文章编号：1003-0077(2015)01-0019-09

基于语料库的明清小说人名与称谓研究

熊 丹[1]，陆 勤[1]，罗凤珠[2]，石定栩[3]，赵天成[1]

（1. 香港理工大学 电子计算学系，香港；
2. 台湾元智大学 中国语文学系，台湾；
3. 香港理工大学 中文及双语学系，香港）

摘 要：在自然语言处理及其应用领域，人名和称谓作为重要的命名实体，是信息处理的关键部分之一。该文从命名实体识别和资讯提取的角度出发，在对 4 部明清古典小说的语料库进行标注的前提下，建构了姓名、字号和称谓作为命名实体的分类及标注系统。人名和称谓总体上分为单一型和复合型，根据复合型的内部组成元素和组合方式，将其进一步分为固定式、同位式、附属嵌套式、灵活嵌套式。结合语料库的完整数据统计，该文对各类型人名和称谓进行了比较分析，并分别展示了 4 部名著在人名、称谓使用上的特点。

关键词：命名实体标注；人名和称谓分类；语料库构建

中图分类号：TP391 **文献标识码**：A

A Corpus-Based Study on Personal Names and Terms of Address in Chinese Classical Novels

XIONG Dan[1], LU Qin[1], LUO Fengzhu[2], SHI Dingxu[3], ZHAO Tiancheng[1]

(1. Department of Computing, The Hong Kong Polytechnic University, Hong Kong, China;
2. Department of Chinese Linguistics & Literature, Yuan Ze University, Taiwan, China;
3. Department of Chinese & Bilingual Studies, Hong Kong Polytechnic University, Hong Kong, China)

Abstract: Personal names and terms of address are important parts of named entities. The recognition of personal names as well as terms of address is ans essential issue in natural language processing. This paper presents a classification and annotation scheme for personal names and terms of address from the perspective of named entity recognition and information extraction on a corpus of four Chinese classical novels. Personal names and terms of address are categorized into simple types and compound types. And the compound-type is further categorized into four subtypes, fixed expressions, appositive constructions, subordinate constructions of affiliation, and other subordinate constructions. This paper also presents a comparative analysis on these types and the characteristics of the four novels based on full statistics of the annotated corpus.

Key words: named entity annotation, classification of personal names and terms of address, corpus construction

1 引言

中国的姓名、称谓文化博大精深、源远流长，古典文学作品往往借助书中角色的姓名、字号及形式多样的称谓来叙述故事、刻画人物形象、显示角色之间的关系，使角色鲜明，语言生动。因此，在使用计算机处理古典文学作品时，对人名和称谓的系统性标注成为有效理解和处理语言的关键因素之一。不同于其他文学和历史层面对称谓的研究，本文从命名实体识别和资讯处理的角度出发，从称谓的内部元素和组合方式入手，对明清小说中的人名和称谓建立了一套分类和标注系统，并将这一系统实际应用到明清古典文学的命名实体标注中。本研究建构的语料库共计 262.35 万中文字（不含标点符号），包括《三国演义》《水浒传》《金瓶梅》和《红楼梦》，在分词的基础上所标注的人名、称谓总数达 17 万条。工作模式主要为人工标注，同时也对现代汉语分词标

收稿日期：2013-04-08 定稿日期：2014-12-09

基金项目：2009 年度蒋经国国际学术交流基金会“历代语言知识库建置计划”(RG013-D-09)

注系统[1]加以训练，进行预处理和后期处理，以提高标注的效率和质量。

本文主要内容如下：第2节简单介绍自然语言处理领域内较通用的语料库加工规范对人名、称谓的处理方法，从而引出本文的研究目的；第3节阐述本研究建立人名、称谓分类及标注系统的理念；第4节通过实例详细说明人名、称谓的分类和标注；第5节展示4本明清小说的完整数据统计，归纳人名、称谓的组成元素和组合方式，探索复合型称谓内部成分的组合规则，并通过对4本小说的数据对比分析其各自的特色。

2 研究概况及目的

称谓是人与人交际过程中必不可少的语言单位，在语言交流中反映了人际关系的复杂程度，而且在不同的历史时期承载着不同的历史人文信息。因此，称谓的标注和识别如果能够通过自然语言处理技术来完成，对于中文信息处理和计算机辅助的文学和历史研究意义重大。然而，目前的中文信息处理技术主要是针对现代汉语，称谓并没有得到重视，也没有对称谓及由称谓和人名组合的名称进行细分。国内较通用的现代汉语语料库加工规范[2]中，将人名作为一类单独的命名实体（总标识符为“/nr”），汉族姓、名分开标注，例如，“张/nrf 仁伟/nrg”、“欧阳/nrf 修/nrg”；对于双姓（含女子冠夫姓）也在切分之后进行标注，例如，“唐/nrf 姜/nrf 氏/nrg”；别名、译名、简称等均标注为“/nr”，如“鲁迅/nr”、“爱因斯坦/nr”，“陈总/nr”；但对于姓名后附加职务或称谓的名称，将职务、称谓只作为普通名词（标识符为“/n”），例如，“李/nrf 主席/n”、“刘/nrf 阿姨/n”、“陈/nrf 老总/n”。“台湾中央研究院”建立的近代汉语（唐以后）标记语料库[3-4]也包括词类标注信息。该语料库包括了《红楼梦》在内的明清文学语料，姓名划归为专有名词（标识符为“Nb”），而称谓也只划归为普通名词（标识符为“Na”），例如，“林(Nb)姑娘(Na)”、“凤(Nb)姐姐(Na)”、“杨(Nb)提督(Na)”。

