

分类号\_\_\_\_\_

密级\_\_\_\_\_

UDC\_\_\_\_\_

编号\_\_\_\_\_

華中師範大學

硕士学位论文

基于统计学的金庸小说个人语言  
风格及疑似作品的证伪研究

学位申请人姓名: 罗男

申请学位学生类别: 全日制硕士

申请学位学科专业: 语言学及应用语言学

指导教师姓名: 沈威 副教授



# 硕士学位论文

## 基于统计学的金庸小说个人语言风格 及疑似作品的证伪研究

**论文作者： 罗男**

**指导教师： 沈威**

**学科专业： 语言学及应用语言学**

**研究方向： 应用语言学**

**华中师范大学语言与语言教育研究中心**

**2020 年 5 月**



# Research on the Personal Language Style of Jin Yong's Novels and the Falsification of Suspected Works Based on Statistics

*A Thesis*

Submitted in Partial Fulfillment of the Requirement  
*For the M.A. in Linguistics and applied linguistics*

By

Nan Luo

Postgraduate Program

Research Center for Language and Language Education  
Central China Normal University

Supervisor: Wei Shen

Academic Title: Associate Professor

Signature

Approved

May. 2020



## 华中师范大学学位论文原创性声明和使用授权说明

### 原创性声明

本人郑重声明：所呈交的学位论文，是本人在导师指导下，独立进行研究工作所取得的研究成果。除文中已经标明引用的内容外，本论文不包含任何其他个人或集体已经发表或撰写过的研究成果。对本文的研究做出贡献的个人和集体，均已在文中以明确方式标明。本声明的法律结果由本人承担。

作者签名： 罗男

日期： 2020 年 6 月 15 日

### 学位论文版权使用授权书

学位论文作者完全了解华中师范大学有关保留、使用学位论文的规定，即：研究生在校攻读学位期间论文工作的知识产权单位属华中师范大学。学校有权保留并向国家有关部门或机构送交论文的复印件和电子版，允许学位论文被查阅和借阅；学校可以公布学位论文的全部或部分内容，可以允许采用影印、缩印或其它复制手段保存、汇编学位论文。（保密的学位论文在解密后遵守此规定）

保密论文注释：本学位论文属于保密，在 \_\_\_\_\_ 年解密后适用本授权书。

非保密论文注释：本学位论文不属于保密范围，适用本授权书。

作者签名： 罗男

导师签名： 沈威

日期 2020 年 6 月 15 日

日期： 2020 年 6 月 15 日

本人已经认真阅读“CALIS 高校学位论文全文数据库发布章程”，同意将本人的学位论文提交“CALIS 高校学位论文全文数据库”中全文发布，并可按“章程”中的规定享受相关权益。同意论文提交后滞后： 半年； 一年； 二年发布。

作者签名： 罗男

导师签名： 沈威

日期： 2020 年 6 月 15 日

日期： 2020 年 6 月 15 日



## 摘要

金庸是二十世纪中国文学最为瞩目的小说家之一，他的武侠小说无疑具有超一流的水准。金庸小说中的爱恨情仇、侠肝义胆一直以来为人们所津津乐道。语言是动人故事与丰富精神的承载者与传递者。金庸小说语言具有自成一家的独特风貌。目前学界对金庸小说的研究角度多集中于文学、美学等方面，尤其缺少运用量化手段展开的相关研究。本文基于统计学的相关理论与方法，利用计量的手段对金庸的15部小说进行全面而客观的分析，以探究其小说的语言风格特征。另外，还对金庸的一部疑似作品《卧龙记》进行了计量分析，并对该部疑似作品的作者是否为金庸进行了判别。

本文主要从字、词汇、句子、段落等四个语法单位来对小说文本的语言风格进行探索。在不同的语法单位上，本文选取了不同的统计指标进行数据的统计。除此之外，还利用基于TF-IDF算法和LSI算法与余弦相似度相结合的方法、情感分析与卡方检验相结合的方法对文本相似度进行了计算与检验。

本文就内容上可以分为两大块：金庸公认的15部作品的语言风格研究和疑似作品《卧龙记》的证伪研究。

就金庸公认的15部作品的语言风格研究而言，统计结果表明，金庸15部小说的平均词长、词长离散度、词汇密度、平均句长、句长离散度、平均段落长度之间存在着差异，并呈现出无序的状态。而在标点符号、用字量、词汇丰富度、词类分布、句长分布等方面具有某些可循的规律。在标点符号的使用上，金庸的15部小说之间具有较高的一致性；用字量方面，金庸小说小长篇内部与超级长篇内部的用字量相对稳定；就词汇丰富度而言，金庸中短篇小说的词汇多样性高于长篇小说，同等量级的小说之间的差异较小；就词类分布情况来看，金庸15部小说中实词内部分布顺序具有一致性，虚词内部分布顺序也相一致；在词长分布与句长分布上，各小说的具体分布图式具有相似性。

另外，文本相似度计算结果表明，短篇小说《越女剑》与其他金庸小说相似度最低，且远远低于其他小说间的相似度。在六部小长篇中，《连城诀》和《侠客行》两部小说与其他四部小说之间相似度较低，而《雪山飞狐》等其他四部小说之间的相似度都相对较高；小说间的相似度与创作时间的先后有着一定程度上的一致性。在六部超级长篇中，“射雕三部曲”之间具有较高的相似度，而连载时间最晚的《鹿鼎记》与其他小说之间的相似度都相对较高。情感分析中，总的来说，金庸小说呈



现出中性的情感倾向。

就疑似作品《卧龙记》的证伪研究而言，将《卧龙记》的统计结果与金庸小说相对比，发现在用字量、词长分布、类符形符比、词汇密度、频次排名前一百的词语、词类分布、平均句长、句长分布等语言特征上，《卧龙记》与金庸小说之间并无显著差异。而在标点符号的使用、平均词长、词长离散度、独现词频率、词频分布、句长离散度、平均段落长度、相似度计算与情感分析方面，两者有着明显不同。其中，文本相似度计算是可信度较高的文本判别方法。综合各部分的统计结果和分析来看，疑似作品《卧龙记》是金庸作品的可能性极低。

**关键词：计量研究；语言风格；金庸小说；《卧龙记》**



## Abstract

Jin Yong is one of the most eye-catching novelists of Chinese literature in the 20th century, and his martial arts novels undoubtedly have superb standards. The love and hatred in the novels of Jin Yong have always been talked about by people. Language is the carrier and transmitter of moving stories and rich spirits. The language of Jin Yong's novels has its own unique style. At present, the academic perspective of Jin Yong's novels is mostly focused on literature, aesthetics, etc., especially the lack of relevant research using quantitative methods. Based on the relevant theories and methods of statistics, this article uses measurement methods to conduct a comprehensive and objective analysis of Jin Yong's 15 novels to explore the language style characteristics of his novels. It also carried out a quantitative analysis of a suspected work "Wolong Ji" by Jin Yong, and judged whether the author of the suspected work was Jin Yong.

This article mainly explores the language style of the novel text from four grammatical units such as words, vocabulary, sentences, and paragraphs. In different grammatical units, this article selects different statistical indicators for data statistics. In addition, the text similarity was calculated and tested using a method based on the combination of TF-IDF algorithm and LSI algorithm with cosine similarity, sentiment analysis and chi-square test.

The content of this article can be divided into two parts: the study of the language style of 15 works recognized by Jin Yong and the falsification of the suspected work "Wolong Ji".

As far as the study of the language styles of 15 works recognized by Jin Yong is concerned, the statistical results show that the average word length, word length dispersion, vocabulary density, average sentence length, sentence length dispersion, and average paragraph length of the 15 novels of Jin Yong exist difference, and presents a disordered state. There are certain rules to follow in terms of punctuation, word size, vocabulary richness, part-of-speech distribution, sentence length distribution, etc. In terms of the use of punctuation marks, Jin Yong 's 15 novels have a high degree of consistency; the amount of words used is relatively stable in the small and super long novels of Jin Yong 's novels; in terms of vocabulary richness, the vocabulary diversity of short-length and medium-length novels is higher than that of long-length novels, and the difference between novels



of the same magnitude is small; in terms of the distribution of parts of speech, the internal distribution order of real words and the internal distribution order of function words in Jin Yong's 15 novels are consistent; In the distribution of long and sentence lengths, the specific distribution patterns of the novels are similar.

In addition, the text similarity calculation results show that the short story "Yuen Jian" has the lowest similarity with other Jin Yong novels, and is much lower than the similarity between other novels. In the six short stories, the similarities between the two novels "A Deadly Secret" and "Ode to Gallantry" and the four other novels are relatively low, while the similarities between the other four novels such as " Flying Fox of Snowy Mountain" are relatively high. The similarity has a certain degree of consistency with the creation time. Among the six super-length novels, "Sculpture Trilogy" has a high degree of similarity, while the latest serialization of "The Deer and the Cauldron" and other novels are relatively high. In sentiment analysis, in general, Jin Yong's novels show neutral emotional tendencies.

As far as the falsification research of the suspected work "Wolong Ji" is compared, the statistical results of "Wolong Ji" are compared with Jin Yong's novels, and it is found that it is ranked in the top in terms of word size, word length distribution, glyph-like symbol ratio, vocabulary density and frequency. There are no significant differences between "Wolong Ji" and Jin Yong's novels in terms of language characteristics such as 100 words, part-of-speech distribution, average sentence length, and sentence length distribution. The use of punctuation marks, average word length, word length dispersion, unique word frequency, word frequency distribution, sentence length dispersion, average paragraph length, similarity calculation, and sentiment analysis are significantly different. Among them, the text similarity calculation is a text discrimination method with high credibility. Based on the statistical results and analysis of various parts, it is unlikely that the suspected work "Wolong Ji" is Jin Yong's work.



## 目 录

<b>摘要</b> .....	I
<b>Abstract</b> .....	III
<b>第一章 绪论</b> .....	1
1.1 研究对象及研究意义 .....	1
1.2 研究现状 .....	1
1.2.1 金庸小说语言研究现状 .....	1
1.2.2 基于统计学原理分析语言风格的研究现状 .....	2
1.3 研究方法 .....	6
1.4 论文结构 .....	7
<b>第二章 语料库的建设</b> .....	8
2.1 语料的选取 .....	8
2.2 语料的加工处理 .....	8
2.3 本章小结 .....	11
<b>第三章 字符和词汇层面的语言风格特征及证伪分析</b> .....	12
3.1 字符层面特征及证伪分析 .....	12
3.1.1 标点符号的使用习惯 .....	12
3.1.2 用字量的统计 .....	14
3.2 词汇层面的特征及证伪分析 .....	15
3.2.1 平均词长及词长分布情况 .....	15
3.2.2 词汇丰富度和词汇密度 .....	18
3.2.3 基于词频的分析 .....	23
3.2.4 词类分布情况 .....	29
3.3 本章小结 .....	31
<b>第四章 句子和段落层面的语言风格特征及证伪分析</b> .....	33
4.1 句子层面的特征及证伪分析 .....	33
4.2 段落层面的特征及证伪分析 .....	37
4.3 本章小结 .....	39
<b>第五章 基于文本相似度计算的语言风格特征及证伪分析</b> .....	40
5.1 基于改进 TF-IDF 算法的文本相似度分析 .....	40
5.2 基于 LSI 算法的文本相似度分析 .....	45



---

5.3 基于情感分析的文本相似度分析 .....	49
5.4 本章小结 .....	54
<b>第六章 结语 .....</b>	<b>56</b>
6.1 本文总结 .....	56
6.2 不足与展望 .....	57
<b>参考文献 .....</b>	<b>59</b>
<b>附 录 .....</b>	<b>63</b>
附 1.各标点符号比重 .....	63
附 2.中科院计算所汉语词性标注集 .....	64
附 3.各类实词比重 .....	65
附 4.各类虚词比重 .....	66
附 5.16 部小说中各类情感倾向的句子数与句子总数 .....	66
<b>致 谢 .....</b>	<b>67</b>



