知識”與“詞彙語法知識”的中國歷代語言知識庫，用以分析、判

* 本文係蔣經國國際學術交流基金會資助的“歷代語言知識庫建置計畫（Building a Diachronic Language Knowledge-base）”（計畫編號：RG013-D-09）的階段性研究成果之一。

1236 香港理工大學電子計算學系，香港九龍紅磡
Department of Computing, The Hong Kong Polytechnic University, Hung Hom, Kowloon, Hong Kong
E-mail: csclaireli@gmail.com, csdxiong@comp.polyu.edu.hk, csluqin@comp.polyu.edu.hk, cstschiu@comp.polyu.edu.hk
4 台灣元智大學中國語言文學系，台灣桃園縣中壢市遠東路 135 號 32003
Department of Chinese Linguistics & Literature, Yuan Ze University, 135 Yuan-Tung Road, Chung-Li, Taiwan 32003, R.O.C
E-mail: gefjulo@saturn.yzu.edu.tw
5 香港理工大學中文及雙語學系，香港九龍紅磡
Department of Chinese & Bilingual Studies, The Hong Kong Polytechnic University, Hung Hom, Kowloon, Hong Kong
E-mail: ctdshi@polyu.edu.hk

斷、提取篇章主題特徵，使計算機具備理解語義、語法與解讀文本篇章主題的能力，并保持不同文體各自獨立的語言知識庫。另外，以語義分類建立語義概念對映機制的語言知識庫設計，既可保存不同時代、不同文體的語言特徵，又能將相同語義概念的詞彙予以貫穿，其成果可應用於語言、文學、資訊科學的研究以及漢語教學等多個領域。

本研究對明清小說命名實體和非命名實體的分詞可作為漸進至文言文分詞的基礎。為了方便表述，本文首先通過實例闡釋明清小說中命名實體的分詞和標註原則，然後以語義規範和語法規範為出發點，將分詞原則歸類為語義相關原則和語法相關原則。另外，鑒於韻文需要採用不同的分詞準則，本階段僅對韻文進行類別標註，暫不進行分詞。本文採用的實例均取自于運行文本《紅樓夢》[1]。

2 分詞和命名實體標註的理念

“歷代語言知識庫建置計畫”的目的是建立含有“詞彙語義知識”、“詞彙語法知識”及“篇章主題分類（特徵）知識”的現代漢語、中國歷代語體文、中國歷代韻文三種語言知識庫，並且以語義概念分類為基礎，建立三種語言知識庫的語義對映機制，最終整合成為貫穿不同時代、不同文體的“中國歷代語言知識庫”。

明清章回小說的語言和現代漢語有些差異，在詞彙和語義方面表現最為明顯。對這些詞彙的正確識別及對詞典的更新有助於後續的自動分詞和語義分析等工作。首先，兩者在詞彙上的區別是顯而易見的，例如，“一心一計”相當於現代漢語中的“一心一意”[2]。更為重要的，有些詞彙表面形式相同，但語義上有很大差異。例如，“老公”是古代對宦官的俗稱，在現代漢語中則指丈夫。也有一些詞彙是現代漢語中不再使用的，例如，《紅樓夢》所寫的“男/女/先兒”中的“先兒”是對古代說書人的一種稱謂，意即說書的“男/女/先生”。明清小說中描述服飾、器具等的詞彙在現代漢語中幾乎不再使用，例如，“五彩絲攢花結長穗宮”、“元狐帽沿”等。另外，明清小說中大多數實體的命名與現代漢語也相差甚遠，例如人物稱謂、地名、建築名、機構名、官職、爵位名等，因此，本研究除了對明清小說進行分詞外，還對這些命名實體進行了系統性的標註。

本研究在沿用北大的分詞體系（俞士汶等，2003）、借鑒中央研究院的分詞標準（經濟部中央標準局印行）的基礎上，根據白話語體文的特徵制定了明清章回小說分詞和命名實體標註準則，採用機器輔助、人工分詞和標註相結合、最後進行人工校對的方式對文本進行處理。對於明清小說中的特有詞彙，則參照了相關專用詞典和詞庫進行匹配。目的在於更好地幫助後續建立明清章回小說與現代漢語的語義對映機制，并為包括韻文、語體文等傳統古漢語的分詞研究及後續的語文分析提供參考。

