基于语料库的武侠与仙侠网络小说文体、词汇及主题对比分析

张三乐　　刘鹏远*　　张虎
北京语言大学　信息科学学院
国家语言资源监测与研究平面媒体中心
北京市海淀区学院路15号, 100083
sanle0409@163.com　liupengyuan@pku.edu.cn　1170226830@qq.com

摘要

网络文学在我国发展迅猛，其数量和影响力呈现逐年上升的趋势，但目前尚无公开的较大规模网络文学作品语料库，鲜见基于语料库对网络文学具体类别作品的定量研究。本文初步建立了一个网络文学语料库，其中包括武侠和仙侠网络小说，使用文本计量、词频统计以及主题挖掘的方法对两类小说的文体风格、具体词汇使用和小说主题进行对比分析。通过比较，我们发现两类小说的文体风格大致相同，它们在词汇的使用和主题上既有共性又各具特色。从微观到宏观，从表面到内容，将定量统计和定性分析相结合，多角度、多层次的对武侠和仙侠网络小说进行比较。

关键词： 网络文学；武侠小说；仙侠小说；文体风格；词汇使用；主题

A Corpus-based Contrastive Analysis of Style, Vocabulary and Theme of Wuxia and Xianxia Internet Novels

Sanle Zhang　　Pengyuan Liu*　　**Hu Zhang**
Beijing Language and Culture University, School of Information Science
Language Resources Monitoring and Reserch Center
15 Xueyuan Road, Haidian District, Beijing, 100083, China
sanle0409@163.com　liupengyuan@pku.edu.cn　1170226830@qq.com

Abstract

Internet literature is developing rapidly in our country, and its number and influence are increasing year by year. However, there is no publicly large-scale online literary corpus, and there are few quantitative researches on specific types of online literature based on corpus. This article has initially established a corpus of online literature, including Wuxia and Xianxia online novels, using text measurement, word frequency statistics and topic mining methods to compare the stylistic style, specific vocabulary use and novel themes of the two types of novels. Through comparison, we find that the styles of the two types of novels are roughly the same, and they share commonalities and distinctive features in terms of vocabulary use and themes. From the micro to the macro, from the surface to the content, it combines quantitative statistics and qualitative analysis to compare Wuxia and Xianxia online novels from multiple angles and levels.

Keywords: Internet literature , Wuxia Novels , Xianxia Novels , Stylistic style , Vocabulary using , Theme

* 通讯作者　Corresponding Author

第十九届中国计算语言学大会论文集，第10页-第19页，海南，中国，2020年10月30日至11月1日。

1 引言

中国通俗文学发展至今，武侠小说始终是其中一个重要类别。金庸、古龙、梁羽生等老一辈的武侠大家所创作的作品曾引起社会上广泛的武侠小说阅读热潮，然而在这些武侠大家退隐之后，武侠作品的创作就陷入低潮期(张珍珍, 2017)。随着网络时代的来临，网络文学在此背景下逐步发展起来，至今已有20余年，其内容和影响力呈现逐年上升的趋势。网络的普及以及数字阅读平台的构建给网络小说提供了创作依托，大批网络写手利用网络平台发表自己的作品，受到广大读者的喜爱与追捧，涌现出大批网络原创小说。新时代网络作者接过金庸、古龙的接力棒，将武侠作品融入新时代元素，使其再度成为一大热潮，同时，由于IP改编影视剧的影响，仙侠小说逐渐走进大众视野，不仅在国内有一大批读者，在国外也很受欢迎。

从文学发展的角度来看，仙侠小说直接脱胎于武侠小说，是在武侠小说的在基础上发展起来的一种新型小说类型，在各大网络文学网站上的点击阅读量居高不下。虽然仙侠小说在网络小说中数量庞大，广受读者欢迎，但是关于它的研究仍处于网络文学研究的边缘地带，与它的发展不匹配(段晓云, 2018)。时至今日，网络文学的研究已取得了较大成就，涌现出一批代表性学者，如黄鸣奋、欧阳友权等，对网络文学的研究做出了巨大的贡献。目前,国内对于网络文学的研究一般以西方的理论研究为基本背景，研究着眼点放在网络文学的个别文本(崔宰溶, 2011)，且研究角度集中在文艺批评、文学特色和文化产业等方面，从定性的角度研究网络文学的特点，多把网络文学当做整体进行研究探讨，尚无公开的较大规模的网络文学作品语料库，鲜见学者基于大规模语料对网络文学具体类别的作品进行定量方面研究。

本文初步建立了一个网络文学语料库，语料来源于国内最大、最有影响力的网络文学网站———起点中文网[0]。该语料库目前包含网络武侠小说和仙侠小说，每种语料大小各100M byte，分别约2380万和2440万词次。基于这个语料库，本文对两类小说文本进行文本计量、词频统计分析以及主题挖掘，试图回答以下问题：