## 第一章 绪论

作家作品的语言风格体现在谋篇布局和遣词造句之中。作家在标点符号运用、词汇选择、句式构造等方面都体现出其独有的特色。这些语言结构又都能够通过一定的量化手段来进行统计与分析，从而揭示出数据背后独特的语言风格特征。前人基于统计学的原理与方法对文本语言风格进行量化的研究已取得较为丰富的成果。本章将从研究对象及意义、研究现状、研究方法和文章结构等四个方面来进行介绍。

### 1.1 研究对象及研究意义

金庸作为武侠小说“一代宗师”，其作品除了在情节的织构、人物的塑造、思想的表达等方面造诣颇深之外，其文字语言的运用更是为人所称道。当前对金庸小说的研究角度主要集中在文学、美学、史学、文艺批评、文化研究等方面，语言学方面的研究数量极少，而使用统计学原理及方法分析其语言风格的可查文献更是寥寥无几。

因此，本文将利用语言学、统计学及数学等领域的相关理论与方法来对金庸 15 部小说进行考察和分析，通过数据展现金庸小说语言风格特征。除此之外，加入其疑似作品《卧龙记》的鉴伪研究，以期在具体语言特征的对比之中，更好地揭示出金庸的语言风格特点，并验证前人所提供的探索的相关统计特征与手段的可行性和有效性。

### 1.2 研究现状

#### 1.2.1 金庸小说语言研究现状

金庸小说语言的研究从目前可查的文献来看，尚且较少，且大多集中于以定性的、内省的方式来进行的研究。研究多从文学的角度对金庸小说语言进行风格概括与描写，陈会明（1999）围绕金庸小说语言浅白凝练的特点进行简要的描写说明。魏仪（2003）从四个方面对金庸小说语言特点进行了简单的论述。张继东（2003）在其硕士论文中从意境美、画面感、幽默感和人物语言个性化等四个方面对金庸小说语言进行了定性描写和论述。王开银（2008）对金庸与古龙小说语言风格进行了对比研究，其中举例说明了金庸不同小说表现出的语言风格，并论述了金庸语言风格的流变。赵杰丽（2011）从金庸小说语言的音乐性方面展开论述。也有从语言学的角度进行的探索，李扬、慕俊杰（2007）从金庸小说中的“道（说）”字入手，指



出金庸小说语言继承了宋代以来话本小说语言的一些特点和元素。闫镇南（2010）借助认知语言学中的隐喻理论和转喻理论，对金庸小说中的爱情表达进行了研究，不失为新颖的研究角度。

关于利用计量统计的方法对金庸小说语言进行的研究数量较少，宋沁璐（2005）在其硕士论文中运用语料库语言学和数理统计的研究方法分析了金庸作品的用字和用词特点，虽然她仅对字频和词频进行了统计分析，但对金庸语言的归纳、举例与分析较为细致与充分。刘颖、肖天久（2014，2015）在对金庸和古龙的语言风格进行对比分析时，选取了六部金庸小说，统计分析了其平均段落长度、词长变化程度、高频词的使用频率，以及标点符号、动词、数词、方位词的使用频率等。刘肖二人的文章相较宋文，对于语言特征的统计更为全面，也更为细致，手段也更为新颖和科学，在通过计量方式揭示金庸作品语言风格上走得更远。

### 1.2.2 基于统计学原理分析语言风格的研究现状

利用统计的方法来对语言进行研究，即是通过实验方法和数学模型对语言进行考察，从而得到可以进行验证的结论。（刘颖，2014）有学者将这类研究称为计算风格学或计量文体学（quantitative stylistics）。其中语言风格，又称风格，对其的研究古已有之，但归属现代语言学范畴的语言风格这一术语则始见于瑞士语言学家索绪尔（Ferdinand de Saussure）的学生巴里（Charles Bally）1905年出版的《风格学概说》一书。后经各家沿用，但其定义却始终存在分歧。本文中提到的语言风格是指“在主客观因素制导下运用语言表达手段的诸特点综合表现出来的气氛和格调”。它可以是时代风格、民族风格、个人风格、语体风格、文体风格等。（黎云汉，2000）

使用统计学的方法与手段来对文本进行研究，以达到文本风格定量描写、作品作家身份判别等目的，其实是建立在语言单位的数量分布能够反映文本特点的假说之上。既然需要统计语言单位的数量分布，那么就需要确定哪些语言单位是可以进行量化统计的。可进行量化统计的语言单位主要可归纳为语音层面、字符层面、词汇层面、句子层面、短语层面、段落层面、篇章层面等。对不同层面的语言单位进行统计，要求不同的语料加工程度。语言单位及其数量分布情况，即是文本的计量特征，在相关文献中又有文本结构特征、语言特征、文体特征、识别特征等之称。虽然有学者指出能够用于文本统计的计量特征多达 500 多种，但在实际的研究中，文本测量所使用的计量特征数量却是有限的。（施建军，2016）常用于量化统计的语言特征，如表 1.1 所示：



表 1.1 语言特征表

语言层面	语言特征
语音层面	声母、韵母、句尾韵、声调等
字符层面	字母、汉字、大小写字母、数字、标点符号、词首字母、词尾字母、词首字、词尾字、空格等
词汇层面	平均词长、词长分布、词长离散度、词汇丰富度、高频词、低频词、独现词、句首词、句尾词、段首词、段尾词、各词类比例、特定类词汇比例等
短语层面	N 元语法、各词性短语比例等
句子层面	平均句长、句长分布、句长离散度、特定句式比例等
段落层面	平均段长、段长离散度等

而统计的方法与手段主要分为统计描述、统计推断、统计模型等三个层次。常用于实现描述的统计量、统计推断的方法以及对语言成分或文本之间关系进行推断的数学模型，具体如表 1.2 所示：（刘颖，2014）

表 1.2 统计方法表

统计描述	频率、概率、平均数、方差和标准差、互信息、Z 评分、Dice 系数、Phi 平方系数、对数似然比、N 元语法、信息熵、极限熵、词语的实用度和通用度以及 Yule 图等
统计推断	反映语言现象的数学规律：Zipf 法则、Menzerath-Altmann 定律、Piotrowski-Altmann 定律和 Fuchs 公式；以及假设检验：参数检验（U 检验、t 检验）和非参数检验（卡方检验和 F 检验）以及方差分析等
统计模型	朴素贝叶斯模型、K-最邻近模型、支持向量机模型、层次聚类和划分聚类等

由于本文的目的是利用统计学原理和方法对金庸小说语言风格进行描写和分析，并且对其疑似作品《卧龙记》进行鉴别判断，因此文献梳理将主要包括利用统计学原理和手段对特定作家作品的语言风格进行探索分析和对作家作品进行鉴别等两个方面的相关研究。

对特定作家作品的语言风格进行分析的文献主要分为两个角度，同一作家作品的语言风格分析、不同作家作品语言风格对比。基于统计的方法对同一作家作品的语言风格进行分析，主要是对标点符号、词长、句长、词汇密度、词汇丰富度、规



定频次的词语分析、实词与虚词总情况分析、高低频词分析、特定词语或词类分析等多个语言特征进行统计，来对作家文本进行探究，观察异同，概括规律。吴礼权（2004）从词语选用与修辞类型两个方面来对鲁迅不同风格的两篇文章进行分析，以探究“庄重”与“幽默”两种风格的定量指标。陈玲（2007）对《哈利·波特》系列英文版小说进行了计量统计研究，并利用 kolmogorov-Smirnov 检验、Kruskal-Wallis 检验、方差检验以及主成因素分析等对统计结果进行了检验。张优（2009）与姜晓艳（2016）都使用了语料库检索与统计相结合的方法，分别对《德伯家的苔丝》与《简爱》中主题词与人物、情节、环境、心理之间的关系进行了探究。张小宇（2016）对鲁迅杂文语言风格进行了探究，研究中还对语料库中未被划分的词语进行了对比，同时使用 N 元文法的方法分析了文本词串的出现规律。余韵（2017）选取了多个语言特征对巴金小说创作分期问题进行了探索。此外，还有仅对作家某一作品中某一类特定词汇进行的统计研究，刘旭鹏（2013）对《平凡的世界》中所有的动词重叠式进行了分类与统计研究。

不同作家作品的对比分析，同样是从字符层面、词汇层面、句子层面以及篇章层面选取多个不同的语言特征来进行统计研究。较早对汉语作家文本进行量化统计研究的是钱锋、陈光磊（1983），二人对比分析了巴金和倪海曙的语言风格，但选取的语料样本较小，仅各选取了两位作家的一篇文本进行对比。吴礼权（2003, 2004）从词语选用、句式特点、修辞文本类型三个方面，选取两篇不同风格的不同作家作品，来对文学鉴赏中“简约”与“繁丰”、“平淡”与“绚烂”风格进行定量的描写。王景丹（2003）比较了八位中国剧作家的风格，对句子层面的句类、特定句式等语言特征进行了统计。刘颖、肖天久（2015）利用多个语言特征统计结果比较了金庸与古龙小说之间的风格差异，另外还进行了文本层次聚类与 k-means 聚类。值得注意的是，研究中还选取了两组分别代表家国责任与个人感受的词语进行文本间的分析对比。时季（2017）选取五十多个语言特征对毕飞宇和苏童的小说进行了聚类比较分析。金迪（2018）对格非和余华小说语言风格进行了对比研究，统计了多个语言特征，并对两位作家文本中的 25 个虚词进行秩和检验。除了对汉语作家作品的对比研究，还有对英语作家文本的对比研究。陈岳红（2008）对澳大利亚两位著名作家的小说《莫里斯格斯特》与《乔治一家的妻子》的语言风格特征进行了对比分析。汪鸿雁（2013）在词汇层面统计分析了美国历任总统就职演说的特点。林敏（2014）选取了一些语言特征对比分析了查尔斯·狄更斯的《双城记》与弗吉尼亚·伍尔夫的《达洛维夫人》。陈菲菲（2018）对美国黑人女性文学的代表作《紫色》与《他们眼望上苍》中的多个语言特征进行了统计，还利用了 Biber 所提出的多维分析方法对作品



进行了对比分析。还有许多不同译本的对比研究，包括非汉语文作品的中译本对比研究与汉语文作品其他语言译本的对比研究。文本翻译在一定程度上其实是译者对原文本的再创作，从这个意义上来说，这类研究也可归于不同作家作品的对比分析的范围。刘泽权等（2011）从词汇和句子层面对《红楼梦》的四个英译本进行了对比研究。蒋跃等人（2016）对人工翻译版本的《傲慢与偏见》与百度在线翻译版本中的被动句进行统计，对比异同。韩红建等人（2016）又选取了更多的语言特征对《傲慢与偏见》人机译本之间的异同进行了分析。黄晖（2017）对丰子恺和林文月两个版本的《源氏物语》译本的语言风格进行了比较分析，但选取的语言特征较少。