3 分詞和標註的準則

3.1 基本術語

- 分詞準則：用於規定將組成句子的漢字串切分為有效的分詞單位。
- 分詞單位：沿用中國國家標準“分詞規範”中的定義（劉源等，1994），指漢語信息處理中使用的、具有確定的語義和語法功能的基本單位。
- 詞條：指《現代漢語語法信息詞典》及《紅樓夢》、《金瓶梅》專用詞典中所收錄的語言成分，包括語素、詞、詞組、成語、習慣語以及標點符號。
- 語素：具有構詞功能的音、義結合體。單音節語素分為：自由語素、半自由語素、不自由語素（或黏著語素）。其中自由語素能單獨成詞，也能作構詞的成分、與別的語素組合成詞，例如：天－晴天；地－地方；人－大人。半自由語素不能單獨成詞，但能作構詞成分、與別的語素組合成詞，例如：語－語言，耳語；民－民眾，子民；歷－歷練，經歷。不自由語素不能單獨成詞，而且組合位置固定，一定在別的語素之前或之後，例如：初（一），初（五）；（作）家，（行）家，（名）家。
- 詞綴：屬於黏著語素的一種，是附著在詞根或詞幹的語素，不能單獨成詞，而且意義十分空虛，一般分為前綴和後綴。例如：“花兒”中的“兒”、“廚子”中的“子”是後綴，“阿哥”中的“阿”、“小姐”中的“小”是前綴。

---

[1] 本文所使用的《紅樓夢》版本，是蔡義江評注的《增評校注紅樓夢》，作家出版社，2007-3-22 出版。

[2] 《紅樓夢》全文，五次使用“一心一計”（第 6、65、69、79、101 回），一次使用“一心一意”（第 98 回），二者語義完全相同。

- 運行文本：本研究項目的運行文本包括《紅樓夢》、《三國演義》、《水滸傳》、《金瓶梅》。

3.2 基本原則

命名實體在自然語言處理中的應用廣泛，而明清小說中的大多數命名實體與現代漢語中的實體存在較大的差異。例如，明清小說中的人物稱謂隨處可見，且組合形式千變萬化、靈活多樣，多數稱謂與現代漢語差異明顯。地名的命名系統在現代漢語中使用【國，省（區，市），市，縣，鎮，鄉，村】的劃分體系，而在明清時期，在全國實施的是【國，省，府（州），縣，鄉、村（图、鎮、市、都、厢）】制，簡稱行省制。另外，明清時期的行政、商業機構也極富特色，明清小說中引用了大量的機構、團體、組織、官職名稱。因此，本研究建立了對明清章回小說中人名（含本名、字號及別名等）、人物稱謂、地名、建築名、機構（團體、組織）名、官職、爵位名的分詞和標註體系，并對這些命名實體進行人工分詞和標註，以幫助後續的語義分析。

明清小說屬於淺近文言文，銜接文言文和現代白話文，有其獨特的語言風格和語法結構，單字詞的使用頻率比雙字詞及多字詞要高很多。在本研究的運行文本中，單字詞的使用頻率大約分別是雙字詞的 13 倍，三字詞的 32 倍，及多字詞的 65 倍。因此，我們的基本分詞原則是：以語義規範為出發點，如分詞之後不會造成語義丟失、轉換、或歧義，則獨立成詞；另外，從語法角度提供一些分詞準則作為輔助規範。

4 分詞和命名實體標註的具體原則

4.1 命名實體的分詞和標註

在對運行文本的大量實例進行歸納、分析、整合後，本研究對嵌套的多層次複合人物稱謂、地名、建築名、機構（團體、組織）名、官職、爵位名建立了一套統一的分詞和標註系統，既保持了規則的一致性，又能對形式多樣的複合命名實體進行靈活處理。在複合命名實體的總體標註方法上，借鑒了北大詞性標註系統（俞士汶等，2003）對複合命名實體的標註法，引用方括號並加相應標註來表示由多詞組合而成的短語型命名實體，並對之做相應的內部分詞和標註。例如：“[史/nr1 大/姑娘]/na1”。