1）在宏观上，网络武侠与网络仙侠两类小说在文体风格上是否相同？为什么？
2）在微观上，两类小说在具体词汇的使用上有哪些异同？各有何种特色？
3）在内容上，两类小说在小说主题上有哪些异同？

2 相关工作

计量风格学产生于1851年英国数学家和逻辑学家Augustus De Morgan的猜想，他认为不同作家的作品风格可以通过隐形的数据特征进行辨别(Herdan, 1964)。近年来，计量风格学更广泛的应用于现当代文学研究领域。

在国外，Chaski (2001)从句法、标点符号、句子复杂度、文本易读性等方面对四位同龄女作者的部分作品进行了分析和比较；Argamon and Levitan (2005)等人认为功能词最容易反映作者的语言风格，并提出了675个能够反映作者风格的功能词；Grieve (2007)则以词首、词尾中字母的频率和包含各个字母的单词频率作为特征对作品进行分析等。在国内，刘颖and 肖天久 (2014)运用文本聚类和N元文法对词长分布、词类等语言特征进行考察，发现《红楼梦》前八十回和后四十回存在较大差异，得出其非一人所作的结论；金迪 (2018)采用数理统计学中定量分析的手段和方法，以计量风格学的视角，从频率统计和假设检验两大角度探究余华和格非小说在词汇和句子层面上的差异，从而分析二者的语言风格。

众多计量风格学的研究是在语料库的基础上对所选择的语言特征项进行分析，语料库在研究中起到了重要的作用。20世纪80年代开始，将语料库运用到文学作品中的研究逐渐升温，为文学研究提供了一个全新的视角，基于语料库的文学研究这一具有鲜明实证研究特征的文学研究领域应运而生(胡开宝and 杨枫, 2019)。

在国外，Stubbs (2005)以康德拉小说《黑暗之心》为语料库进行研究，发现其主题词为不确定的实词、虚词、以及抽象名词和带有否定前缀形容词的名词词组；Mahlberg (2007)以狄更斯的23部作品为语料库，分析其中的高频词簇，发现和身体部位相关的词簇常常可以推动故事情节发展。在国内，刘宇凡et al. (2011)等人将唐代以来的文学作品按不同时期分类建立语料库并对其进行字频分析，发现唐代以来人们使用汉字的习惯处于不断变化之中，时期越相近，汉字的使用习惯越一致; 陈建生and 王岩 (2016)将厄普代克所著“兔子系列”小说语料库与厄普代

[0]起点中文网：www.qidan.com。

克其他类型小说语料库中的关键词进行比较分析，发现第三人称代词高频出现，且多与心理动词搭配。涂梦纯and 刘颖 (2019)以余华和莫言的各5部小说为语料库，对二者的用词特征进行详尽的分析，讨论了量词、拟声词等词，发现了莫言用词丰富、情感充沛以及文言化、乡土化的特征，而余华与之相比白话、冷静、讽刺的风格。

由上可知，随着时代的进步和技术的发展，计量风格学和语料库语言学越来越多的应用于文学作品研究领域，为文学作品研究提供了新的视角和方法，但研究多集中在传统经典的文学作品，对新兴的网络文学领域却极少涉及。

在网络文学研究方面，欧美的网络文学研究成果已经相当可观，他们的网络文学研究趋势集中在文学实验的理论性研究。Bolter (1991)创造性的用解构主义理论观照“超文本”得叙事特点，指出超链接使“超文本”具有多重阅读路径，彻底颠覆了传统的线性阅读的叙事模式，从而体现了解构主义取消中心、无限变更的开放活动的观念；Aarseth (1997)提出“制动文本理论”，为广大网络文学研究者提供了新的视角，掀起了网络文学研究的热潮；近年来，德、美合作出版了网络文学研究系列论文集，至今已出版四集，从不同的视角对网络文学进行全面的探讨，代表着当今西方网络文学研究的前沿(王艳, 2016)，如Beyond the Screen。在国内，网络文学研究也在迅速发展。黄鸣奋 (2002)阐释了网络文学贵在鲜活、追求互动的网络根本特质；欧阳友权 (2004)在哲学意义上探讨网络文学，以“本体论”和“现象学”讨论网络文学的终极意义。后来，越来越多的学者从新颖的观点和角度对网络文学进行研究，如王黎 (2010)从女性主义的角度去分析网络文学；张珍珍 (2017)描述了网络武侠小说兴起的文化背景，讨论了网络武侠小说和传统武侠小说的内在关联，分析了网络武侠小说在类型创造、传播方式上的转变；段晓云 (2018)从文学空间切入，探讨网络仙侠小说中文学空间的描写。