在我们对4本名著进行标注的过程中发现，称谓不能简单地处理为普通名词，因为不论是单独使用还是和姓名连用，称谓都发挥命名实体的功能。而且，小说中对同一个人物的称谓会随着该人物身份、地位、所处的场合、交流的对象、甚至当时作者想要体现的感情色彩而不断变化。称谓可单独使用，也可通过不同形式灵活组合而成，例如，可以从姓名、字号中截取一部分再加上头衔组合而成。这种复合型称谓各元素之间的组合关系相当复杂。如果将其作为命名实体来标注，就需要对这些称谓进行系统性的分类、并对其组合关系进行分析，既要确保古典文学标注语料库的建设过程中，采用统一的原则进行分词和标注，又要兼顾灵活性而有利于文学和历史的后续研究，例如，建立文本内及不同文本间相关命名实体的关联和基于命名实体为人物属性建立档案等。因此，本文将称谓作为一类主要的命名实体、从资讯处理的角度进行分析，不同于一般的文学和历史研究中对称谓的分类，但同时也考虑如何在资讯平台上为文学和历史研究提供方便。本文主要从称谓的内部元素和组合方式入手，将人名和称谓进行整合分类，探索复合型称谓的组合规律，并将总结的规则实际应用到4本名著的命名实体标注中。

3 人名、称谓分类及标注系统的设计理念

3.1 称谓的界定

长期以来，关于称谓的概念、范畴，一直存在多种看法，没有定论（如郑尔宁等介绍的现代汉语称谓研究的几种主流观点）[5]。关于面称（直接当面称呼）与引称（间接指称性称谓）之间的关系，也出现了很多探讨，其中不乏具有代表性的研究[6]。本文对称谓的定义不予深入探究，而是采用一个广义的概念，既包含人与人之间言语交际中所使用的直接称呼，也包含提及他人时使用的指称性名称。从这一意义来看，本文将用于指称、显示人物身份和角色定位的官职、爵衔都纳入称谓之列。从词类的角度来看，本文研究的称谓仅包括名词和名词性短语，不包括代词。另外，鉴于本文的研究是从资讯处理的角度出发标注命名实体，因此仅将特指某一人物、并根据上下文语境能判断其所指人物的称谓作为命名实体，例如，《红楼梦》中的“姑娘”，如果能够判断其所指的对象，则加称谓标注，而“一/个/姑娘/领着/他”、“姑娘/们”等非特指之称谓，则当普通名词处理，不加称谓标注。

3.2 人名、称谓的分类方式

自古以来已有很多对称谓的研究，但针对不同的研究目的，对称谓的分类方式也各有侧重。

Braun[7]在对不同语言中的称谓进行比较研究时，从词类的角度将称谓大体分为代词称谓、动词形式的称谓和名词形式的称谓。综观古今，一些较有影响力的汉语称谓专著和词典[8-14]，其中有些工具书对古今中外的称谓兼收并蓄，分门别类地收录了几千甚至 3 万余条称谓，对于称谓的分类，其角度和细微性均有所不同。例如，基于指称对象的身份一般分为家族亲属、社交、职业职官、民族宗教等；基于称谓的使用形式分为习称、别称、统称、通称、俗称等；基于情感色彩、雅俗褒贬分为尊称、贬称、昵称、谀称、雅称、贱称等；基于称谓的使用年代分为古称、今称。鉴于本研究的结果需要应用于古典汉语信息处理和语料库建构，本文将人物的姓名、字号和各类称谓糅合汇总、再从其内部构成及组合方式逐层逐级进行分类，建立人名和称谓的分类及标注系统。

3.3　人名、称谓的分词及标注的基本原则

语料库的分词系统，遵从的是本项目根据白话语体文的特征制定的明清章回小说的分词准则，基本原则是“致力于在做到切分后不造成语义丢失、转换、引申或歧义的情况下，切分到最小完整语义单位”[15]。该切分系统基本沿用了北京大学的现代汉语分词体系[2]，并借鉴了“台湾中央研究院”的分词标准[16]。人名、称谓的标注主要沿用北京大学词性标注系统[2]，对于该系统中没有的类型则新增标识符。鉴于语料库的建构要求全文分词的一致性，因此对称谓的分词采用语料库整体分词原则，例如，“[三/姐姐]/na2”、“[国舅/老爷]/na2”、“[冠军/将军]/nu1”、“[忠武/侯]/nu2”。需要注意的是，古典小说中包括一些现代汉语中已经不再使用的古称，例如，“足下”、“衙内”、“房下”，用作称谓时不能切分。

本研究将“姓”、“名”、“姓＋名”、“字”、“姓＋字”作为不同类别分别标注，例如，“刘/nr1＃”、“备/nr2＃”、“刘备/nr3＃”、“玄德/nr4＃”、“刘玄德/nr5＃”，因此无需对“姓＋名”和“姓＋字”类的人名再进行分词。对复姓和多姓，使用“//”予以区分，如“诸葛//亮/nr3＃”、“[张//王/nr1 氏]/na1”。但对于由不同的分词单位[17]组合而成的称谓，则需要进行分词。如果称谓中包含其他类型的命名实体，如地名、机构名，则以嵌套方式保留其独立标识符。

另外，虽然文学作品中多数人物为作者所虚构，但也会引用历史人物和其他文学作品中塑造的人物，为了便于本研究后续历代语言知识库的贯穿，分别使用“＃”、“＊”和“&”表示历史真实人物（以《二十四史》为依据）、神话传说虚构人物和引用其他文学作品的人物，例如，《红楼梦》中出现的“陶渊明/nr3＃”、“如来佛/nr6＊”、“李逵/nr3&”等。因此，本文所提到的人名、称谓包括小说塑造的人物及小说中引用其他文献的人物。