作家作品鉴别研究包括两个方面，一方面是对佚名作品的作家判断，另一方面是对疑似抄袭作品的判定。对佚名作品的作家判别最早可追溯至 1851 年，英国著名数学家、理论代数奠基人德•摩根（Augustus de Morgan）提出利用平均词长来鉴别《圣经新约》中写给各地主教的 14 封书信是否都出自使徒保罗之手。摩根统计了希罗多德和修昔底德不同卷宗作品的平均词长，进行交叉对比后发现，同一作家作品的平均词长间差距小，而不同作家作品间的平均词长差距较大。由此，他将此计量特征用于判断《新约圣经》中的 14 封书信是否都由同一人所著。另外还有许多著名例子，如统计学家 Yule 对西方经典宗教作品《追随基督》(*de Imitatione Christi*) 作者身份的判别，学者 Wake 对柏拉图《第七封书信》真伪的判断，Mosteller 等人对《联邦党人文集》(*The Federalist Papers*) 中作者归属问题的研究。汉语佚名作家判别最早出现于《红楼梦》的前八十回与后四十回的作家身份判别研究中，后也成为研究成果较为丰硕的方向。除了对《红楼梦》中一百二十回的语言特征进行频率统计和假设检验以外，张运良（2009）、施建军（2011）、肖天久与刘颖（2014,2015）等人还分别在选取不同语言特征的基础上，利用 K-最邻近模型、支持向量机模型、层次聚类等方法来对文本进行判别研究。疑似抄袭作品的判定方面，李惠、刘颖（2013）利用主成分分析和随机森林的方法来对郭敬明与庄羽的小说进行了抄袭判定的研究。吉志薇（2014）利用改进后的 TF-IDF 算法与余弦相似度对郭敬明与庄羽的小说作品进行了相似度计算，从而判定抄袭与否。宋明媚、林丽清（2017）利用卡方检验来对郭敬明与庄羽小说作品中的高频虚词与高频情节关键词进行检验，从而判定作家作品是否存在抄袭行为。在作家作品的鉴别方面，有两个重要的因素，一是语言特征的取舍，二是各种统计模型的使用，两者之间是相互交织的。陈大康（1987）在对《红楼梦》后四十回作者的判别研究中提出了 47 个虚词作为特征项，后施建军（2016）用其中的 44 个虚词对《儒林外史》与《儿女英雄传》进行了聚类



实验，得到了肯定的结果。易勇等（2007）基于空间向量模型与朴素的贝叶斯模型来提取特征并对李白和杜甫的诗词进行作者识别实验。王少康等（2011）利用点积相似度算法和改进的 KL 距离算法来对 10 位现代汉语著名作家作品的语句节奏进行计算，以此作为区别性特征来识别作家作品。陈芯莹等（2012）在统计两个已知作家的文本中多个语言特征的基础上，找出其中差异较大的数据组，并将其作为特征项来对未知作家的文本进行聚类实验，根据实验结果得出七个能够具有区别的语言特征。范亚超等（2018）利用降噪自编码器深度模型提取文本的语言特征，并使用支持向量机模型来对《西游记》中的诗词进行作者识别实验。

综上可知，基于统计的特定作家作品的分析中，同一作家语言风格的分析研究明显少于不同作家作品的对比研究，其中英语作家的作品对比研究相对较多。这一方面是因为利用计量的方法对语言特征进行研究在西方起步早，相关的理论与实践都较为丰富。另一方面则是由于汉语的独特性，在借鉴西方已有的方法与理论时，会出现不匹配的情况，从而使得研究成果的输出较为低效。同时，也由于目前在对中文文本进行预处理方面，特别是分词与词性标注上，还存在许多有待解决的问题，因此也导致后期文本的统计较为低效。目前对于汉语较为行之有效的语言统计特征与方法虽有了许多有益的探索，但还是存在许多问题有待解决。而对于汉语作家作品的判别，国内学者虽在模型的使用与改进上有了许多有益尝试，但用于具体作品的研究数量还是较少。而在用于统计的语言特征方面，缺乏高效且通用的计量特征。同时，不论是特定作家作品分析还是作家作品判别，在具体研究时，方法较为单一，缺乏多角度、多方法的运用与实践。

### 1.3 研究方法

本文基于语料库语言学、统计语言学、计量语言学、计算语言学等相关领域的理论，利用定量分析与定性分析相结合的方法对小说文本中的不同层级的语言特征进行统计、计算与解释。以实证的结果说明与检验金庸小说语言风格之间的异同，并对疑似作品《卧龙记》进行分析与判别。

(1) 基于语料库的统计与计算。本文将在建立有效的语料库的基础之上，利用统计学中描述统计的方法对从字符、词汇、句子到篇章的各层级的语言特征进行描写与分析；除此之外，采用计算语言学的已知可行的算法对文本进行计算与分析，并利用推论统计的相关方法进行检验。基于语料库的统计与计算可以为文本的语言风格提供客观的可检验的数据支持，有利于从不同的角度观察文本的语言风格。

(2) 基于数据结果的描写与对比分析。在利用统计与算法得到相关的数据之



后，对数据结果进行相应的制表绘图，从语言学的角度描写数据情况，并从不同维度对数据进行观察和规律总结。同时，结合定性分析对文本间的差异进行对比，发现异同。基于数据结果的描写与对比分析，有利于在数据的基础上进一步揭示文本语言风格的数据表现，以及文本语言风格之间的一致性与差异性。

## 1.4 论文结构

本文将通过量化的手段在探究金庸 15 部小说及疑似作品《卧龙记》的语言风格特征的基础之上，进一步判别《卧龙记》的真伪。论文共分为六个章节：第一章为绪论，包括研究目标和研究意义、研究现状、研究方法与论文结构三个部分；第二章为语料库的建设，这一章主要介绍研究所需的语料库的建立过程与结果，包括语料的选取与加工；第三章为字符与词汇层面的语言风格特征及证伪分析，分别选取字符与词汇层面具有代表性的可量化特征对金庸 15 部小说及疑似作品《卧龙记》进行统计与分析；第四章为句子与段落层面的语言风格特征及证伪分析，同样也是选取了相关语言特征进行数据统计与分析；第五章为基于文本相似度计算的语言风格特征及证伪分析，分别通过 IF-IDF 与 LSI 算法和余弦相似度相结合来对文本之间的相似度进行计算与分析，另外，还利用情感分析的方法与卡方检验对文本相似度进行具体分析；第六章为结语，总结全文，反思不足。



## 第二章 语料库的建设

在进行统计分析之前，需获取客观真实的数据。建立语料库是对文本进行量化分析的第一步，也是文本语言风格分析的数据来源。根据语料库语言学的相关理论与研究方法，本文将从语料的选取和语料的加工处理两方面来建立金庸的 15 部小说及其疑似作品的语料库，为后文的风格分析提供研究基础。

### 2.1 语料的选取

金庸小说共计 15 部，分为三个版本。最早的版本即连载版，也称作旧版，是报纸上直接刊载的版本。金庸于 1955 年到 1972 年间在报纸上陆续连载完成了其 15 部武侠小说。70 年至 80 年间，金庸对连载版进行了删改和修订，后授权出版。此修订后版本就是所谓“修订版”。1994 年由金庸授权三联书店在中国大陆出版，故也称此版为“三联版”。最后是“新修版”，又称为“世纪新修版”，金庸在 1999 年到 2007 年间修改完成，这一版本在中国大陆由广州出版社和花城出版社出版发行。本文的研究语料来自于百度网盘中一套完整的 PDF 格式的金庸全集，版本为三联出版社 94 年版。

《卧龙记》又名《岳小玉传》，是一部由无名氏所著的武侠小说，曾被署名为金庸之作。中国奥林匹克出版社曾于 1989 年出版发行四册本。通过搜索，在网络文学阅读网站上可找到《岳小玉传》，对比曾出版的《卧龙记》内容，确定两者为同一书。本文的语料电子版来自于网络文学阅读网站刊登的完整版《岳小玉传》。为避免行文混乱，后文中疑似作品名统一为《卧龙记》。

### 2.2 语料的加工处理

在初步获取金庸小说全集及《卧龙记》的电子版后，本文将对这些语料进行进一步的加工和处理，主要分为以下步骤：语料的获取与整理、分词和词性标注、熟语料校对。

#### (一) 语料的获取与整理

在获得初始语料后，还需要对文本内容进行加工处理，以得到最能反映小说语言真实面貌的研究语料。在初步建立的语料中，仅需保留小说的正文内容，其他内容，如每部小说的封面、目录、出版信息、重印说明、改版说明、自序、后记等都被全部删除。因为部分语料是获取的电子版本，我们对电子版的语料还进行了随机



抽样的核验与校对，确保所得语料的准确性。除此之外，在正文中，还需将一部分对于文本分析有干扰或无用的内容进行删除，如小说中的脚注、小说末尾的说明注释，小说的篇章标题和章节序号，阿拉伯数字、外文符号和特殊符号等。本文在语料建设中，删除的内容有：

(1) 小说的章节序号和章节目次。金庸小说中短篇小说一般没有回目，都以“一”、“二”、“三”等数字作为章节的题目，而长篇小说则以类似传统章回小说的对句、短语、诗词等作为章节目次。如，《书剑恩仇录》“第一回 古道腾驹惊白发 危峦快剑识青翎”、《射雕英雄传》“第一回 风雪惊变”、《鹿鼎记》“第一回 纵横钩党清流祸 峭茜风期月旦评”等。这些内容，在语料的整理过程中都进行了删除，以减小在词频和词类等方面分析的影响。

(2) 注释说明的文字。注释性文字一般是对小说中的内容进行补充说明，对小说语言的分析并无直接的作用，因此也在语料中将其删除。如，《书剑恩仇录》中最后的注“注：一、陈家洛之母姓徐名灿，字湘萍，世家之女，能诗词，才华敏瞻，并非如本书中所云为贫家出身。……”

### (二) 分词和词性标注

由于中文的独特性，其在能够进行量化统计之前，首先需要进行分词和词性标注。只有经过分词和词性标注的语料才能够为下一步的统计提供数据支持。本文采用北京理工大学海量语言信息处理与云计算工程研究中心分词软件ICTCLAS来对初步建成的语料库进行分词加工和词性标注。本文采用的词性标注集详见附录1。

### (三) 熟语料校对

尽管分词软件的结果已有较高的准确性，但还是不可避免地存在一些分词和标注上的错误。因此，在对语料进行分词和词性标注的加工后，还需要对其进行人工的筛查与校对。在对结果进行检查的过程中，发现有如下几个在分词与词性标注中常见的问题，先列举如下：

(1) 当分未分。这种情况指的是应该切分的词语并未被切分，并正确标注。示例如下：

错误：身高/n 膀/ng 宽/a, /wd 一/m 脸/q 精悍/a 之/uzhi 色/ng。/wj

更正：身/ng 高/a 膀/ng 宽/a, /wd 一/m 脸/q 精悍/a 之/uzhi 色/ng。/wj

说明：句中“身高”与“膀宽”对举，“高”与“宽”分别是修饰“身”与“膀”的形容词，因此将“身高”划分为表示身体长度的名词并不恰当。

错误：段誉道/nr: /wp “/wyz 很/d 好/a, /wd 你/rr 练/v 罢/y, /wd

更正：段誉/nr 道/v: /wp “/wyz 很/d 好/a, /wd 你/rr 练/v 罢/y, /wd



说明：“段誉”是小说中的人名，而“道”表示说，为动词。因此句子中将“段誉道”划分为人名并不正确。

(2) 当合未合。这种情况指的是不应该被切分的词语却被切分，标注可能有偏差。示例如下：

错误：戚/nr1 长发/n 教/v 给/p 他/rr 和/cc 戚芳/nr 的/ude1 剑/n 法/n, /wd

更正：戚长发/nr 教/v 给/p 他/rr 和/cc 戚芳/nr 的/ude1 剑/n 法/n, /wd

说明：“戚长发”为小说中的汉语人名。句中将“戚”与“长发”切分成两个词，并不恰当。

错误：杨/nr1 铁心/v 见/v 一/m 壶/q 酒/n 已/d 喝/vg 完/vi 了/y, /wd

更正：杨铁心/nr 见/v 一/m 壶/q 酒/n 已/d 喝/vg 完/vi 了/y, /wd

说明：“杨铁心”为小说中的汉语人名。句中将“杨”与“铁心”切分成两个词，并不恰当。

(3) 分合不当。这种情况指的是连续的字符，该分在一起的并未分在一起，不该分在一起的却分作一个词语，从而导致词性标注错误。示例如下：

错误：师祖梅/nr 念笙/nr2 早/ad 瞧/v 出/vf 三/m 个/q 徒/l儿/n 心术不正/al, /wd

更正：师祖/n 梅念笙/nr2 早/ad 瞧/v 出/vf 三/m 个/q 徒/l儿/n 心术不正/al, /wd

说明：“梅念笙”是小说中的人物名。句中将“师祖”与“梅”划为一词，表示人名，并不恰当。

错误：想当年/vl 徽/ng 宗道君/nr 皇帝/n 一心/d 只/d 想/v 长生不老/vl, /wd

更正：想当年/vl 徽宗/n 道君/n 皇帝/n 一心/d 只/d 想/v 长生不老/vl, /wd

说明：“徽宗”为“宋徽宗”的简称，与“道君”一起用于修饰“皇帝”。句中将“宗道君”划为一词，并标为人名，并不正确。

在分词软件的选取上，即便我们采用了由北京理工大学海量语言信息处理与云计算工程研究中心开发的，获得过中国大数据自然语言处理方向第一名的分词软件ICTCLAS，但分词结果仍然不是十分理想。由于语料众多，我们只着重校对和更正了人名和地名上的分词错误。通过以上步骤，最终获得加工后的成熟语料库。各小说的语料库具体情况如表 2.1：