综上所述，网络文学作为一种新兴的文学形式，广泛的受到人们的喜爱，也吸引了众多学者的目光。目前，网络文学定性研究在我国已取得不小的成就，这些研究多集中在文艺批评、文学特色和文化产业、基础学理研究等方面。然而，很少有学者从计量、统计等定量的角度对网络文学作品的特点进行探究，也尚无公开的较大规模的网络文学语料库。本文初步建立了一个网络文学语料库，包括网络原创武侠小说和仙侠小说，使用计量、词频统计和主题挖掘的方法，多层次、多角度的对这两类网络文学文本的特点进行探究。

3 数据

本文初步构建了一个网络文学语料库，语料库构建步骤大致如下：

- 数据获取。编写爬虫软件，随机抓取了起点中文网中上架于2015年至2017年、作品分类为“武侠”和“仙侠”的两类小说。
- 数据清洗。对每一类别抽取了100M字节的文本，对其中存在的网站标语、乱码等进行处理，对语料进行整理，去掉其中的重复行，将标点符号由半角转为全角。
- 分词和词性标注。整理完语料之后，采用分词和词性标注工具jieba[1]对文本进行分词和词性标注。由于网络小说中的词汇和句法较为随意，会有一些未登陆词，在后续实验中不予考虑。

最终，本文的研究数据基于上述语料，共约7000万字，语料库数据基本情况如表1所示。

Table 1: 网络文学语料库数据分布

文本类别	总字数	总词数	标点数	句子数	段落数
武侠小说	35825816	24405279	30041194	1043739	369989
仙侠小说	35034702	23845448	29511768	1027756	364526

4 文本风格比较

本文从四种最常见的计量指标以及词类的分布对网络武侠小说和仙侠小说进行了风格计量，并使用统计检验的方法检验二者之间是否存在显著性差异，从而比较二者的文体风格。

[1]https://github.com/fxsjy/jieba

4.1 基本计量指标比较

本文从词汇丰富度、文本可读性、句子离散度、句子破碎度和词类分布这些角度进行计量比较：

- 词汇丰富度

  词汇丰富度是指作者在文本中使用词汇的丰富程度。本文选择词频广度和型例比作为考察词汇丰富度的参数特征。词频广度即计算高频词之外的词语的比例，本文设定高频词为覆盖率90%的词语。型例比，是指文本中不同的词语在所有词语中所占的比例，公式如下：

$$TTR = Types/Tokens$$

- 文本可读性

  文本可读性最初由教育学家Dale提出。文本可读性高可理解为文本简单，文本复杂度低。文本可读性可以从平均词长、平均句长来初步考察。平均词长是所用词汇的平均字数，平均句长是所用句子包含的平均字数。

- 句子离散度

  句子离散度即文本中句子的句长偏离平均句长的长度，可以提现出文本节奏，计算公式如下：

$$D = \sqrt{\frac{1}{n}\Sigma\left(L_i - L_0\right)^2}$$

  公式中D表示句子离散度,L_i表示每句句长，L_0表示平均句长，n表示句子的总数。

- 句子破碎度

  破碎度即在句子中停顿的次数，可以侧面体现文本的语体色彩。计算公式为：句子破碎度=句子停顿次数[2]/句子总数。

我们在对网络武侠小说和仙侠小说进行基本计量之后，将两类小说各平均分成十份小语料，每份约10M，使用spss软件对每一类计量指标结果进行独立样本T检验，两类语料计量和统计检验结果见表2。

Table 2: 网络文学文本计量数据及统计检验结果

		武侠小说	仙侠小说	P值
词汇丰富度	型例比	0.151	0.186	0.541
	词频广度	0.875	0.962	0.259
文本可读性	平均词长	1.468	1.469	0.076
	平均句长	34.324	34.089	0.831
句子离散度		37.500	24.909	0.042
句子破碎度		2.424	2.328	0.225

通过表格可以看出，网络武侠小说和仙侠小说的型例比、词频广度、平均词长、平均句长和句子破碎度的P值都大于0.05，没有达到显著水平，即两类小说的词汇丰富度、文本可读性、句子破碎度没有显著差异。

在此之外，两类小说句子离散度的P值为0.042，说明两类小说的句子离散度存在显著差异。通过比较，发现武侠小说的句子离散度比仙侠小说更高，说明武侠小说中句子长短不一，跳跃性大，使文本富于节奏变化，使读者阅读起来长短不一、错落有致，有着抑扬顿挫、跌宕起伏的感受，阅读体验感强；而仙侠小说的句子离散度较低，说明其文章节奏较为缓和，读者阅读时有较为严肃、平缓的体验感。