下节通过实例详细阐述人名、称谓的分类及标注，本文所有实例均取自于已标注的 4 本名著。

4　人名、称谓的分类及标注

4.1　总体分类系统

由于社会结构、文化背景的差异，在不同的时代、地域、以及社会群体中，称谓具有明显的特征。而小说为了凸显其艺术效果，使用的称谓更是变化多样。例如，《金瓶梅》中的蔡京，虽然不是小说的主要人物，却使用了多种指称方式。既有直接用单姓“蔡”和姓名“蔡京”进行指称的，也有用官职指代的，例如，“左丞相”、“大学士”、“吏部尚书”、“太师”等。下属、仆役称呼他时会用“老爷”、“蔡老先生”、“蔡太师”、“太师爷”、“老太师”、“太师老爷”、“蔡太师老爷”等，而内相们私下谈论时则贬称为“老贼”。另外，本文采用的语料文本虽然是明清时期创作的小说，但其故事所处的时代背景、社会环境都不尽相同，不同程度地折射出秦汉、唐宋、明清等多个时期的文化形态和社会风貌，而且故事人物的社会角色千差万别，因此语料中出现的称谓非常丰富。

基于对 4 部名著人名和称谓的综合分析，本文从其组成元素和组合方式的角度进行了综合分类，总体上分为单一型和复合型两大类，顾名思义，前者由人名、称谓本身独立承担指称功能，后者由多个成分叠加或嵌套组合而成。复合型称谓的内部组合方式非常灵活，有的是由多个独立使用的单一型人名、称谓叠加而成，例如，“[令郎/先生]/na2”，其中“令郎”和“先生”都可以用来作为独立的称谓；有的是截取人名的一部分、再和称谓合并而成，例如，“[凤/nr2 姐姐]/na1”；还有的是由人名附加修饰、描述语组合而成，例如，“[周瑞/nr3 家/的]/na1”，其中“家/的”不能独立作为称谓，一般附加于人名后组合成复合称谓。经过对语料中的人名、称谓进行归纳分析，本文从其内部构成及其组合方式入手，分为以下类别（图 1）。

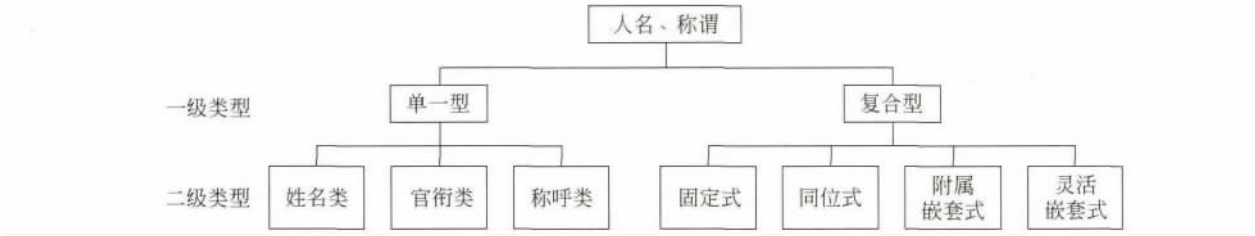

图1 人名、称谓的总体分类

4.2 单一型细分

如图1所示，单一型人名、称谓分为3大类，表1中对其进行了细分，并用实例说明。为了便于理解，取自语料中的实例均保留语料库中的标识符号和标注形式，表1还列出北京大学现代汉语语料库[2]采用的相应标识符，以便参照。

表1 单一型人名、称谓细分

二级类型	三级类型	本语料标识符	北大标识符	定义及说明	实例
姓名类	姓	/nr1	/nrf	包括单姓、复姓。	关/nr1＃ 诸葛/nr1＃
	名	/nr2	/nrg	包括单音节、双音节。	备/nr2＃ 黛玉/nr2
	姓＋名	/nr3	/nrf /nrg	人物的姓＋名，复姓用“//”区分。	林黛玉/nr3 司马//相如/nr3＃
	字	/nr4	/nrg	通常为双音节，有少量单音节。	士隐/nr4 平/nr4(单音节“字”较少见，此为《三国演义》中的太医吉平：姓/吉/nr1 ，/名/太/nr2 ，/字/称/平/nr4)
	姓＋字	/nr5	/nrf /nrg	人物的姓＋字，复姓用“//”区分。	甄士隐/nr5 诸葛//孔明/nr5＃
	别名	/nr6	/nr	所有除本名之外的名称，包括别号、谥号、帝王庙号、昵称、绰号、不能分为姓和名的外族名等。复姓用“//”区分。	卧龙先生/nr6＃(别号) 昭烈皇帝/nr6＃(谥号) 宋徽宗/nr6＃(庙号) 颦儿/nr6(由黛玉表字“颦颦”而来的昵称) 耶律//雄奴/nr6(绰号) 彻里吉/nr6(姓和名无法区分)
官衔类	官职	/nu1	/n(普通名词)	具有特指意义(与某个或某些人物有对应关系①)的官职名，例如，“司徒/nu1 王允/nr3＃”；泛指的官职名不加标注，例如，“出/了/一/个/郎中/缺”。	大司马/nu1 太守/nu1
	爵位封号	/nu2		包括帝王根据血缘亲疏、功劳等授予的爵衔、尊号，含对皇室、宗室女子的封号。	郡王/nu2 贵妃/nu2
称呼类		/na2	/n	人与人交往中基于血缘关系、社会地位、身份、宗教等各种因素对某一特定人物的称呼，既包括当面交流时直接称呼对方所使用的名称，也包括提及他人时的间接指称，不含自称。	老祖宗/na2(常用于直称) 祖母/na2(常用于引称) 老太太/na2(直称、引称均可)