表 2.1 16 部小说语料库情况表

序号	作品	连载的起止时间	总字数	总词数
1	书剑恩仇录	1955.2.8-1956.9.5	431862	317079
2	碧血剑	1956.1.1-1956.12.31	350411	256447
3	射雕英雄传	1957.1.1-1959.5.19	758449	567052
4	雪山飞狐	1959.2.9-1959.6.18	110159	82419
5	神雕侠侣	1959.5.20-1961.7.8	810681	613357
6	飞狐外传	1960-1961	373676	278544
7	倚天屠龙记	1961.7.6-1963.9.2	817568	615020
8	鸳鸯刀	1961	29344	21728
9	白马啸西风	1961	57500	41769
10	天龙八部	1963.9.3-1966.5.27	1021443	761534
11	连城诀	1963	192732	144019
12	侠客行	1966.6.11-1967.4.19	309020	233656
13	笑傲江湖	1967.4.20-1969.10.12	827171	609851
14	鹿鼎记	1969.10.24-1972.9.23	1021369	739590
15	越女剑	1970	13761	10648
16	卧龙记	时间不详	361062	262960

需要说明的是，短篇小说、中篇小说以及长篇小说的划分主要依据文本的总字数，除此之外还与文本的人物多寡、情节繁简等方面密切相关。本文中对金庸已知的 15 部小说进行短篇、中篇以及长篇小说的划分时，主要依据学界已有的观点(孔庆东，2009)，即划分为一部短篇《越女剑》，两部中篇《鸳鸯刀》、《白马啸西风》，六部小长篇《雪山飞狐》、《连城诀》、《侠客行》、《碧血剑》、《飞狐外传》、《书剑恩仇录》，六部超级长篇《射雕英雄传》、《神雕侠侣》、《倚天屠龙记》、《天龙八部》、《笑傲江湖》和《鹿鼎记》。在这里，小长篇和超级长篇在字数上都归属于长篇小说，但在金庸小说中，这两个分类中的小说，在字数上却有数十万的差距，因此将其分为两类较为合理。在下文根据篇幅分类的具体分析中，将以此为标准。

### 2.3 本章小结

本章主要对研究所需的语料库及其加工过程进行了必要的说明，包括了语料的选取和加工两个部分。在语料的加工处理部分对语料的获取与整理、分词和词性标注、熟语料校对等方面作了较为详细的说明。在下一章中，将会以本章建成的语料库作为研究中用于数据统计的样本，对 16 部小说的字符和词汇层面的语言风格进行分析。



## 第三章 字符和词汇层面的语言风格特征及证伪分析

字符与词汇层面的语言特征能够揭示出作家作品的语言风格。字符与词汇层面具有丰富的统计特征，本章将在字符层面选取标点符号与用字量两个计量特征来进行统计与分析，而在词汇层面，则将从平均词长与词长分布情况、词汇密度与词汇丰富度、词频分布情况、各词类使用情况等方面来对小说语言风格进行统计与分析。

### 3.1 字符层面特征及证伪分析

在汉语作品的定量分析研究中，字符层面的特征分析一般选取的是汉字和标点符号两个重要特征。对于汉字而言，主要是通过统计不同作品中使用的不同汉字总量来对作品进行研究分析。而标点符号的统计主要集中在各种标点符号的使用数量及标点符号之间的间隔分布情况上。本节将分别从文本的标点符号使用情况及用字量两方面来对作品风格进行分析。

#### 3.1.1 标点符号的使用习惯

在现代汉语中，标点符号的使用虽然受到语法等方面因素的限制，却也存在着极大的自由发挥的空间。标点符号的选择和使用能够在一定程度上体现出作家个人的风格特色。也就是说，通过对标点符号使用习惯的统计分析来对作家风格进行区别。标点符号的使用习惯主要表现在各标点符号不同的使用频率上。

本文对金庸 15 部作品及疑似作品《卧龙记》中各标点符号出现的频率进行了统计。统计的标点符号包括：顿号、逗号、冒号、双引号、单引号、破折号、分号、句号、感叹号、问号、省略号等。统计结果详见附录 2。将结果绘制为折线图，如图 3.1：



图 3.1 标点符号使用情况折线图

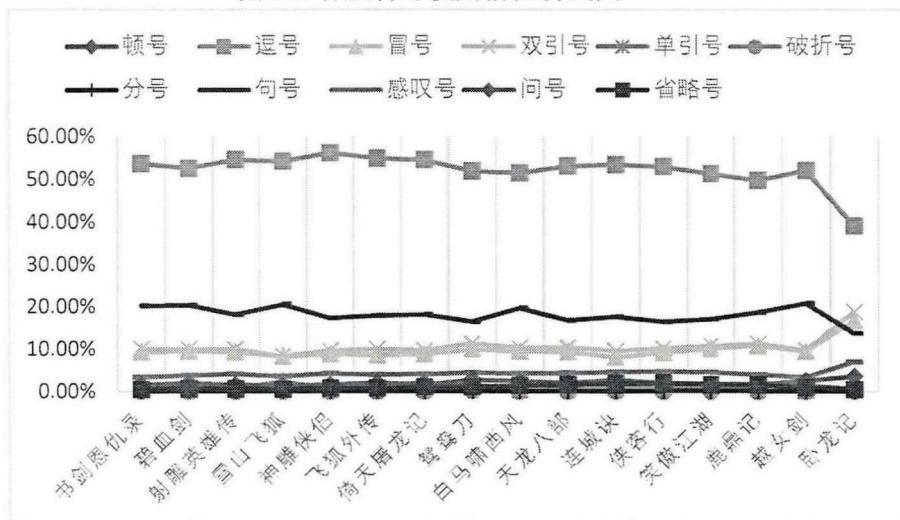


图 3.1 中的横坐标是各部小说名称，纵坐标则是在各部小说中各标点符号在总标点符号数中所占百分比。由图可知，金庸小说中各标点符号的占比基本遵循：逗号>句号>冒号与引号>其他标点符号。在金庸的 15 部小说及疑似作品《卧龙记》中，逗号和句号使用频率远高于其他标点符号。逗号与句号一般对应文本中的陈述句类，与小说叙事属性相符。另外，文本中冒号与双引号则多对应人物之间的对话场景，16 部小说中人物的对话明显占有重要位置。

与疑似作品《卧龙记》相比，金庸的 15 部小说中，同一标点符号的百分比在作品之间的差异并不明显。《卧龙记》中逗号和句号的使用频率却明显低于其他小说，而问号和感叹号的比例则又相对有所升高。与此同时，《卧龙记》中的冒号和双引号使用频率也远高于其他小说。

冒号与双引号的使用数量在一定程度上反映了作品中人物对话的场景和数量。《卧龙记》全文约四万余字，与同等量级的《碧血剑》相比，其冒号和双引号的使用数是后者的一倍之多。在《卧龙记》中人物间对话众多，且你来我往，短小简练，多在同一时空发生，人物并无太多的心理和行为的刻画和描写，从《卧龙记》中选取一段人物对话，展示如下：

“万万不可！”邝火怒道：“你若没本领治好郭大堡主的掌伤，大可速离此地，若要毒杀于他，邝某绝不答允！”

公孙咳声道：“邝庄主何出此言？不才几时说过要毒死郭堡主了？”

邝火道：“你不是说要用毒药喂给郭堡主吗？”

公孙咳道：“是呀！但这毒药尽管可以毒死咱们这里每一个人，但郭堡主喝了下去，却反而只会有益无害！”



邝火陡地呆住，道：“这又是什么道理？”

公孙咳瞪了他一眼，道：“你若中了血花莲掌力，也不会给这种毒药毒死！”

邝火奇道：“这又是什么道理？”

岳小玉也瞪了他一眼，道：“你还不明白吗？这自然是以毒攻毒，两种毒性相生相克的结果了。”

公孙咳哈哈大笑，道：“聪明！聪明！一点就透！”（《卧龙记》）

由于对话多而短小，且人物间的互动往来又充满张力，由上例可看出，感叹号和问号出现频繁。此类示例在文中比比皆是，这也一定程度上解释了《卧龙记》中感叹号和问号使用频率高于其他小说的原因。

综上，在小说各标点符号的使用情况中，金庸的 15 部小说之间具有较高的一致性，而疑似作品《卧龙记》虽然基本符合金庸小说标点符号使用占比基本顺序，但标点符号的实际占比情况却与金庸小说有着较为明显的差异。

### 3.1.2 用字量的统计

用字量指作家在一定量的作品中所使用的不同汉字数量的比例，即不同用字数与总字数之比。用字量能在一定程度上体现出作家在作品中汉字使用的丰富度。表 3.1 是金庸作品及其疑似作品的用字量统计情况：

表 3.1 16 部小说用字情况表

项目 作品	不同用字数	总字数	用字量
越女剑	1425	13761	10.36%
鸳鸯刀	1910	29344	6.51%
白马啸西风	2117	57500	3.68%
雪山飞狐	2683	110159	2.44%
连城诀	2963	192732	1.54%
侠客行	3224	309020	1.04%
卧龙记	3322	361062	0.92%
飞狐外传	3394	373676	0.91%
碧血剑	3583	350411	1.02%
书剑恩仇录	3663	431862	0.85%
笑傲江湖	3786	827171	0.46%
神雕侠侣	3863	810681	0.48%
倚天屠龙记	3907	817568	0.48%
射雕英雄传	4013	758449	0.53%
鹿鼎记	4071	1021369	0.40%
天龙八部	4085	1021443	0.40%



由表 3.1 可知，总字数不同的作品，其不同用字数也存在差异，总体上呈现出逐渐上升的趋势。而就用字量而言，随着小说总字数的增加，用字量则不断下降。六部小长篇内部与六部超级长篇内部，其用字量相对稳定。从《笑傲江湖》到《书剑恩仇录》之间，从《雪山飞狐》到《白马啸西风》之间，都有着明显较大的涨幅。同时，三部中短篇小说之间的涨幅最为显著。用字量的明显增长处基本对应了金庸小说中短篇、小长篇与超级长篇之间的临界。