文學作品中除了引用歷史人物和地名外，還會有塑造的人名和地名，以及神話傳說中虛構的人名和地名，為了後續研究的方便，這三种不同類型的人名、人物稱謂、地名、建築名在標註時，通過使用特殊符號予以區分：

- 運行文本中引用的真實實體：除遵從相應的分詞和標註原則之外，在標註後加“#”進行特別標識，例如：“西子/nr6#”，“雲南/ns2#”。
- 運行文本中引用其他文學作品塑造的人名、人物稱謂、地名、建築名：除遵從相應的分詞和標註原則之外，在標註後加“&”進行特別標識，例如《紅樓夢》中出現“孫行者/nr6&”。
- 神話虛構的人名、人物稱謂、地名、建築名：除遵從相應的分詞和標註原則之外，在標註後加“*”進行特別標識，例如：“警幻仙子/nr6*”，“離恨天/ns4*”。

下文通過一些實例對命名實體的具體分詞和標註原則進行說明。

4.1.1 人名：標註集符號為“/nr”

人名分為六類:【姓】、【名】、【姓+名】、【字】、【姓+字】及【別名】，分別標註為“/nr1”、“/nr2”、“/nr3”、“/nr4”、“/nr5”、“/nr6”。

- 【姓】：分為單姓、雙姓、三字以上姓，此子類包含所有特指某人物的姓，對這一子類加“/nr1”標註。由於此子類僅涵蓋特指某人物的姓，不涵蓋“姓+名”和稱謂，因此，單姓和複姓採用相同的標註原則。例如：“薛/nr1 林/nr1 二/人”，“蘭/nr1# 與/司馬/nr1# 雖/同/名”。對於不是特指某人物的姓，則不加標註，例如：“賣/與/薛/家”。
- 【名】：此子類包括所有特指某人物的本名，對這一子類加“/nr2”標註。例如：“黛玉/nr2”，“文君/nr2#”。
- 【姓+名】：此子類包含所有特指某人物的姓名。如果為單姓，則不需分詞，採用“/nr3”標註整個姓名，例如：“王熙鳳/nr3”。如果為複姓，採用“//”切分姓和名，例如：“司馬//相如/nr3#”。如果包含多個姓（如女子冠夫姓），無論是單

姓還是複姓，採用“//”切分，例如：“[張//王/nr1 氏]/na1”。這樣設計是爲了便於提取應用、并保持標註的一致性，假如出現諸如“複姓//名”、“複姓//單姓//單名”、“單姓//複姓//名”等類複合姓名，均可以採用統一的規則。

- 【字】：此子類包含所有特指某人物的字，對這一子類加“/nr4”標註。例如：“字/存周/nr4”。
- 【姓+字】：此子類包含所有特指某人物的姓+字。如果為單姓，則不需分詞，採用“/nr5”標註整個姓+字，例如：“賈存周/nr5”。如果為複姓、或包含多個姓，和複合【姓+名】採用統一的規則。
- 【別名】：此子類包括所有除本名之外的名稱，含別號、化名、筆名、謚號、廟號、因避諱等原因改用他名等。對這一子類加“/nr6”標註。例如：“顰兒/nr6”，“漢高祖/nr6#”，“警幻仙子/nr6*”。如果別名中包含單姓，則不需分詞，例如：“花袭人/nr6”。如果為複姓、或包含多個姓，和複合【姓+名】採用統一的規則。

4.1.2 人物稱謂：標註集符號為“/na”

明清以及之前的小說中使用了各式各樣的人物稱謂，如不對這類命名實體加標註，則會在實際應用中導致歧義或信息丢失。本研究將明清小說中的人物稱謂分為兩類:【（姓、名、姓+名、字、姓+字、別名、官職、爵位）+稱謂】、【稱謂】，分別標註為“/na1”、“/na2”。