[2]黄柏荣and 廖序东 (2002)在《现代汉语》中指出，“点号主要用来表示句子中的各种停顿”，并将点号分为句中点号（逗号、顿号、分号和冒号）以及句末点号（句号、感叹号、问号）。本文在计算句子破碎度中采用了这一标准。

4.2 词类分布比较

词类是词的语法分类，是词在语法结构中表现出来的类别。不同的词在文本中起着不同的作用，在文本风格分析中词类的使用频率是构成文本风格的重要特征之一[3] 。我们使用编程对网络武侠小说和仙侠小说的部分词类进行统计，然后同样将两类小说各平均分成十份，使用spss软件对其进行独立样本T检验，从而比较两类小说的异同。

Table 3: 网络文学文本的词汇分布和统计检验结果

实词	武侠小说	仙侠小说	P值	虚词	武侠小说	仙侠小说	P值
名词	0.1255	0.1242	0.165	拟声词	0.0005	0.0007	0.068
动词	0.1188	0.1193	0.464	连词	0.0170	0.0171	0.064
形容词	0.0176	0.0189	0.046	介词	0.0167	0.0170	0.771
数词	0.0240	0.0253	0.011	助词	0.0460	0.0510	0.175

由上可知，经过统计检验，网络武侠小说和仙侠小说的名词、动词、拟声词、连词、介词和助词的P值都大于0.05，说明两类小说在这些词类的使用频率上没有显著差异；形容词和数词的P值小于0.05，说明两类小说在形容词和数词的使用频率上差异较为显著。

形容词在汉语中充当修饰成分，在小说中的作用主要是对人物的刻画、对环境的渲染以及对场景的描写。数词使用时常常与量词搭配作定语，也起到修饰的作用，在小说本文中使用数词，说明其更加注重细节描写，给人以真实感。由表可知，仙侠小说的形容词和数词使用频率较高，如：

例：一缕琴声随风传来，缓如溪水流泉，脆如珠落玉盘，叮叮咚咚空灵有质。随着琴声渐清，一丝歌声却是悱恻辗转，酥人心扉.....（选自仙侠小说）

说明仙侠小说中人物刻画和环境描写更加丰富、细致，使文章内容生动形象，提高了读者的阅读体验，更能吸引读者眼球，一定程度上说明了仙侠小说越来越受读者喜爱的原因。

总之，通过对网络武侠小说和仙侠小说进行计量和统计检验，发现除了句子离散度、形容词和数词使用频率有差异之外，其他指标都没有显著差异，说明两类小说的文体风格基本相同。这可能是由于仙侠小说是在武侠小说的基础上发展起来的，且二者都属于网络文学文本，其用词、用句、行文结构等都大致相似，因而文体风格上没有显著差异。

5 具体词汇使用比较

本文从词频统计的角度考察网络武侠小说和仙侠小说的词汇使用情况，探究两类小说的用词风格。由于实词在文本中主要承担表达意义的作用，有具体的词汇意义，所以我们选择使用频率较高且有丰富词义的名词、动词、形容词进行对比分析。根据语料库的词性标注结果，我们通过编程抽取了两类小说的名词、动词和形容词，去除停用词、人名、地名以及“说、有”之类的常用词，统计两类小说按照频率排序的共用词和独用词的使用情况。

5.1 不同词类下的共用词

按照频率列出前500个高频词中名词、动词和形容词的共用词，各取前20词。统计结果如下：

通过对不同词类的共用词进行比较，可以发现网络武侠小说和网络仙侠小说在具体词汇使用上的共性。我们将这些高频的共用名词、动词按照语义进行粗略分类，名词可以分为武功武器词、身体部件词、人物关系词以及其他，动词可以分类为使令动词、动作动词以及心理动词。另外，按照郭伊迪 (2012)中的分类将统计出的共用形容词进行分类，可以分为度量形容词、情绪形容词和色彩形容词。

在网络武侠小说和网络仙侠小说的共用名词中，人物关系词占比最高，符合小说注重刻画人物关系的特点。在人物关系词中，如“弟子、师父、大哥”等，发现描写对象多为男性，而且关系多为师徒、父子、师兄弟关系。这说明男性往往是小说中的主要角色，且对师徒、父子和兄弟情义描写较多，突出了两类小说中侠肝义胆的人物情感和侠义色彩。在武功武器词中，“剑”的使用频率最高，其他词也多与“剑”相关，如“剑法、剑气、长剑”，另外还有“刀、气