疑似作品《卧龙记》的用字量约为 0.92%，与金庸同等量级的作品之间并无明显差异，也符合上文所述的金庸小说用字量基本规律。用字量在一定程度上能够反映出金庸作品中对汉字使用的总体情况，同时也可以总结出一定的规律，即用字量随着小说的增加呈现有规律的下降，小长篇之间用字量差异不大，超级长篇之间的用字量也较为相似。但并不排除上表中所展现的总字数与用字量之间的规律也适用于其他作家作品。因此，以用字量作为鉴别疑似作品的特征量，还需进一步的探究。

### 3.2 词汇层面的特征及证伪分析

词是砌成文本的基石，是构成语言特征的重要角色。利用统计的方法对文本中的词汇使用情况进行量化分析，可以揭示出不同文本之间的语言风格差异。前文可知，词汇层面可用于测量的特征繁多，本文主要从平均词长和词长分布情况、词汇密度和词汇丰富度、词频分布情况、各词类使用情况等几个具有代表性的计量特征来对小说语言风格进行统计分析。

#### 3.2.1 平均词长及词长分布情况

##### (一) 平均词长与词长离散度

文本中总字数除以总词数所得值即为平均词长。而词长离散度，指的是文本中所有词语的长度偏离平均词长的程度。统计学中，离散度指的是对不同数值之间的差异性的测量。标准差则是常用的用于反映数据之间差异的变异性量数之一。本文即利用标准差公式来对词长离散度进行计算。公式如下所示：

$$S = \sqrt{\frac{\sum(x_i - \bar{x})^2}{n}} \quad (\text{其中 } i=1,2,3,\dots,N) \quad (\text{公式 1})$$

在公式 1 中， $S$  为词长离散度， $n$  为文本中的词语总数， $\bar{x}$  则表示平均词长。由此可计算本文中所涉及的 16 部小说的词长变化程度。金庸 15 部小说及疑似作品《卧龙记》的平均词长及词长离散度的统计结果如表 3.2 所示：



表 3.2 平均词长与词长离散度统计表

作品\项目	平均词长	词长离散度
越女剑	1.5192	0.583
神雕侠侣	1.5488	0.583
倚天屠龙记	1.5541	0.558
侠客行	1.5590	0.557
雪山飞狐	1.5640	0.544
飞狐外传	1.5718	0.560
射雕英雄传	1.5722	0.552
连城诀	1.5802	0.579
天龙八部	1.5843	0.582
鸳鸯刀	1.6001	0.558
笑傲江湖	1.6035	0.553
书剑恩仇录	1.6040	0.550
碧血剑	1.6076	0.580
白马啸西风	1.6139	0.599
鹿鼎记	1.6465	0.517
卧龙记	1.6559	0.595

由表 3.2 可知，金庸 15 部小说的平均词长范围约在 1.5192-1.6465 之间，但小说之间的差异并无明显规律。其中，《越女剑》的平均词长最短，《鹿鼎记》的则最长。而《卧龙记》的平均词长为 1.6559，略超出于金庸小说平均词长的最高值。

平均词长用于反映作家作品的语言风格最早由英国著名数学家、理论代数奠基人德·摩根 (Augustus de Morgan) 提出，后在作家作品的特征分析中被多次使用，有学者指出平均词长体现出了文本的可读性，平均词长的长短反映出了文本中词语的长短，而词语的长短则很大程度上决定了阅读文本的难易程度。但也有学者认为，由于语言的特殊性，中文词语中多为单音节词和双音节词，在音节文字中能够发挥区别文本作用的平均词长，在中文中并没有显著的区别效果。(施建军, 2016) 由上表 3.2 中的统计数据可看出，实际的统计结果也反映出了这一指标存在的粗糙性和局限性。

词长离散度越大，表明作者所使用的词语在长度上的变化越大，反映出文本在词语长短变化上的多样性和灵活性。由上表可对比金庸在各部小说中用词的灵活程度。金庸小说的词长离散度区间约在 0.517-0.599 之间，其中《白马啸西风》的词长变化最大，《鹿鼎记》则最小。《卧龙记》的词长离散度为 0.595，落在金庸小说词长



离散度区间。但与其同等量级的金庸小说相比，即与《碧血剑》等六部小长篇相比，其词长变化程度最大，用词更为灵活多变。从这个角度上来说，《卧龙记》的词长离散度与金庸小说存在差异。

除此之外，观察表 3.2 会发现，词语长度变化情况与文本的长短并无直接关系，如，同为短篇小说的《越女剑》和《鸳鸯刀》，在词长离散度上存在着较大的差距，而同为超级长篇的《鹿鼎记》和《天龙八部》之间也存在差异。另外，词长离散度与创作时间也并无明显关联，如，创作时间虽相近，但《越女剑》与《鹿鼎记》其词长离散度则分别居于两端。因此，除了可以确定金庸小说离散度区间之外，其小说的词长离散度并无其他显著规律。

## （二）词长分布情况

词长分布情况指的是文本中不同音节词语出现的情况。作家在进行文本创作时，对不同长度词语的使用有着个人的取舍和独特的偏好，因此，统计文本中的不同长度词语的使用情况，可以作为揭示作家作品语言风格的一种尝试。

本文将对金庸的 15 部作品及其疑似作品中的单音节词、双音节词、三音节词、四音节词及四音节以上词的使用情况进行统计，并列出其频率，具体情况如表 3.3：

表 3.3 词长分布情况表

项目 作品	单音节词频 率	双音节词频 率	三音节词频 率	四音节词频 率	四音节以上 词频率
书剑恩仇录	68.46%	27.60%	3.21%	0.72%	0.01%
碧血剑	67.89%	28.36%	2.98%	0.75%	0.01%
射雕英雄传	69.80%	27.46%	1.97%	0.76%	0.01%
雪山飞狐	69.90%	27.32%	2.03%	0.74%	0.01%
神雕侠侣	70.90%	26.84%	1.46%	0.79%	0.01%
飞狐外传	69.46%	27.70%	2.05%	0.77%	0.01%
倚天屠龙记	70.41%	27.06%	1.73%	0.79%	0.01%
鸳鸯刀	69.32%	27.14%	2.63%	0.90%	0.00%
白马啸西风	66.74%	29.62%	2.91%	0.73%	0.01%
天龙八部	69.25%	28.28%	1.58%	0.88%	0.01%
连城诀	69.43%	28.09%	1.68%	0.78%	0.01%
侠客行	70.98%	26.69%	1.44%	0.88%	0.01%
笑傲江湖	68.65%	28.13%	2.16%	1.05%	0.01%
鹿鼎记	66.99%	28.84%	3.27%	0.88%	0.02%
越女剑	73.06%	25.31%	0.94%	0.69%	0.01%
卧龙记	67.56%	28.52%	2.97%	0.93%	0.02%



在金庸 15 部小说及其疑似作品中，单音节词的使用频率大约在 65%-75% 之间，几乎占整部作品的三分之二，双音节词则在 25%-30% 之间，三音节及其以上的词语仅仅占到词语总数的 5% 左右。总的来说，金庸小说的词长分布情况是，随着音节数的增加，词语数量会大幅减少。

由表 3.3 可知，不论是单双音节词的使用频率，还是三音节以上的词语，金庸的 15 部作品之间都有着差异，小到 0.03%，大到 6%，中间的跨度较大，并没有具体的规律可循，如果以此为区间，很难有效地鉴别出伪作与原作。

最早使用词长分布这一特征来揭示作家作品风格的学者曾对莎士比亚、狄更斯、福迪等多位作家的作品进行词长分布情况分析。这一方法后被引入用于汉语作家作品的分析中。但有学者指出，汉语的词长一般在五个音节以内，而英语词长则通常在十三个字符内，汉语较之后者少了近三分之二的观测值，因此词长使用频率在汉语文本风格特征的分析中所能发挥的作用十分有限。（施建军，2016）由上文的观察可知，词长分布用于作家作品的分析和判别确实存在一定的局限性。

### 3.2.2 词汇丰富度和词汇密度

#### （一）词汇丰富度

词汇的丰富度指的是在文本中词语种类的多样性。在一定量的文本中，不同的词语出现得越多，也就意味着文本的词汇越丰富，同时也说明相同词语的重复情况越少。本文将以类形符比和独现词比例这两个统计量来说明文本的语言风格特征。

类形符比（Type Token Ratio, TTR），最早由 Johnson 提出，类符（Type）指的是文本中所有不同的词语，形符（Token）则是指总的词语数。具体公式如下：

$$TTR = \frac{Types}{Tokens} \quad (公式 2)$$

然而由于这一公式受到文本长度的影响较大，当文本长度增加时，其使用新词的频次会大大下降。因此，当文本长度不一致时，计算所得的类形符比不能较为准确地反映出文本之间的差异。于是有学者对这一公式进行了改进，有对样本获取方法和计算程序进行改进的，如 Laufer 和 Waller 提出的 Equal TTR、Malvern 和 Richards 的 Split TTR；（金迪，2018）有对公式直接进行改动的，如 Guiraud 修改的  $TTR = \frac{Types}{\sqrt{Tokens}}$ ，Carroll 的  $TTR = \frac{Types}{\sqrt{2}Tokens}$ ，以及 Herdan 的  $TTR = \frac{\log Types}{\log Tokens}$  等。（陆芸，2012）

本文将会同时使用公式 1 与 Herdan 改进的对数类形符比，即公式 3（如下）来对 16 部小说中的词汇使用情况进行统计，结果如图 3.2 所示：



$$TTR = \frac{\log Types}{\log Tokens} \quad (\text{公式 3})$$

图 3.2 类字符符比折线图

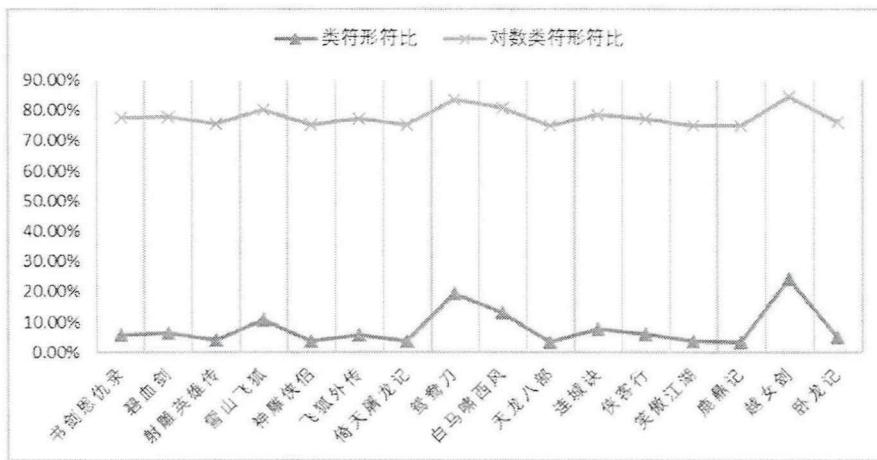


图 3.2 中横轴为小说名，纵轴为各小说的类字符符比。由图 3.2 可知，使用公式 2 所得类字符符比中，《鸳鸯刀》和《越女剑》处有明显的波峰，其类字符符比分别为 0.19 和 0.24，远高于其他作品。它们分别是一部短篇小说和中篇小说。相对而言，在使用公式 3 所得的对数类字符符比中，这两部小说的波峰明显较前者趋缓，小说之间因文本长度造成的差异也极大缩小。由此可见，小说的长度确实对类字符符比的计算有着影响，而使用改进后的对数类字符符比，的确可在一定程度上弥补这种不足。

在金庸的 15 部小说中，中短篇小说的词汇多样性高于长篇小说，但同等量级的小说之间的差异较小。疑似作品《卧龙记》与同等量级的金庸作品《侠客行》和《飞狐外传》等之间词汇丰富度也具有较高的相似性，三者的对数类字符符比分别为 0.76、0.77、0.77。