① 指称关系是受时空限制的，在特定的时间地点，某个官衔和某个人物有一一对应的关系，但时过境迁，担任这个官职的人物变了，指称关系会相应变化。

4.3　复合型细分

单一型称谓可独立用于指称，也可作为复合型称谓中的单元成分。本工作对各类复合型称谓采用统一的标注系统：使用“[ ]”总括，如内部成分的类别与复合称谓的类型相异，则保留其独立标识符。基于内部成分的组合关系，复合型称谓可分为4大类，本节通过实例进行描述。

4.3.1　固定式组合

这一类型是由多个成分组合而成的较固定的名称，其内部成分一般不分开使用，或分开后仅作为简称使用，例如，

• 以美号赐封的爵位和封号：帝王封爵时赐予的美号和爵衔组合而成的名称，具有特指性、较固定性，例如，“[北静/郡王]/nu2”，“[顺平/侯]/nu2”。因为“北静”和“郡王”作为单一成分均为爵位，与其复合称谓一致，而无需再加独立标识符。

• “名号＋将军”组合而成的武将官职：对有军功者授予“将军”官衔时会冠以名号，例如，“[奋威/将军]/nu1”，“[冠军/将军]/nu1”。

4.3.2　同位式组合

这一类型由多个存在同位关系的成分堆叠而成，其内部成分一般为同一类型，可分开后独立使用，例如，“[父亲/大人]/na2”，“[都太尉/统制]/nu1”。

4.3.3　附属嵌套式组合

这一类型由两个存在附属、主次、支配或依存关系的成分组合而成，其内部成分可能为不同类型，但具备依存关系。主要包括：

• 主次关系：较常见的主次关系如“[[北静/王]/nu2 妃]/nu2”、“[丞相/令史]/nu1”。

• 管辖地＋官职：人物的官职经常和其管辖地连用，为了不使这一信息丢失，将其作为一个复合型命名实体，例如，“[扬州/ns2＃ 刺史]/nu1”。

• 封地＋爵位封号：如果封爵时赐予了封地，爵位、封号名用作称谓时通常会附带封地名，例如，“[乌程/ns2＃ 侯]/nu2”。“乌程”为地名，因此保留其地名标识符(/ns)，而“侯”则无需重复爵位标识符(/nu2)，系统可默认识别。

• 机构＋官职：小说中提到官职时，往往还会采用“机构＋官职”这一组合形式，为了保持两者之间的关联，便于后续的信息提取，将这两个命名实体作为一个复合型命名实体，例如，“[吏部/nt 尚书]/nu1”。

4.3.4　灵活嵌套式组合

这一类型包括所有其他由两个或两个以上的成分灵活嵌套组合而成的复合型称谓，其内部成分可以是单一型人名、称谓，也可以是以上几种复合型称谓。无论其内部成分多么复杂，都可逐层剖析成单一型人名、称谓后使用统一的标注规则进行处理。从其内部组合方式划分，灵活嵌套式组合可进一步分为8类，在表2通过实例说明。对灵活嵌套式组合的复合型称谓使用“[ ]”总括，并加“/na1”作为总标识符。如这一组合的内部成分为单一型“称呼类”实体，无需再加单一型“称呼类”实体标识符“/na2”，系统可默认识别，例如，“[西门/nr1 老爹]/na1＃”

表2　灵活嵌套式组合细分

三级类型	实　例	说　明
人名＋称呼	[西门//庆/nr3 大人]/na1，[玄德/nr4＃公]/na1＃，[九天玄女/nr6＊ 娘娘]/na1＊	由各种形式的姓名、字号(包括"姓名类"所有子类)附加称呼组成。
官职＋称呼	[太尉/nu1 恩相]/na1	由官职附加称呼组成。
爵位封号＋称呼	[[临安/伯]/nu2 老太太]/na1	由爵位、封号附加称呼组成。
人名＋官职	[高/nr1 太尉/nu1]/na1＃，[兀颜//光/nr3 上将军/nu1]/na1	由各种形式的姓名、字号附加官职组成。
人名＋官职＋称呼	[[赵/nr1 枢密/nu1]/na1 相公]/na1	此例中，其内部成分本身已是复合型称谓。
人名＋爵位封号	[史/nr1 侯/nu2]/na1，[琼英/nr2 郡主/nu2]/na1	由各种形式的姓名、字号附加爵位、封号组成。
人名＋爵位封号＋称呼	[[元/nr2 妃/nu2]/na1 姐姐]/na1	此例是由姓名的一部分加封号、再附加称呼组合而成的复合型称谓。
机构名＋称呼	[吏部/nt 公]/na1	由人物任职的机构和称呼组合而成。