- 【（姓、名、姓+名、字、姓+字、別名、官職、爵位）+稱謂】：此子類包括所有特指某人物，由任何形式的姓、名或官職、爵位等加上稱謂組合而成的複合人物稱謂。對於此類命名實體，使用方括號并加“/na1”標註，內部以最小完整語義為單位進行分詞，并加相應標註。例如：“[寶/nr2 二/爺]/na1”，“[周瑞/nr3 家/的]/na1”，“[林/nr1 御史/nu1]/na1”。
- 【稱謂】：此子類僅包括具有特指意義的稱謂，即特指某人物、并根據運行文本的上下文語境能判斷其所指人物的稱謂，含人與人交往中基於血緣關係、社會地位、職業、官銜、宗教等各種因素的稱呼，不包括自稱。對這一子類加“/na2”標註，內部以最小完整語義為單位進行分詞。例如：“老爺/na2 在/大/書房/等/[二/爺]/na2 呢”，“驚/了/老太太/na2 的/駕/了”。若稱謂為泛指、指多人、或根據運行文本的上下文語境無法判斷其所指人物，則不加標註。例如：“並/公子/小姐/金安”，“雇/的/小/丫頭”。

4.1.3 地名：標註集符號為“/ns”

地名分為四類：國名；省、府（州）、縣、鄉、村、街、巷；山脈、河流、湖、海、岛；其他地名。分別標註為“/ns1”、“/ns2”、“/ns3”、“/ns4”。例如：“暹羅國/ns1#”，“姑蘇/ns2#”，“大荒山/ns3*”，“太虛幻境/ns4*”。

4.1.4 建築名：標註集符號為“/nv”

建築名分為三類：府邸、官邸、院落、亭園、廳房；寺廟、當舖、茶樓、酒館、旅店等公眾場所；其他建築名。分別標註為“/nv1”、“/nv2”、“/nv3”。例如：“大觀園/nv1”，“葫蘆廟/nv2”，“翠煙橋/nv3”。

對於複合建築名，和其他複合命名實體採用統一的規則，即使用方括號并加相應標註，內部以最小完整語義為單位進行分詞、并加相應標註。例如：“[[忠靖/侯]/nu2 史/nr1府]/nv1”。

4.1.5 機構、團體、組織名：標註集符號為“/nt”

對於複合機構、團體、組織名，和其他複合命名實體採用統一的規則。例如：“[欽天監/nt 陰陽司]/nt”。

4.1.6 官職、爵位名：標註集符號為“/nu”

- 官職名：標註為“/nu1”，對於複合官職名，和其他複合命名實體採用統一的規則。例如：“同知/nu1”，“[巡/鹽/御史]/nu1”。
- 爵位、封號：標註為“nu2”，包括具有特指意義的爵位、封號，含由帝王根據血緣親疏、功勞等授予的爵銜、尊號，以及對皇室和宗室女子的封號。對於複合爵位名，和其他複合命名實體採用統一的規則。例如：“[忠靖/侯]/nu2”，“[北靜/王妃]/nu2”。

4.1.7 韻文：標註集符號為“/yw”

韻文的標註分為五個子類：賦、對聯、詩歌、詞曲，及其他（包括酒令、偈語等）。分別標註為“/yw1”、“/yw2”、“/yw3”、“/yw4”、“/yw5”。

明清章回小說中出現較多韻文，鑒於韻文的文體特點，對韻文的分詞需採用不同的準則。因此，本階段僅對韻文進行類別標註，韻文的分詞將依據後續“歷代語言知識庫建置計畫”的韻文分詞規範另行處理。

4.2 以語義規範的分詞原則

以語義來規範的總體原則為：當語義由較小的語素直接組合而成時，採用切分成最小完整語義的分詞單位作為基本原則。例如：“此/事”，“侍/湯/奉/藥/”，“一/段/陳跡/故事”。