独现词指的是文本中仅出现一次的词语，也即是频次为 1 的词语。在文本中，单频次的词语所占比例越大，说明其所用词语的重复率越低，词汇也就愈加丰富和多样。与类字符符比一样，独现词频也受到文本长度的影响，为提高有效性，独现词频也采用对数计算（金迪，2018），公式如下：

$$I = \frac{\log \sum Number(H)}{\log \sum Number(W)} \times 100\% \quad (\text{公式 4})$$

其中，H 指文本中频次为 1 的词语数量，W 为文本中总词语数。16 部小说的统计结果如表 3.4 所示：



表 3.4 独现词数与独现词频统计表

作品	项目	独现词	总词数	独现词频
鹿鼎记	20803	738830	73.58%	
笑傲江湖	17580	609856	73.38%	
天龙八部	22515	761211	74.00%	
倚天屠龙记	20547	614808	74.50%	
神雕侠侣	20964	613137	74.67%	
射雕英雄传	20768	566910	75.04%	
卧龙记	9661	262839	73.53%	
侠客行	11136	233134	75.39%	
飞狐外传	13583	278486	75.91%	
书剑恩仇录	15656	316984	76.25%	
碧血剑	13561	256349	76.40%	
连城诀	8585	143945	76.26%	
雪山飞狐	6892	82408	78.08%	
白马啸西风	3909	41762	77.74%	
鸳鸯刀	3101	21710	80.51%	
越女剑	1975	10645	81.83%	

小说中的独现词多与小说的主题相关，一般为对人物、场景、动作、故事背景等的刻画与描写。选取 16 部小说中的独现词举例如下：

《鹿鼎记》：切菜刀、砧板、庞然大物、温顺、秦失其鹿、秦朝、灶头、蔺相如、秦王、问鼎中原、金、定王、专指、余姚人、连袂、杭州府、隐士、羊膏、绝唱、削发为僧等；

《笑傲江湖》：灭门、和风熏柳、马勒、泼喇喇、黄汤、兀鹰、獐子、尽兴、兔肉、正经事、野味、箭法、酒炉、头束、脸儿、叶落归根、家乡话、痘、爷们、挥金如土等；

《天龙八部》：绑架、吃闲饭、四书五经、推己及人、父母官、公断、老百姓、气冲冲、搓揉、明净、血瘤、措辞、药囊、深藏不露、孟述圣、搪塞、微言大义等；

《倚天屠龙记》：歉意、井口、井水、心浮气躁、井栏圈、山层、屏崖、风烟飘渺、烦俗、灰衣、守戒、捆绑、仁善、出言无状、粗眉大眼、不可收拾、外孙女儿等；

《神雕侠侣》：爱宠、撒娇、男仆、好人家、客房、血印、难逃一死、藉藉无名、凌霄花、踏坏、花架、银桂树、生人、不管三七二十一、不徐不疾、出人意外、怒叱、佛等；



《射雕英雄传》：斜阳、男男女女、残瓦、枉死鬼、芳魂、逃之夭夭、散场、说唱、黄酒、豆腐干、却之不恭、筷儿、舖、搜括、吹牛拍马、李邦彦、嫖院、能征惯战等；

《侠客行》：滴溜溜、刺眼、鼻涕、策马、郑重其事、崩坏、白虹贯日、寥寥无几、苦功、分毫、收帐、耐烦、指不胜屈、破天荒、慈悲、不知轻重、震骇、丝毫无损等；

《飞狐外传》：刀枪剑戟、神定气闲、后叉步、撩掌、抱虎归山、大痛、后襟、真功夫、轰隆隆、猥亵、垫步、探马、推掌、稳重、徐大爷、吃不了兜着走、得意之作等；

《书剑恩仇录》：无垠、海波、心胸、披襟、袒居、圆柔、亲题、亭台楼阁、大石碑、观海、鞍、金堤、渚、欢心、勋业、美言、碧堂、赤栏、曲桥、天香坞等；

《碧血剑》：嘉靖、内乱、卒、心疾、妄言、绞杀、披荆斩棘、胜览、云渤海国、唐人、沧海、山盘地、佛教、船、遗风、屡试不第、经营、天资聪颖、观光、风物、打点等；

《连城诀》：南郊、瓦屋、晒谷场、旱烟袋、嘉许、神光、炯然、凛凛、娇嗔、作斜、劈势、败局、赖皮、天花、考究、抱拳、称道、高足、问卜、布囊、薄礼、赏面等；

《雪山飞狐》：脚程、铁蹄、邪门、微胖、髭须、貂皮、鞭梢、箭杆、抄接、茫茫、追踪、高头大马、威武、唿哨、忽喇喇、淡黄、鞍头、青绸、千里迢迢等；

《白马啸西风》：哀求、寒冰、蜷曲、长枪、回旋、浑似、陡然间、雁翎、并举、劈刺、屈服、手刃、心如刀割、祷祝、保佑、转瞬间、结义兄弟、神刀、关西、瘦瘦长长等；

《鸳鸯剑》：劲装、并肩而立、山寨、剪径、小贼、唯恐、大模大样、打量、尖削、先考、墓、衣衫褴褛、烟雾、深湛、劲敌、轶闻、婆婆、首级、倏然、晋北、大同等；

《越女剑》：秋水、悍勇、窥伺、狠命、鼓噪、狠辣、蚕食、甚或、绰绰有余、毙命、血渍、固不待言、齐整、名扬天下、试探、好手、钦服、轻飘飘、陡然间、薄雾等；

《卧龙记》：旅雁、高山、龙门、繁荣、售卖、花灯、粥、水泄不通、大白天、墟、盛大、节日、风调雨顺、胡天胡地、蹦跳、有头有脸、市井之徒、热腾腾、肉条、箸、面条、老顽固等；

为了更清晰地呈现各小说独现词频率的情况，绘制柱状图 3.3 如下：



图 3.3 独现词频柱状图

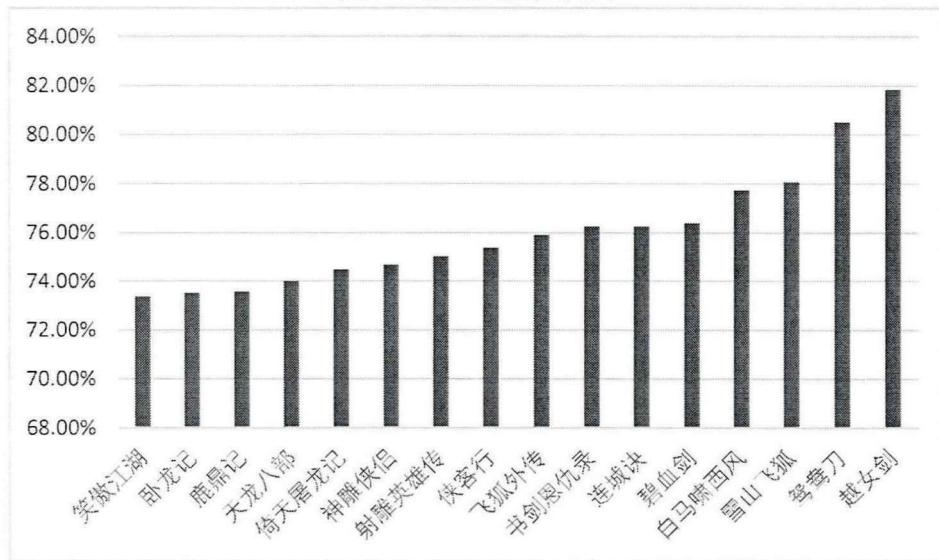


图 3.3 中横轴为各部小说名，竖轴为小说的独现词频率。独现词的频率也能反映文本中词语使用的多样性和丰富性。由图 3.3 可知，金庸的小说中，独现词频基本遵循中短篇小说大于小长篇，小长篇大于六部超级长篇的规律，即总词数越多，独现词频越低。这在一定程度上说明了小说总词数对小说用词丰富度情况的影响。

那么，比较同等量级的金庸作品会发现：在《越女剑》等三部中短篇小说中，总词数最多的《白马啸西风》的独现词频率远低于《越女剑》与《鸳鸯刀》，说明其在用词上的多样性远低于后两者；而《越女剑》的词汇丰富度远高于金庸其他小说。在六部小长篇中，《雪山飞狐》的独现词频明显较高，其他五部小说之间的差异则较小，其中《侠客行》略低。六部超级长篇之间的独现词频差异也较小，其中《射雕英雄传》最高，《笑傲江湖》与《鹿鼎记》较低。

小说《卧龙记》的独现词频约为 73.53%，词汇丰富程度较低。虽然其独现词频落在金庸小说独现词频的区间内，但其频率却与长篇小说相近，而与其他同等量级的小说差异则较为明显，并不符合前文所述的金庸小说独现词频的规律。

## （二）词汇密度

词汇密度指的是文本中出现的实词与总词数之比。词汇密度用于衡量文本中的信息量，而文本信息的主要承载者便是实词。因此，统计实词在文本中的比例，可以反映出作家作品中的信息密度。词汇密度越大，文本所承载的信息密度越大。同时，词汇密度并不受文本长度的影响，具有一定的稳定性。然而，汉语中的实词与虚词的划分，一直以来都存在着不同的观点。由于划分标准和目的的不同，导致虚



实划分的结果也不尽相同。因此，在统计时必须选取一种划分标准与结果。本文采取《现代汉语》(邢福义、王国胜主编, 2011) 中对实词和虚词的划分，统计的实词为名词、动词、形容词、代词、数词、量词等六类。统计结果如表 3.5 所示：

表 3.5 实词总数与词汇密度统计表

项目 作品	实词总数	总词数	词汇密度
白马啸西风	26697	41762	63.93%
卧龙记	170678	262839	64.94%
连城诀	94450	143945	65.62%
侠客行	154556	233134	66.29%
笑傲江湖	405470	609856	66.49%
天龙八部	508355	761211	66.78%
神雕侠侣	411772	613137	67.16%
雪山飞狐	55825	82408	67.74%
鹿鼎记	500597	738830	67.76%
飞狐外传	188843	278486	67.81%
倚天屠龙记	418666	614808	68.10%
射雕英雄传	387131	566910	68.29%
碧血剑	175762	256349	68.56%
鸳鸯刀	14976	21710	68.98%
书剑恩仇录	223488	316984	70.50%
越女剑	7579	10645	71.20%

根据表 3.5 可见，金庸小说词汇密度在 63.93%-71.20% 之间，最高与最低之间的差幅约为 7.27%，其中《越女剑》的词汇密度最大，《白马啸西风》则最小。各小说的词汇密度与小说的篇幅、创作的时间也无可观测的直接联系，呈现出较为无序的状态。虽然《卧龙记》的词汇密度在 16 部小说中处于较低的位置，但与其他小说之间的差距，并不比已知的金庸小说作品之间的差距大；同时，与其同等量级的《连城诀》也具有相似性。

### 3.2.3 基于词频的分析

#### (一) 词频分布

词语的使用频次及各频次的分布情况能够反映出作者独特的语言风格特征。除了上文中所提到的关于词汇丰富度的量化指标以外，各频次词语的概率分布也可用于考察文本词汇的丰富性。相对低频词语的出现概率越高，文本的词汇就越丰富多样。Yule 图是由统计学家尤尔 (G.Yule) 提出的用于统计作家风格特征的方法，主