说明：

1. 以上各种组合的内部成分先后顺序不定，例如，“官职＋称呼”组合，其内部成分的顺序也可能是“称呼＋官职”，如“[义士/提辖/nu1]/na1”。

2. 以上组合中，任何一种内部成分的数量不定，例如，“人名＋称呼”组合中，可能出现多个称呼，如“[[西门/nr1 先生]/na1 大人]/na1”。

中的“老爹”是一个单一型称呼，无需再加独立标识符。如内部成分为其他类型实体，则需保留其独立标识符，例如，“[西门/nr1 提刑/nu1]/na1”。

5 数据分析

本节将基于这4部小说的特性，通过语料库数据对人名、称谓的使用情况进行详细分析。

5.1 综合数据分析

表3显示了人名、称谓作为命名实体对比语料库总规模的数据。4部小说中，《红楼梦》的总词汇量最大，《三国演义》最少，数量相差31%。4部小说中包括人名、称谓的命名实体占总词汇量的9.21%，由此可见，人名、称谓在语料库建构中的作用不可忽视。在这4部小说中，《三国演义》的词条数最少，其人名、称谓的比率最高，达到12.20%，主要原因是作为历史小说，文中出现的姓名和官衔都最多。

表3 人名、称谓在语料库中的比率

	三国演义	水浒传	金瓶梅	红楼梦	四部语料
语料库总词条数/万条	37.00	48.64	45.42	53.46	184.52
人名、称谓的总数/万条	4.51	4.61	4.04	3.84	17.00
人名、称谓的比率/%	12.20	9.48	8.89	7.18	9.21

5.2 各类型人名、称谓的分布

本文根据4部小说中人名、称谓的构成方式，将其分为单一型和复合型两大类，其中单一型又分为姓名类、官衔类、称呼类；复合型进一步分为固定式、同位式、附属嵌套式、灵活嵌套式。图2展示了这些类型分别在四部小说中的频率归一化分布，即对每部小说中的各类人名、称谓出现的总次数进行统计，例如，“林黛玉/nr3”在《红楼梦》中出现了279次，则计为279。以各部小说中人名、称谓的总次数作为分母，计算出这些类型在该小说中所占的百分比。

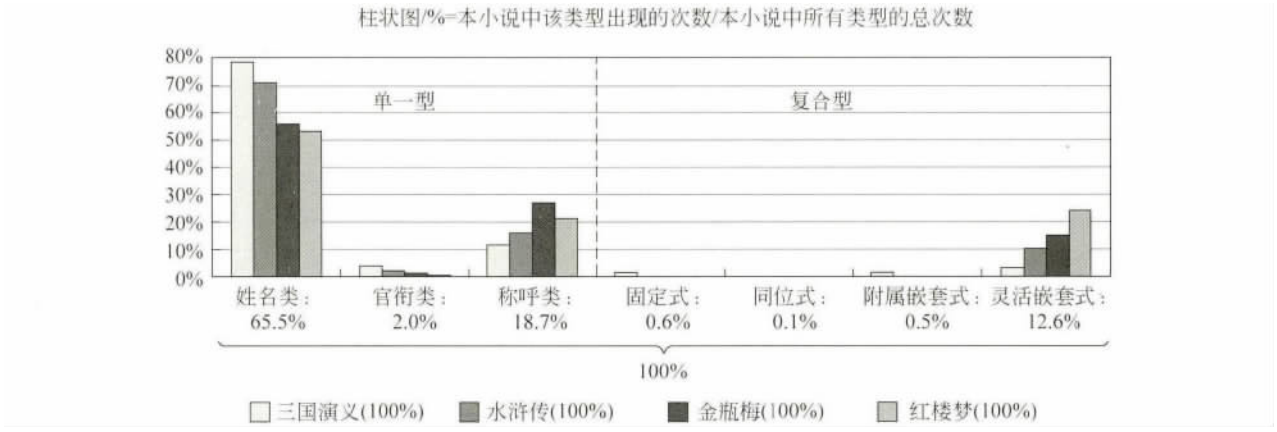

图2 各类型人名、称谓的分布

从总体上看，4部小说中出现最多的均为“姓名类”，其中一个主要原因是这4部小说采用第3人称叙述故事，使用姓名的语境很多。另外，古代人名的形式多种多样，除了组成现代人名的姓和名之外，还存在字、别号等其他形式的名称，可选择性强，因此使用姓名的频率很高。在复合型称谓中，4部小说都是“灵活嵌套式”的比例最高，因为这种组合能帮助塑造人物形象，增强小说语言的吸引力。

从4部小说各自不同的特性上看，《三国演义》和《水浒传》是历史小说，因此“官衔类”的比例比《金瓶梅》和《红楼梦》高。而同为历史小说，《三国演义》比《水浒传》的官方人物多，因此“官衔类”的比例更高。《金瓶梅》作为平民文学，“官衔类”的比例比作为贵族文学的《红楼梦》更高，是因为围绕西门庆出现了较多官场人物。另外，《三国演义》中的“固定式”、“附属嵌套式”比例均高于其他3部小说，这是

因为其“固定式”包括很多由“名号＋将军”组合而成的武将官职，“附属嵌套式”中包括较多由“管辖地＋官职”组合而成的复合称谓。《金瓶梅》中的“称呼类”（即单一型称呼）比例最高，因为小说中围绕西门庆构成了繁杂的人际关系网，描绘了当时的市井风情，因此使用非正式称呼的语境较多。

《水浒传》中的“同位式”高于其他 3 部小说，因为文中的江湖豪客之间常常使用这一方式的称谓以示尊敬，例如，“[庄主/太公]/na2”、“[先锋/哥哥]/na2＃”。

为了审视各种类别的使用分布，表 4 展示了各类型人名、称谓的实例使用率，即用这些实例出现的总次数除以其个数，这一数据体现单位类型上的实例使用率。由此数据可见，《红楼梦》的“姓名类”使用率最高，因其在小说的第 3 人称叙述中使用较多，其中“宝玉/nr2”在全文中出现了超过 3 900 次。《三国演义》中“同位式”的使用率较低，全文仅出现了两例（“[宗兄/将军]/na2＃”和“[大司马/将军]/nu1”），各使用了一次，因为其作为历史演义小说，更多地使用了“固定式”（如“名号＋将军”）、“附属嵌套式”（如“管辖地＋官职”）。