[3]道格拉斯·比伯、苏珊·康拉德、兰迪·瑞潘 (2012)《语料库语言学》，刘颖、胡海涛译，北京：清华大学出版社，2012年，第43页。

Table 4: 武侠和仙侠网络小说按照频率排序的共用词

名词	武功武器词	剑、刀、剑法、长剑、剑气、气息
	身体部件词	手、眼睛、脸、
	人物关系词	弟子、师父、大哥、前辈、父亲、师弟、长老、掌门
	其他	风、山、门派
动词	使令动词	派、令
	动作动词	死、杀、跑、出手、救、抓住、拉、打、修炼、盯、追、愣、跳、躲
	心理动词	怕、担心、放心、喜欢
形容词	度量形容词	深、快、大
	情绪形容词	好、平静、激动、紧张、难、简单、轻松、诡异、干净、厉害、尴尬、逍遥、犹豫、强大
	色彩形容词	白色、红色、黑色

息”，说明两类小说中最常使用的武器是“剑”，“刀”次之，而且都注重内功。在身体部件词中，如“手、眼睛、脸”，说明其细节描写丰富。在其他词中，“门派”一词符合两类小说的特点，有共同信仰和武功继承的人同处于一个派系，门派和门派之间形成敌友关系，构成小说的故事网络。另外，其他词中还包括“风、山”这样的自然风景词，表明小说环境烘托较为丰富，提高了作品的阅读性。

在两类小说的共用动词中，动作动词占比最高，符合小说注重动作刻画的特点。两类小说使用了大量相同的动作动词，说明二者在动作情节的描写上有一定的相似性，例如“死、杀、出手、救”，说明二者都有着极为丰富的打斗、征战、死伤的情节；又如“打、跑、抓住、拉、追”等，用不同的动词细致的表示不同的动作形态，丰富了小说内容；对人物表情也有一定描写，如“盯、愣”；值得注意的是“修炼”一词，一般在文中搭配“武功、法力”，体现出二者对武功描写着墨较多。在心理动词中，如“怕、担心、放心”等，说明其注重心理描写，有着较为丰富的情绪表达。另外还有一些使令动词，从中可以看出“派遣”、“命令”的行为较多，如“派、令”，也表明这两类小说中普遍存在着地位等级关系。

在两类小说的共用形容词中，情绪形容词的占比最高，其中有一些对人物情绪的刻画，如“平静、激动、紧张”等，还有一些对事物的主观判断，如“好、简单、难”等。度量形容词的高频使用说明小说中有一些对事物性质的叙述，如“深、快、大”等。另外，两类小说中的还有一些色彩描写，“如白色、黑色、红色”，丰富了小说色彩，增强了小说的趣味性和形象性。

总之，通过比较发现，网络武侠小说和网络仙侠小说都注重人物刻画和动作描写。两类小说的男性角色较为突出，普遍存在着地位等级关系；在动作描写上有一定的相似性，常用“剑”作为武器，有着丰富的武打情节，注重细节刻画和环境烘托。

5.2 不同词类下的独用词

按照频率列出前500个高频词中名词、动词和形容词的独用词，各取前20词。统计结果如下：

通过对不同词类的独用词进行比较，可以看出网络武侠小说和网络仙侠小说在具体词汇使用上各具特色。同样，我们将这些高频的名词、动词和形容词的独用词按照语义进行粗略分类，名词可以分为武功武器词、门派派系词、人物动物词以及其他，动词全部划分为动作动词，形容词划分为颜色形容词和情绪形容词。

在网络武侠小说和网络仙侠小说各自的独用名词中，我们可以发现网络武侠小说中多使用冷兵器和内功，例如“刀法、枪、内力、轻功”等；而网络仙侠小说则多使用法力、法术，如“灵力、法术、法力”等，体现出武侠小说传统、写实和仙侠小说创新、虚幻的特点。由上共用词可知“门派”一词的在两类小说中得到高频使用，在独用词中我们就可以看出两类小说在门派派系词在具体使用上的不同。可以看出武侠小说中的门派多使用“盟、教”构词，如“教主、红衣教、金兰盟”等，而仙侠小说则多使用“宗、族”，如“宗门、人族、魔族”，具有神化色彩，另外，武侠小说的门派词使用要多于仙侠小说，说明武侠小说中的派系较多。在人物动物词中，武侠小说中使用了“公主、陛下、夫人”等词，体现出其真实性和历史性；而仙侠小说中人造、虚构的人物较多，例如“仙人、凡人、女娲”等，除了人物词，仙侠小说中的动物词使用要多与武侠小