要是通过统计高频到低频词语的概率分布情况，来观察小说词汇的使用面貌。不同频次的词出现的概率计算公式如下：

$$p(i) = \frac{n_i \times i}{N} \quad (\text{公式 5})$$

其中*i*表示词语出现频次，*n<sub>i</sub>*表示出现频次为*i*的词语数量，*N* 表示文本中的总词数。(刘颖, 2014)通过计算不同频次的词语出现的概率，可以画出相应的曲线图，即 Yule 图。文本将金庸 15 部小说分为短篇小说、中篇小说、小长篇以及超级长篇四类，来与疑似作品《卧龙记》的词频分布进行比较。金庸小说及疑似作品的 Yule 图如图 3.4 所示：

图 3.4 词语各频次的概率分布图

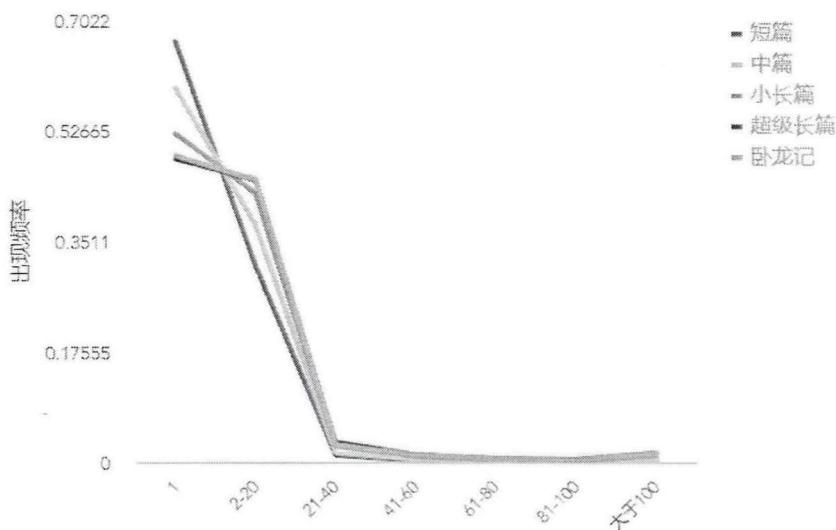


图 3.4 中横轴为词语频次，纵轴为各频次词语所对应的出现频率。从上图的结果来看，金庸小说词汇丰富度情况为：短篇小说>中篇小说>小长篇>超级长篇。金庸的短篇小说、中篇小说、小长篇与超级长篇的词频分布之间存在明显的差异，而《卧龙记》的词频分布与超级长篇小说的平均值之间存在一定的相似性。

假设检验 (hypothesis testing) 是对差异性进行检验和推断的过程，其基本依据是归谬法和概率原理。假设检验的过程一般是假设事件  $H_0$  成立，然后通过数据计算，如果计算结果导致了小概率事件发生，则证明该假设不成立，反之，则证明假设成立。假设又分为原假设与备择假设，原假设一般为预测对比值无差异的假设，备择假设则为预测对比值有差异的预测。在统计学中，小概率事件一般为发生概率小于或等于 0.05 的事件。(刘颖, 2014)



卡方检验 (Chi-Squared Test) 是最常用的非参数假设检验之一。卡方检验有  $2 \times 2$  卡方检验、 $2 \times C$  卡方检验、 $R \times C$  卡方检验、配对卡方检验、分层卡方检验等具体的数据检验分析方法。其中， $R \times C$  卡方检验可用于检验各组数据之间是否存在差异。它必须满足三项假设，即假设 1，存在两个无序多分类变量；假设 2，具有相互独立的观测值；假设 3，样本量足够大，最小的样本量要求为分析中的任一期望频数大于 5。其检验的具体过程为：

(1) 提出假设：

原假设：x 与 y 没有明显差异；

备择假设：x 与 y 有明显差异。

(2) 计算皮尔逊卡方统计值  $\chi^2_0$ 。

(3) 选取显著水平  $\alpha$ ，根据计算所得的皮尔逊统计值  $\chi^2_0$  和自由度  $df$ ，进而得到相应的 P 值。当  $P < \alpha$ ，则拒绝原假设，反之则承认原假设。

基于本文中的数据情况，采用  $R \times C$  卡方检验来对短篇小说、中篇小说、长篇小说和疑似作品《卧龙记》进行检验，并提出原假设：它们之间没有明显的差异。卡方检验的结果如表 3.6 所示：

表 3.6 卡方检验表

卡方检验

	值	自由度	渐进显著性 (双侧)
皮尔逊卡方	1879.786 <sup>a</sup>	24	.000
似然比	1980.477	24	.000
线性关联	1367.743	1	.000
有效个案数	423312		

a. 0 个单元格 (0.0%) 的期望计数小于 5。最小期望计数为 10.34。

由表 3.6 的检验结果可知，皮尔逊卡方值为 1879.786，自由度为 24， $P$  值  $< 0.05$ ，则原假设不成立，《卧龙记》与金庸的短篇小说、中篇小说、小长篇小说的词频分布有比较明显的不同。

虽然  $R \times C$  卡方检验能够观察到各组数据之间是否存在差异，但各组之间具体的差异情况，需要使用 Post hoc testing 检验来进一步考察。Post hoc testing 检验，即根据调整后的标准化残差 (adjusted standardized residuals) 判断各组的差异。利用 Post hoc testing 检验可以得到词频的范围与作品交叉表情况，具体如表 3.7 所示：



表 3.7 词频的范围与作品交叉表

范围	计数	作品					总计
		短篇	中篇	小长篇	超级长篇	卧龙记	
1	计数	1975	7010	69415	123177	9661	211238
	期望计数	1473.6	5879.4	66399.1	127584.5	9901.4	211238.0
	占作品的百分比	66.9%	59.5%	52.2%	48.2%	48.7%	49.9%
2-20	调整后残差	18.5	21.1	20.0	-27.7	-3.5	
	计数	923	4441	56859	114593	8911	185727
	期望计数	1295.6	5169.3	58380.2	112176.3	8705.6	185727.0
21-40	占作品的百分比	31.3%	37.7%	42.7%	44.8%	44.9%	43.9%
	调整后残差	-13.9	-13.7	-10.1	15.3	3.0	
	计数	33	180	3394	8221	552	12380
41-60	期望计数	86.4	344.6	3891.4	7477.3	580.3	12380.0
	占作品的百分比	1.1%	1.5%	2.6%	3.2%	2.8%	2.9%
	调整后残差	-5.8	-9.1	-9.8	13.9	-1.2	
61-80	计数	7	51	1157	3047	233	4495
	期望计数	31.4	125.1	1412.9	2714.9	210.7	4495.0
	占作品的百分比	0.2%	0.4%	0.9%	1.2%	1.2%	1.1%
81-100	调整后残差	-4.4	-6.8	-8.3	10.2	1.6	
	计数	4	27	610	1585	105	2331
	期望计数	16.3	64.9	732.7	1407.9	109.3	2331.0
大于100	占作品的百分比	0.1%	0.2%	0.5%	0.6%	0.5%	0.6%
	调整后残差	-3.1	-4.8	-5.5	7.5	-4	
	计数	3	17	346	1040	76	1482
总计	期望计数	10.3	41.2	465.8	895.1	69.5	1482.0
	占作品的百分比	0.1%	0.1%	0.3%	0.4%	0.4%	0.4%
	调整后残差	-2.3	-3.8	-6.7	7.7	.8	
	计数	8	56	1280	4011	304	5659
	期望计数	39.5	157.5	1778.8	3417.9	265.3	5659.0
	占作品的百分比	0.3%	0.5%	1.0%	1.6%	1.5%	1.3%
	调整后残差	-5.1	-8.3	-14.4	16.2	2.5	
	计数	2953	11782	133061	255674	19842	423312
	期望计数	2953.0	11782.0	133061.0	255674.0	19842.0	423312.0
	占作品的百分比	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%

观察表 3.7 可知，《卧龙记》与金庸的超级长篇小说在多数词频范围内存在趋势的一致性，但是在部分词频范围内仍然存在显著性差异。

以单现词的占比为例，《卧龙记》的调整后残差为-3.5，金庸超级长篇小说的调整后残差为-27.2，《卧龙记》的单现词占比略高于金庸的长篇小说，趋势相近。相对于金庸的短篇小说、中篇小说和小长篇小说的 18.5、21.1、20.0 而言，前者（《卧龙记》与金庸长篇小说）的单现词比例远小于后者（金庸的短篇小说、中篇小说和小长篇小说）。再比如，考虑词频在 21-40 范围内的词，《卧龙记》调整后残差为-0.4，表示《卧龙记》的词频在 21-40 范围内的词与期望没有显著性差异，而金庸的短篇小说、中篇小说和小长篇小说的调整后残差值分别为-5.8、-9.1、-9.8，三者的值都



为负且绝对值都大于 3，表明金庸的短篇小说、中篇小说和小长篇小说中词频在 21-40 范围内词的实际比例远低于其期望，相对于《卧龙记》，金庸的短篇小说、中篇小说和小长篇小说中范围在 21-40 范围内的词的比例显著性地偏小。而金庸的超级长篇小说中词频在 21-40 范围内词的调整后残差为 13.9，值为正且绝对值大于 3，表明金庸的超级长篇小说在词频在 21-40 范围内的词所占的比例远高于《卧龙记》。

通过图 3.4 和表 3.7 可知，《卧龙记》在各词频范围内的频率与金庸的超级长篇小说具有一定的相似性，与金庸的短篇小说、中篇小说和小长篇小说的相似性极低。但是，考虑到《卧龙记》与金庸超级长篇小说在字数上有巨大差异，二者在篇幅上完全不属于一个量级，综合考量，在各词频分布这个量化指标上，金庸是《卧龙记》的作者的可能性极低。

## （二）高频词

高低频词语的统计与分析有助于揭示作家的内容表达。高频词是文本中出现次数最多的一部分词语，即频次最高的部分词语，与文本内容密切相关。也就是说，不同文本之间的内容差异可以在一定程度上从高频词的考察中反映出来。但目前对于高频词的定义与界限尚未有统一的结论。本节中所讨论的高频词指的是，将文本中的词语以其出现频次为依据，从高到低进行排序，位序在前一百的词语。

在金庸 15 部小说的前一百高频词中，共同出现的词语有 39 个，分别是：了、的、他、道、是、我、你、在、人、这、一、也、那、又、得、不、来、去、说、有、她、都、着、要、上、已、到、叫、说道、之、却、给、个、大、但、只、过、声、对。而在包括《卧龙记》在内的 16 部小说的前一百高频词中，共同出现的词语有 38 个。《卧龙记》中缺少的词语为：对。