表 4　各类型的实例使用率

二级类型	三国演义	水浒传	金瓶梅	红楼梦
姓名类	16	24	23	26
官衔类	12	12	7	6
称呼类	20	17	23	20
固定式	3	3	1	3
同位式	1	3	2	2
附属嵌套式	3	1	1	2
灵活嵌套式	6	9	8	16

5.3　姓名类数据分析

鉴于“姓名类”在小说中出现的频率最高，因此本小节对其数据进行进一步分析。图 3 展示了“姓名类”的 6 个子类在 4 部小说中的频率分布，即对每部小说中“姓名类”各子类出现的次数进行统计，计算各子类在该部小说的“姓名类”中所占的百分比。

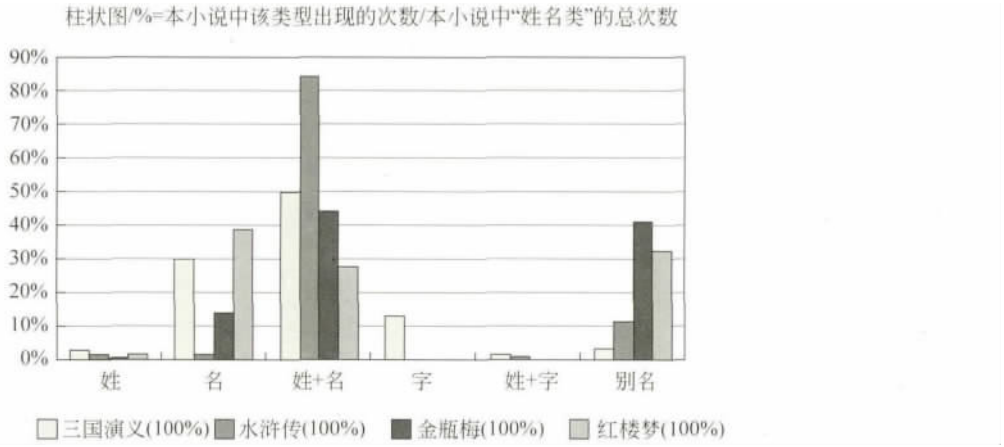

图 3　“姓名类”的分布

从总体数据来看，“姓＋名”的比例最大，这和现代日常交流中称谓的使用规律一致。古典小说中使用别名的频率很高，因为古人对他人的称呼，以称“别号”为最尊；对亡者的称呼，以称“封号、谥号”为最尊。

从 4 部小说各自的使用情况看，《红楼梦》中“名”的比例最高，文中采用第 3 人称叙述故事时，对主要角色“宝玉/nr2”、“黛玉/nr2”、“宝钗/nr2”的指代常常使用“名”。《水浒传》中讲述故事时则多数使用“姓＋名”，其中“宋江/nr3＃”出现的次数达 3 800 多次。《金瓶梅》描写的是世情生活，因此别名出现的比例较高。

图 4 展示了 4 部小说中不同音节的“姓”、“名”、“字”的分布情况。对各部小说中“姓”、“名”、“字”出现的次数分别进行统计，计算这些子类在该部小说的“姓”、“名”、“字”总数中所占的百分比。《三国演义》中的单音节“名”所占比重很大，这在一定程度上反映了当时的姓名文化。与之形成对比的是，《人民日报》现代汉语语料库中“单姓双名”的情况远远多于“单姓单名”[18]。另外，“字”是中国古代姓名文化中的重要元素，通常为双音节，仅《三国演义》中出现了少量单音节“字”，例如，太医吉平，“字/称/平/nr4”。

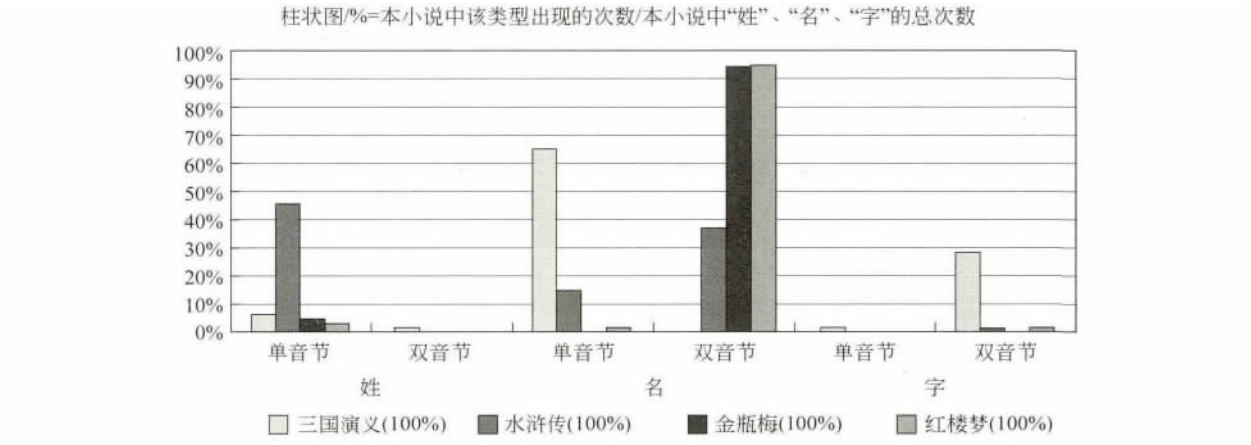

图4 不同音节"姓"、"名"、"字"的分布

5.4 灵活嵌套式组合数据分析

"灵活嵌套式"是复合型称谓中比重最大的一类,其组合灵活多变、内部成分相对复杂,在语料中长度也很突出,例如,"[[元/nr2 妃/nu2]/na1 姐姐]/na1",由"名"的一部分加封号组合而成的复合型称谓作为其内部成分,再附加单一型称谓"姐姐"组成多层次的复合型称谓。图5显示了第4.3.4节所描述的8类"灵活嵌套式"称谓在整个语料库中出现的总频率的比例分布,由此可见,使用最多的组合是"人名+称呼",这在一定程度上也是因历史上人名形式的多样性所致。