Table 5: 武侠和仙侠网络小说按照频率排序的独用词

		武侠小说	仙侠小说
名词	武功武器词	暗器、刀法、穴道、枪、兵刃、内力、修为、轻功	灵力、法术、法力、炼气、法器
	门派派系词	教主、红衣教、丐帮、盟主、金兰盟	宗门、人族、魔族
	人物动物词	公主、女儿、老头、夫人、陛下、猴	仙人、凡人、女娲、妖兽、鹤、妖怪、僵尸、野兽、
	其他	镖局	灵魂、宝物、血影、血光
动词	动作动词	争宠、娶、围攻、切磋、厮杀、打擂、拔剑、逃走、下马、联姻、拿下、联手、扫视、飞扬、失踪、打探、追赶、出鞘、刺杀、埋伏	修行、飞行、叫魂、飞升、尸变、提升、炼丹、炼制、吸收、吞噬、穿越、渡劫、传送、喷出、斩杀、砸断、燃烧、弥漫、毁灭、下狱
形容词	色彩形容词	深红、暗橙	金色、紫色、蔚蓝、血色、丹色、青色
	情绪形容词	慈悲、仗义、娇羞、柔情、冷峻、胆怯、敏捷、刚猛、焦躁、羞涩、精明、妩媚、孤傲、肃然、稀奇、雄厚、精壮、悦耳	混沌、硕大、古朴、清凉、浓郁、浩瀚、清冷、稚嫩、强横、充沛、炽热、轻易、坚挺、猖狂

说，如“妖兽、鹤、野兽”等，在现实的基础上进行虚构，体现出其创造性和奇幻、志怪色彩，提高了小说的趣味性。在其他词中，武侠小说中有“镖局”一词，同样体现其写实的特点；仙侠小说中有“灵魂、宝物、血影、血光”，这些字眼更具有暗黑色彩，营造出恐怖神秘的场景。

在两类小说各自的独用动词中，可以发现，武侠小说更加注重对打斗方式的描写，如“围攻、刺杀、拔剑、追赶”等，更关注传统武打和江湖世界；在武侠小说中构造的故事更加贴近生活，如“争宠、娶、打擂、联姻”，体现了当时的社会特色，重现了当时的社会场景，具有社会性和真实性，拉近了读者与故事的距离，产生亲切感。仙侠小说则不同，仙侠小说中使用了一些动能较高的动词，如“斩杀、砸断、吞噬”，给人以冲击感；在武功方面突破传统武功技法，如“叫魂、飞行、飞升”，还使用了如“尸变、穿越、渡劫”等词，突破现实世界，具有虚构色彩，给人以新奇之感。

在两类小说各自的独用形容词中，我们可以发现，武侠小说中的情绪形容词中多是对人物情感、性格的描写，如“慈悲、仗义、娇羞”等，而仙侠小说则多为对事物的主观判断，如“混沌、硕大、古朴”。另外，两类小说含有一些色彩形容词，如武侠小说中有“深红、暗橙”，仙侠小说中则更加丰富，如“金色、紫色、蔚蓝”等颜色词，颜色词层次更加丰富，种类较多，为读者构建了色彩斑斓的世界，增强了小说的趣味性和形象性，提高了读者的阅读体验。

总之，通过比较，可以看出两类小说在具体使用词汇上的不同。网络武侠小说更加关注传统武打和现实世界，多使用冷兵器和内功，门派多为“盟、教”，且派系相对较多，具有真实性、社会性和历史性，还将笔墨更多的用于人物刻画，让读者从字里行间体会人物情绪；网络仙侠小说则突破传统武功技法和现实世界，多使用法力、法术，门派多为“宗、族”，具有创造性、虚构性和灵异、志怪、暗黑色彩，常给小说营造出恐怖神秘的场景，注重人对事物的描写和对周边世界的评价，还使用了较为丰富的色彩形容词，更具趣味性。

6 主题比较

LDA(latent dirichlet allocation,隐狄利克雷分配)，是一种采用词袋模型的文档主题生成模型，能够把文本中的主题自动汇集。它的基本思想是假设所有的文档存在K个隐藏主题，一篇文档的每个词都是以一定概率选择了某个主题，并从这个主题中以一定概率选择了某个词语，不断的抽取隐含主题及其特征词，直到遍历完文档中的全部单词。

本文使用LDA模型对网络武侠小说和仙侠小说两类文档集合进行主题建模，挖掘文本隐含的主题信息，对两类小说的主题进行比较。

6.1 语料预处理

在原有语料分词和词性标注的基础上，我们去除了语料中的停用字，过滤了标点符号以及无意义的词，如数词、量词、虚词等等，以避免对模型最终结果的影响，降低噪音。然后，根据小说文本的特点，把小说按照不同的章节分隔开来，作为独立的文档。

6.2 主题比较分析

我们使用LDA模型对两类小说进行主题建模，指定主题数为150个，设置每个主题打印出最能描述该主题的前20个词。由于篇幅限制，将其中部分主题进行总结并制成词云进行比较，图1是武侠小说的两类主题，图2是仙侠小说的两类主题。