以下通過實例對具體原則和特殊情況加以解釋。

- 含“之、了、的、於、眾、只、被、也、亦、所、而、得、時、者”等語素的結構，在明清小說中使用頻率較高，由於與之組合的詞語可替換性強、詞性繁多，予以切分。例如：“之/說”，“攜/了”，“回答/的”。但是，對於連詞和固定用語，則不予切分，例如：“由於”，“只要”，“了不得”。
- 帶指示代詞“這、那、每、各、諸、此、某”的結構予以切分。例如：“這/石”，“每/人”，“諸/事。但是，需要結合運行文本語境判斷，對於非指代或非獨立指代，均不切分。例如：“那怕”，“因此”，“彼此”。
- 和方位詞“前、後、左、右、上、下、裏、中、內、外、畔、旁、邊”所組合的詞條，若表示單純地點方位，予以切分。例如：“門/前”，“檐/下”。但若組合本身具有新的引申意義、和方向無關，或沒有相對應的表示相反語義的反義詞，則不予切分，例如：“心上”、“心下”、“心內”、“心中”都是指“心裏”，而並沒有相應的“心外”，故不切分。
- 含否定成分“不、沒、非、無、勿”的詞條，予以切分。例如：“不/可”，“非/一般”，“無/處”。但是，非獨立否定則不予切分。例如：“不但”，“沒有”，“無故”。可根據兩條原則判斷這類詞條是否切分：第一，基於有否相對應的相反用語，例如：“不/顧”有對應的“你/顧/你/的”，所以予以切分；第二，基於多義或歧義的考慮，根據運行文本的語境判斷，例如：“再/無/一些/遺漏/不當/之/處/了”和“獨/老祖宗/na2 不/當/說/，/不/當/信”。
- 成語、結構式疊詞、習慣用語保留合併。例如：“穿紅著綠”，“隱隱/的/有/座/廟宇”，“不得不”。另外，還需要注意明清小說中區別于現代漢語的固定用語，例如：“一勢兒”，“每日家”。但是，如果疊詞含有中插成分，則予以切分。例如：“享/一/享”，“坐/了/坐”。
- 語義相反或相對、相同或相近的雙字並列結構，因連用程度強，不予切分。例如“離合”“賢愚”；“喜樂”，“抄錄”。
- 含“兒、然、子、著、些、麼”等後綴的詞條，因後綴不具備獨立語義，不予切分。例如“今兒”，“這/會子”，“這些”。

4.3 以語法規範的分詞原則

分詞主要遵循語義相關原則，以下從語法角度描述的分詞原則作為輔助規範。例如，“差人”可用作名詞，也可用作動賓結構，用作名詞時不予切分，用作動賓結構時則予以切分。

- 動賓結構的組合，由於其語義大多可以來自動詞及隨後的名詞結構，以分開為主。例如：“擷/花”，“留/飯”，“理/朝廷”。對於固定表達和習慣用語，則可保留合併。例如：“修方”，“沏茶”。
- “形容詞+名詞”的雙字偏正結構，由於切分後多數語義不會發生變化，因此以分開為主。例如：“奇/物”，“細/簪子”。
- 名詞、動詞的主謂、述補結構，由於其語義相對獨立，以分開為主。例如：“腰/圓”，“耳/赤”；“吹/倒”，“吃/盡”。
- 數詞、量詞、名詞組成的結構，由於內部成分的意義相對獨立，也予以切分。包括“第+數詞+量詞”結構、“數詞+量詞（+名詞）”結構、“數詞+名詞”結構等。例

如：“第/二/日”，“幾十/兩”，“幾百/株/（花）”，“三/尺/來”，“初/七/日”，“二/更”，“三/人”。

- “副詞+動詞”的雙字偏正結構，如果整體語義較為固定，則不予切分。例如：“厚愛”，“明示”。
- 趨向動詞（上，下，過，回，進，出，起，歸，到，出，走）+（來，去，入）的述補雙字結構，由於相組合而成的詞條用法較固定，不予切分。例如：“上來”，“下去”，“進入”。
- 意義已經固化的詞條、成語、歇後語、或習慣語，不予切分。例如“一疊聲”，“接二連三”，“牽五挂四”，“巴不得”，“貴人多忘事”。