图5的子饼图是对"人名+称呼"这一子类所做的进一步分析。结果显示,"姓+称呼"的比例最大,因为在对话中使用较多,这也是小说语言的特征之一。其中,"姓+称呼"中包括多姓的情况,例如,女子冠夫姓,例如,"[张//王/nr1 氏]/na1"、"[西门//吴/nr1 氏]/na1"。"姓名+称呼"中,也可能不使用全名,而是姓加上名的一部分,例如,"[王/nr1 凤/nr2 姐]/na1"。

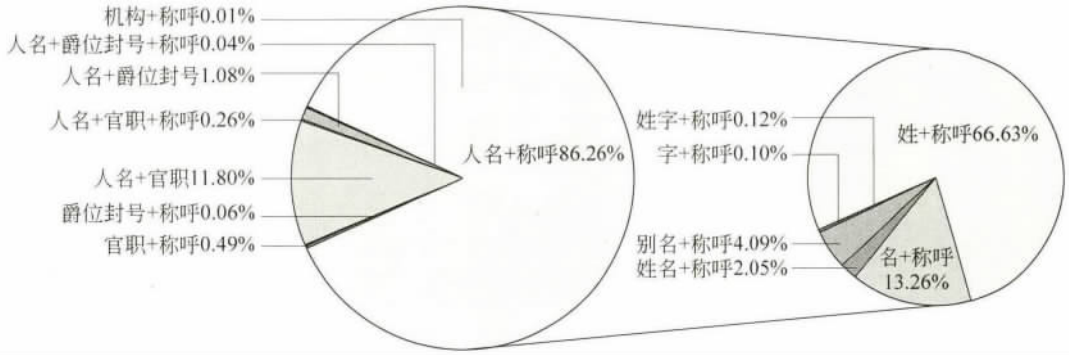

图5 "灵活嵌套式"在语料库中的总体分布

图6展示了"灵活嵌套式"各子类在4部小说中的分布。由此数据可见,4部小说中都是"人名+称呼"的比例最高,其中《红楼梦》中"人名+称呼"的比例高于其他3部小说,主要因为"[贾/nr1 母]/na1"、"[凤/nr2 姐]/na1"这类形式的称谓在第3人称叙述中出现的次数很多。《水浒传》中"人名+官职"的比例最高,因为其故事中涉及较多官场人物,对官员的称谓常常使用这种形式,例如,"[高/nr1 太尉/nu1]/na1"出现了超过200次。"机构+称呼"的组合较为少见,仅《金瓶梅》中出现了3次"[吏部/nt 公]/na1",这也反映了《金瓶梅》语言的生动、不拘形式。

6 结语

在以往的汉语分词和标注中,称谓通常被作为普通名词处理。但称谓无论是单独使用,还是和姓名等组合使用,都发挥着命名实体的功能。本文基于古典文学语料库对人名、称谓作为命名实体进行全面、综合性的分析,填补了以往命名实体在汉语分