Figure 1: “受害”和“敛财”主题

Figure 2: “修炼”和“降妖”主题

根据主题聚类的结果，我们将武侠小说中的主题总结为：受害、敛财、出征、联姻、门派等，将仙侠小说中的主题总结为：修炼、降妖、打斗、魔界、天界等（见附录）。

可以发现，网络武侠小说和仙侠小说在主题上有一定的相关性，比如两类小说中都出现了有关“武打”的主题，如“受害”和“打斗”主题。在这类主题中出现了许多关于细节描写的主题词，如“转身、胸骨、瞳孔、脚尖”以及“眼珠、汗毛、双手”等等，说明两类小说在描写打斗场面时会细化到人物的面部表情和身体变化；又比如两类小说中都出现了有关“派别”的主题，网络武侠小说中有“门派”主题，而仙侠小说中则有“魔界”和“魔界”主题，也体现出一定的相关性和对应性。

同时，两类小说的主题也有不同。网络仙侠小说中的主题多与“仙、魔”相关，主题词中也多以“仙、魔”构词，如“成仙、魔头、魔神”等等，说明仙侠小说多以神、人、魔三界为背景讲述故事，着重“修仙练功”和“降妖伏魔”等情节，更加新颖、新奇，具有神化色彩；而网络武侠小说的主题涉及范围较广，除了武功、打斗之外，还涉及到钱财、婚姻、君臣等社会的各个方面，说明武侠小说多以人民生活为背景讲述故事，各个场景都与人民生活息息相关。如“敛财”主题，其中有“打耳光、上交、商人”等主题词，更加传统、熟悉，具有现实色彩。

总之，通过对两类小说的主题进行挖掘，我们发现两类小说在主题既有一定的对应性，也有一些不同之处，从二者的主题词中也可看出两类小说的异同。两类小说都出现了一些相似的

主题，但是网络武侠小说的主题涉及范围更广，而网络仙侠小说的主题则更加集中。另外，两类小说都注重细节描写，但是网络武侠小说的主题词更加传统、熟悉，具有现实色彩，而仙侠小说的主题词更加新颖、新奇，具有神化色彩。

7 结论

网络文学的发展至今已有20余年，随着网络的普及以及数字阅读平台的构建，网络文学以其方便、经济的特点越来越受到人们的关注和喜爱，其内容和影响力呈现逐年上升的趋势。

本文建立了一个网络文学语料库，包括武侠和仙侠网络小说。通过对两类小说文本进行计量、词频统计以及主题挖掘，从宏观到微观，多层次、多角度的比较了武侠和仙侠网络小说的异同，回答了我们提出的三个问题。

从宏观上，我们对网络武侠小说和网络仙侠小说进行计量风格分析，发现两类小说的文体风格基本相同，这可能是由于仙侠小说是在武侠小说的基础上逐步发展起来的，用词、用句、行文结构都大致相似，因而文体风格没有显著差异。

从微观上，我们对两类小说的名词、动词和形容词进行统计分析，发现两类小说在具体词汇使用上既有共性又各具特色。从两类小说使用的词汇上可以看出，两类小说都注重人物刻画和动作描写，小说中男性角色较为突出，且普遍存在地位等级关系；有丰富的武打情节，在动作描写上有一定的相似性，常用“剑”作为武器，注重细节刻画和环境烘托，具有集体性、等级性和侠义色彩。网络武侠小说更加写实、传统，具有真实性、社会性和历史性，还将笔墨更多的用于人物刻画，让读者从字里行间体会人物情绪；网络仙侠小说则突破传统武功技法和现实世界，关注异世界，多使用法力、法术，具有创造性、虚构性和灵异、志怪、暗黑色彩，常给小说营造出恐怖神秘的场景，注重人对事物的描写和对周边世界的评价，还使用了较为丰富的色彩形容词，更具趣味性。

从内容上，我们使用LDA主题模型对两类小说进行主题挖掘，发现两类小说主题的异同。两类小说的主题具有一定的相关性，也有一些不同之处。两类小说出现了一些类似的主题，然而武侠小说的主题更加广泛，主题词更加传统、写实，具有现实色彩；而仙侠小说的主题则更加集中，主题词更加新颖、新奇，具有神化色彩。同时，也表明LDA模型可以应用于大规模小说语料的主题挖掘。

在后续工作中，我们将进一步扩充网络文学语料库，多角度、多层次，使用多种方法对各类网络文学进行研究。

致谢

感谢各位匿名评审老师和论文辅导老师的帮助。本论文受教育部人文社会科学研究规划基金资助项目（18YJA740030）和北京语言大学研究生创新基金项目（20YCX153）资助。

参考文献

Espen J Aarseth. 1997. *Cybertext: Perspectives on ergodic literature*. JHU Press.

Shlomo Argamon and Shlomo Levitan. 2005. Measuring the usefulness of function words for authorship attribution. In *Proceedings of the 2005 ACH/ALLC Conference*, pages 4–7.