5 質量保證與效果

運行文本的分詞和標註主要分為三個階段，採用自動分詞系統訓練、自動分詞處理、之後進行嚴格的人工標註來完成。首先，爲了使現有的現代漢語自動分詞系統（Q. Lu et al.，2004）能夠適應明清小説的文體和命名實體，本項目先利用一些現有的古典文學詞彙以及小批量人工分詞和標註的命名實體得到的詞彙來更新自動分詞系統的詞彙知識和標註的規則，再對文本進行預處理：結合相關詞典對文本進行分詞，使用系統對文本中的命名實體進行統一標註。這一階段完成後，可導出經過系統輔助分詞和標註的文本，據實驗統計，此時的準確率達 90%。第二階段，嚴格按照規範對系統預處理後的文本進行人工標註和校驗。由於一些詞條的處理必須根據語義和上下文語境進行判斷，系統無法實現，這一階段對於保證質量至關重要。此階段至少對全文檢查兩次，重點在於消除因一詞多義而產生的歧義、保證一致性、解決命名實體標註遺漏、錯誤等問題。這一階段由一人完成，因為多人操作可能會導致人為的不一致。校驗完成後，由另一人員進行檢查，據檢查結果統計，準確率為 99. 44%，遺留的問題均為分詞問題，主要包括兩類：第一類問題約 0. 31%，是由於不同的人員對詞條的語義或分詞規範有不同的理解而導致的，修改這些詞條、進一步明確規範後能使語義更加清晰；第二類問題約 0. 25%，主要是校驗時對詞典中的某些詞條疏忽導致的，因為分詞規範的要求更加細化，需要對詞典中的詞條進行修改。該三階段並非一次性行爲，在人工標註時，如發現自動分詞和標註的問題，會補充相應的詞彙知識和標註規則，以不斷提高自動分詞標註的性能。我們的的目標是確保最終完成的語料庫準確率超過 99. 5%。

6 結語

本研究在盡可能借鑒北京大學和臺灣中央研究院分詞和標註標準的基礎上，以語義分析為主要出發點，結合明清章回小說中人物稱謂變化繁多、韻文與語體文以及實體名與小說塑造名、神話傳說虛構名交叉使用、用詞簡潔、單字詞頻率高等等特徵，建立明清小說分詞和命名實體標註準則，對《紅樓夢》、《三國演義》、《水滸傳》、《金瓶梅》採用機器輔助結合人工分詞和標註、并進行人工校對的方式進行處理。在分詞方面，致力於在做到切分後不造成語義丟失、轉換、引申或歧義的情況下，以切分到最小語義單位為基本理念；在命名實體標註方面，力求對嵌套的多層次複合人物稱謂、地名、建築名、機構（團體、組織）名、官職、爵位名建立統一的標註準則，其目的是既保證規則的一致性，又能滿足複合命名實體多種組合的需要。本研究對詞條的細分及對各類命名實體的標註可作為後續語義分析的基礎、以更好地建立明清章回小說與現代漢語的語義對映機制，亦希望為韻文、語體文等傳統古漢語文體的切分、標註研究提供參考。

参 考 文 献

[1] Q. Lu, S. T. Chan, R. F. Xu, T. S. Chiu, B. L. Li, and S. W. Yu. A Unicode based Adaptive Segmentor . Journal of Chinese Language and Computing, 2004, 14 (3): 221-234.

[2] 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註注音. Journal of Chinese Language and Computing, 2003, 13(2): 121-158.

[3] 夏迎炬、于浩、西野文人，人民日報語料庫命名實體分類研究. Computational Linguistics and Chinese Language Processing, Vol 10, No.4, December 2005: 533-542.

[4] 中文資訊處理分詞規範，經濟部中央標準局印行.

[5] 中國國家標準 GB13715 “信息處理用現代漢語分詞規範”。見：劉源等，信息處理用現代漢語分詞規範及自動分詞方法。北京：清華大學出版社，1994 年第 1 版.