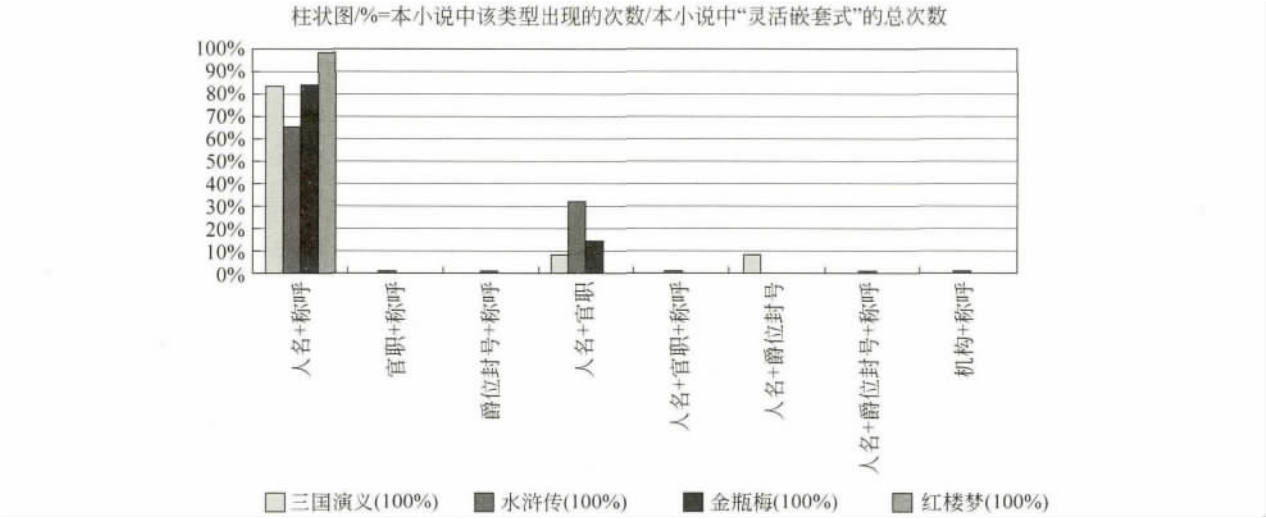

图 6　"灵活嵌套式"在四部小说中的分布

词和标注中对称谓的缺项。鉴于明清小说语料中人名、称谓的复杂性，本文从命名实体识别和资讯提取的角度对其进行逐层分类，以帮助识别、处理和提取这一类型文学作品中的人名和称谓。在后续的研究中，可以和更早期时代的语料或现代语料进行比较分析，明确不同时代的差别。另一项颇有意义的工作是在本文分类研究的基础上建立文本内及跨文献、贯穿历代语言知识库的指称对象的关联，进一步为文学和历史的研究提供基础关联信息。

参考文献

[1] Lu Q, Chan S T, Xu R F, et al. A Unicode based Adaptive Segmentor[J]. Journal of Chinese Language and Computing, 2004, 14(3): 221-234.

[2] 俞士汶，段慧明，朱学锋，等. 北大语料库加工规范：切分·词性标注·注音[J]. Journal of Chinese Language and Computing, 2003, 13(2): 121-158.

[3] 魏培泉，谭朴森，刘承慧，等. 建构一个以共时与历时语言研究为导向的历史语料库[J]. Computational Linguistics and Chinese Language Processing, 1997, 2(1): 131-145.

[4] 中央研究院近代汉语语料库[DB/OL]. http://early_mandarin.ling.sinica.edu.tw/

[5] 郑尔宁. 近二十年来现代汉语称谓语研究综述[J]. 语文学刊，2005，2：120-122.

[6] Dickey E. Forms of address and terms of reference[J]. Journal of Linguistics, 1997, 33(2): 255-274.

[7] Braun F. Terms of Address: Problems of patterns and usage in various languages and cultures[M]. Berlin, New York, Amsterdam: Mouton de Gruyter, 1988.

[8] 李学勤主编，(晋)郭璞注. 尔雅注疏[M]. 北京：北京大学出版社，1999：116-123.

[9] (清)梁章钜. 称谓录[M]. 长沙：岳麓书社，1991.

[10] 杨应芹，诸伟奇. 古今称谓词典[M]. 合肥：黄山书社，1989.

[11] 陆瑛. 简明称谓辞典[M]. 广西：广西民族出版社，1989.

[12] 韩省之. 称谓大辞典[M]. 北京：新世界出版社，1991.

[13] 吴海林. 中国古今称谓全书[M]. 哈尔滨：黑龙江教育出版社，1991.

[14] 吉常宏. 汉语称谓大词典[M]. 石家庄：河北教育出版社，2001.

[15] Xiong D, Lu Q, Lo F J, et al. Specification for Segmentation and Named Entity Annotation of Chinese Classics in the Ming and Qing Dynasties[C]//Proceedings of the Chinese Lexical Semantics (CLSW2012 Revised Selected Papers), Lecture Notes in Computer Science, Volume 7717. Berlin, Heidelberg: Springer, 2013: 280-293.

[16] 台湾经济部中央标准局. CNS14366，中文资讯处理分词规范[S]. 台湾：经济部中央标准局，1996.

[17] 国家技术监督局. 中华人民共和国国家标准 GB13715，信息处理用现代汉语分词规范[S]. 北京：中国标准出版社，1992.

[18] 夏迎炬，于浩，西野文人.《人民日报》语料库命名实体分类的研究[J]. Computational Linguistics and Chinese Language Processing, 2005, 10(4): 533-542.

(下转第 43 页)

2009,51:116-129.
[6] 乐明. 汉语篇章修辞结构的标注研究[J]. 中文信息学报, 2008,22(4):19-23.
[7] 孔庆蓓. 从修辞结构理论看叙述语篇和描写语篇的区别[J]. 南开语言学刊, 2008,2:92-104.
[8] 杨晓虹,杨玉芳. 汉语语篇修辞结构边界韵律表现[J]. 清华大学学报(自然科学版), 2009,49(S1):1375-1379.
[9] 胡苑艳,陈莉萍. 修辞结构理论与汉语篇章结构[J]. 长春大学学报, 2011,21(1):39-43.
[10] 傅间莲,陈秀群. 基于规则和统计的中文自动文摘系统[J]. 中文信息学报, 2006,20(5): 10-16.
[11] 袁毓林. 信息抽取的语义知识资源研究[J]. 中文信息学报, 2002,16(5):8-14.

赵建军(1976—),博士,主要研究领域为韵律学,认知语言学。
E-mail: zhaojianjun768@163.com

杨玉芳(1950—),博士,研究员,主要研究领域为心理语言学。
E-mail: yangyf@psych.ac.cn

杨晓虹(1984—),博士,助理研究员,主要研究领域为言语认知。
E-mail: yangxh@psych.ac.cn

(上接第 27 页)

熊丹(1980—),硕士,主要研究领域为词汇语义学。
E-mail: csdxing@comp.polyu.edu.hk

罗凤珠(1955—),博士候选人,副教授,主要研究领域为中国古典诗词,数位人文,文学地理学。
E-mail: gefjulo@mail2000.com.tw

陆勤(1960—),博士,教授,主要研究领域为计算语言学,词汇语义学,中文信息处理,基于自然语言处理技术的信息抽取和知识发现。
E-mail: csluqing@comp.polyu.edu.hk