Jay D Bolter. 1991. *Writing space*. Erlbaum.

Carole E Chaski. 2001. Empirical evaluations of language-based author identification techniques. *Forensic Linguistics*, 8:1–65.

Jack Grieve. 2007. Quantitative authorship attribution: An evaluation of techniques. *Literary and linguistic computing*, 22(3):251–270.

Gustav Herdan. 1964. Quantitative linguistics.

Michaela Mahlberg. 2007. Clusters, key clusters and local textual functions in dickens. *Corpora*, 2(1):1–31.

Michael Stubbs. 2005. Conrad in the computer: examples of quantitative stylistic methods. *Language and Literature*, 14(1):5–24.

刘宇凡, 郭金忠, and 陈清华. 2011. 唐代以来汉语文学作品中的字频演变. 中文信息学报, 25(3):93–98.

刘颖and 肖天久. 2014. 《红楼梦》计量风格学研究. 红楼梦学刊, (4):25.

崔宰溶. 2011. 中国网络文学研究的困境与突破. Ph.D. thesis, 北京大学博士学位论文.

张珍珍. 2017. 网络武侠小说的发展及其特色. Master's thesis, 青海师范大学.

欧阳友权. 2004. 网络文学本体论纲. 文学评论, 6:69–74.

段晓云. 2018. 网络仙侠小说文学空间研究. Master's thesis, 兰州大学.

涂梦纯and 刘颖. 2019. 余华与莫言长篇小说的计量统计和分析. 中文信息学报, 33(2):131–142.

王艳. 2016. 西方网络文学研究综述. 创新与探索*:* 外语教学科研文集.

王黎. 2010. 女性网络文学作者的创作倾向. Master's thesis, 山东大学.

胡开宝and 杨枫. 2019. 基于语料库的文学研究: 内涵与意义. 浙江大学学报*(*人文社会科学版*)*, 5(5):130.

道格拉斯·比伯、苏珊·康拉德、兰迪·瑞潘. 2012. 语料库语言学. 清华大学出版社.

郭伊迪. 2012. 基于语义角度的形容词分类研究. Master's thesis, 黑龙江大学.

金迪. 2018. 基于语料库的格非, 余华小说计量风格学研究. Master's thesis, 南京师范大学.

陈建生and 王岩. 2016. 厄普代克"兔子系列" 小说特点的语料库文体学研究. 牡丹江大学学报, 25(9):22–24.

黄柏荣and 廖序东. 2002. 现代汉语.

黄鸣奋. 2002. 网络文学之我见. 社会科学战线, 4:15–16.

A 附录：武侠和仙侠小说的部分主题

	主题	武侠小说	主题	仙侠小说
1	受害	禀报、太子、刺杀、拦截、转身、陷阱、杀害、山贼、喷出、弥漫、瞳孔、纷飞、胸骨、脚尖、主持、凄厉、血水、功力、威胁、披风	修炼	师父、修炼、法宝、成仙、道友、真人、大师、教主、禅师、飞剑、光明、只能、施法、化作、祭炼、身体、飞升、化成、斗法、仙剑
2	敛财	防守、来路、打耳光、资金、抢夺、惯例、集散、上交、商人、主动权、缓和、余味、交给、鬼迷心窍、金子、引流、交入、小贩、呛着、紧跟	降妖	犹如、人族、抓住、危险、屋顶、金牌、村民、目光、击中、铃声、城墙、天师、妖兽、脸颊、飞刀、瞳孔、肩膀、菩萨、刀鞘、收服
3	出征	说道、安排、来到、皇上、格格、车马、师父、姑娘、战场、皇帝、回来、宝刀、说完、看着、离开、父亲、义军、吩咐、将军、总舵主	打斗	鲜血、眼珠、带头、弟弟、苍穹、汗毛、丧命、阻拦、直扑、双手、刺杀、飞剑、了结、流血、大片、穿过、带队、指使、飞掠、嚎叫
4	联姻	兵马、身份、去路、迎战、退身、女人、摆手、使臣、解除婚约、公主、嫁入、男人、马车、收买、帮忙、皇上、征讨、将领、边境、协议	魔界	弟子、光明、魔头、法力、神魔、炼成、放出、魔法、魔教、出手、佛光、灰尘、魔神、敌人、宝物、石像、神光、正宗、血影、尊者
5	门派	方志、师父、弟子、想到、不知、全真教、见到、功夫、不由、修习、神功、掌门、少林、想着、实在、丐帮、担心、徒弟、教主、内功	天界	祖师、修行、猴子、不由、佛祖、真人、修为、不知、看着、天地、人类、老道、古灵精怪、菩萨、天庭、微笑、混沌、师父、点头、雀儿