前一百的高频词中，各部小说的单独出现的词语数量不尽相同，具体情况如下：

《碧血剑》共 4 个：袁承志、青青、何铁手、金蛇；

《神雕侠侣》共 7 个：杨过、小龙女、李莫愁、法王、蒙古、陆无双、此时；

《射雕英雄传》共 8 个：洪七公、欧阳锋、黄药师、丘处机、功夫、欧阳克、后、梅超风；

《倚天屠龙记》共 8 个：张无忌、张翠山、谢逊、赵敏、周芷若、少林、明教、殷素素；

《飞狐外传》共 9 个：程灵素、袁紫衣、马春花、赵半山、福康安、胡斐、道、掌门人、凤天南

《连城诀》共 11 个：狄云、戚芳、万震山、丁典、僧、万圭、花铁干、血刀、水笙、吴坎、血刀老祖；



《笑傲江湖》共 11 个：令狐冲、岳不群、剑法、盈盈、林平之、岳灵珊、恒山、田伯光、师妹、仪琳、嵩山；

《天龙八部》共 12 个：虚竹、萧峰、段誉、乔峰、王语嫣、慕容复、阿朱、段正淳、木婉清、段、阿紫、无；

《鹿鼎记》共 13 个：韦小宝、皇上、康熙、皇帝、太后、公主、吴三桂、小、倒、侍卫、双儿、不过、问道；

《雪山飞狐》共 14 个：曹云奇、苗若兰、胡一刀、金面佛、阮士中、宝树、陶子安、刘元鹤、田青文、竟、苗大侠、范帮主、赛总管、陶百岁；

《侠客行》共 15 个：石破天、雪山、白万剑、丁珰、帮主、石清、丁不四、谢烟客、帮、史婆婆、夫妇、内力、贝海石、少年、长乐；

《书剑恩仇录》共 18 个：陈家洛、张召重、徐天宏、霍青桐、文泰来、余鱼同、乾隆、李沅芷、周绮、陆菲青、周仲英、红花、香香公主、骆冰、清兵、不敢、如、忙；

《白马啸西风》共 19 个：李文秀、老人、苏普、阿曼、苏鲁克、计、汉人、陈达海、车尔库、强盗、迷宫、白马、拉齐里、毒针、瓦耳、华辉、两个、跟、哈萨克人；

《鸳鸯刀》共 22 个：萧中慧、袁冠南、卓天雄、周威信、林玉龙、任飞燕、瞎子、盖一鸣、萧半和、江湖、刀法、夫妻、镖师、头、双刀、逍遙子、喝道、伸手、言、宝刀、太岳四侠、书生、鸳鸯刀；

《卧龙记》共 31 个：岳小玉、公孙、可以、铁、这个、许不醉、酒、吗、诸葛、神秘、老鼠、挂、还是、狂风、常、珠、布、忽然、看、所、以、醉、练惊虹、一定、现在、老、怎样、虽然、咳、就算、婆婆。

《越女剑》共 34 个：范蠡、剑士、这、青衣、勾践、锦衫、吴国、吴、西施、剑士、风、薛烛、伍子胥、竹棒、越国、文种、卫士、师兄、大夫、吴士、小人、大王、名、宝剑、吴王、铸剑、范大夫、越、剑术、戳、王者、身子、利剑、四；

理论上而言，在高频词中，各个文本中同时出现的词语数量越多，一定程度上意味着文本在内容表达上的相似度也越高，而相反的文本中单独出现的词语数量越多，说明文本间的差异越大。由上述内容可知，金庸小说各文本间共同出现的词共 39 个，占比 39%，与之相比，《卧龙记》仅有一词的微小差异。同时，各文本中共现的词多为常用词语，总体而言，各文本间的内容相似度较低。从各小说中并未交叉出现的词语方面来看，各小说中的这类词语多为小说中的主要人物名，除此之外还有地名、兵器名、国名、民族、称谓、武功等等，极少数的虚词和其他实词。由



此可见，前一百的高频词中，小说中单独出现的词语与小说的主题高度相关，仅可通过其一窥小说内容的人物、地名、相关情节等叙事方向，尚难以以此为区别文本语言风格特征的依据。

### 3.2.4 词类分布情况

词类是词的语法分类，是词在语法结构中表现出来的类别。不同词类在文本中的使用频率，是构成文本语言风格的重要特征之一。以往的研究中表明，实词的使用情况在不同的作家作品以及不同语域中具有差异，比如名词在学术语域中使用频率要高于小说与口语语域，而在小说与口语中使用更多的是动词，这与不同语域其擅长的和强调的方向不同有关。（刘颖，2014）另一方面，虚词，在语言风格的量化研究中，一般被看作是最能体现作家语言风格的特征。因为这一类词较少承载实在的语义信息，与文本的主题无过多的关联，而与作家的写作习惯则有着密切关系。因此，统计金庸小说及其疑似作品中各词类的使用频率，有助于更好地揭示金庸小说的语言风格。本文将分为实词与虚词两部分，分别进行统计和分析。

#### （一）实词

如前文所述，本文采用《现代汉语》（邢福义、王国胜主编，2011）中对实词和虚词的划分，统计的实词包括名词、动词、形容词、代词、数词、量词等六类。各实词频率是指各类实词数占总词数之比，具体情况见附录 2。将各类实词的占比制作成折线图后，可以更加直观地看到各类实词的占比以及它们在文本之间的差异。详见下图 3.5：

图 3.5 各类实词比重折线图

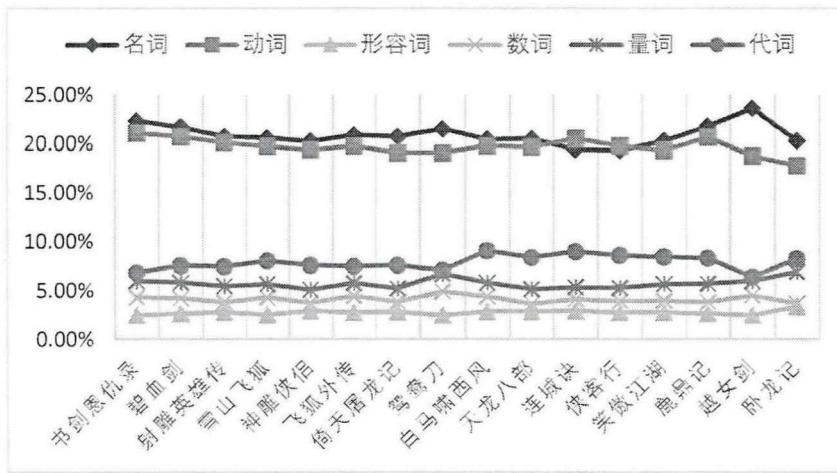


图 3.5 中的横坐标是各部小说名称，纵坐标则是在各部小说中各类实词在总词



数中所占百分比。由图 3.5 可知, 六类实词在金庸小说中的比重次序为: 名词>动词>代词>量词>数词>形容词, 其中名词与动词占有较大比重, 远超其他实词。《连城诀》中动词占比稍高于名词, 而《天龙八部》与《侠客行》中动词与名词的占比差异微小; 《鸳鸯刀》和《越女剑》的量词与代词占比几乎相同。疑似作品《卧龙记》的六类实词占比的高低顺序与其他 15 部小说一致, 但其中动词占比明显低于其他 15 部小说, 形容词和数词的占比差异较小。下文分别从《书剑恩仇录》与《卧龙记》中选取了一段挥剑交锋的打斗场景:

霍青桐侧身一让, 不顾来敌, 挥剑又割断布带一端。哪知敌人剑法迅捷, 不容她缓手去拾包袱, 又是一剑拦腰削来。霍青桐无法避让, 挥剑挡格, 双剑相交, 火花迸发。她心中一震, 敌人武功不弱, 顾不得仔细琢磨, 伸左手又去拾那包袱。敌人长剑如影随形, 直刺她左腕。霍青桐左手一缩, 食中两指捏了剑诀, 右手剑直递出去, 抬头看时, 接连三番阻她拾包袱之人是个美貌少年, 认出就是昨日途中无礼呆看那人, 不禁心头火起, 刷刷刷三剑都是进手招数, 两人斗在一起。(《书剑恩仇录》)

这人一身衣衫洁白如雪, 飞掠下来的姿势更是美妙异常, 祁紫天一见之下, 脸色不禁大变, 再也不等待下去, 手中长剑倏地“嗤”的一声, 就向铁老鼠胸前刺去。

他外号称为“厉剑追魂”, 这时候一剑刺出, 使的便是杀手招数, 一时间只见剑影森森, 走势矫疾无伦, 铁老鼠非要急速闪躲不可。

铁老鼠在兵刃上的造诣, 也许不如祁紫天, 但他擅长轻功, 身法自是灵捷无比, 一见长剑急刺过来, 身形已立刻向上飞跃几逾一丈。(《卧龙记》)

从上面的文段中可以看到, 《书剑恩仇录》中的打斗, 一方步步紧逼, 一方借势打势, 情势在千钧一发之间变化万千。这种紧张感和激烈感得益于密集的动词运用, 如“侧身一让”、“挥”、“割”、“缓手”、“削来”、“避让”、“挡格”、“相交”、“迸发”、“直刺”、“直递”等等动词和动词短语, 让人目不暇接, 不禁屏息凝神, 随着人物的交锋而心惊肉跳。反观《卧龙记》, 同样是打斗场景的呈现, 动词的使用较为简炼, 多有留白, 情势就显得较为缓和, 让人有喘息之地。

## (二) 虚词

《现代汉语》(邢福义、王国胜主编, 2011) 中虚词包括: 副词、介词、连词、助词、拟音词等五类。各虚词频率是指各类虚词数占总词数之比, 具体统计情况见附录 3。将图表数据制作成折线图后, 将更直观地呈现出各类虚词的占比以及在文本之间的差异。详见下图 3.6:



图 3.6 各类虚词比重折线图

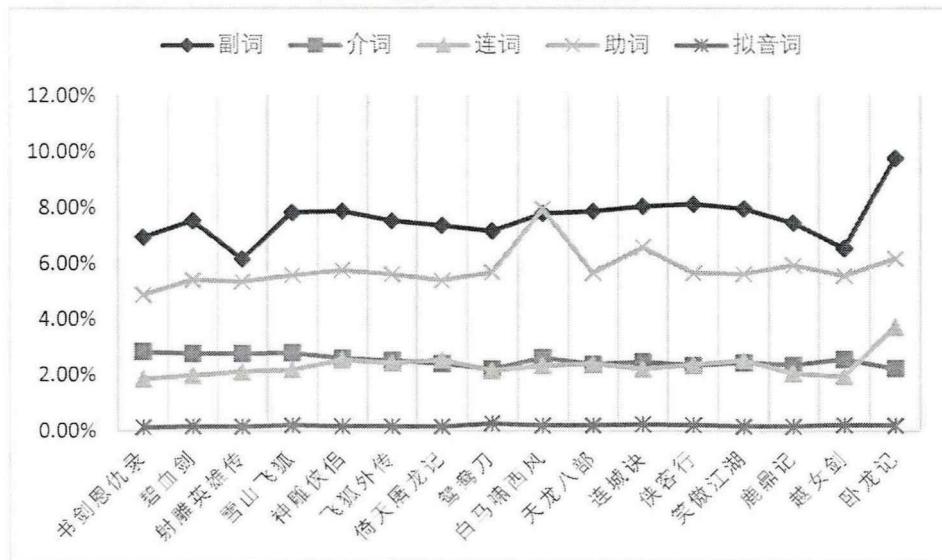


图 3.6 中的横坐标是各部小说名称，纵坐标则是在各部小说中各类虚词在总词数中所占百分比。由图 3.6 可知，在金庸小说中，五类虚词占比的基本情况为：副词>助词>介词>连词>拟音词。其中，《白马啸西风》的助词与副词的占比相似，助词的占比远高于金庸其他小说作品。另外，除《书剑恩仇录》等五部作品介词占比稍高于连词占比外，金庸其他小说中的介词占比与连词占比非常相似。而疑似作品《卧龙记》的各类虚词占比基本符合金庸小说虚词占比的高低顺序，但在连词与介词的占比上，与金庸小说有着明显的区别。在同一词类的比较中，金庸小说《射雕英雄传》与《越女剑》中副词的占比相对低于其他小说，而疑似作品《卧龙记》的副词比重则远高于金庸 15 部小说；除此之外，《卧龙记》的连词占比也明显高于其他金庸作品。

### 3.3 本章小结

本章从字符和词汇两个层面对 16 部小说进行了统计分析，其中字符层面从标点符号的使用和用字量两个方面进行了统计，词汇层面则选取了平均词长、词长离散度、词长分布、类符形符比、独现词频率、词汇密度、各词频出现频率、前一百频次的词语以及各词类的分布情况等方面进行了统计。

在标点符号的使用上，金庸的 15 部小说之间具有较高的一致性，疑似作品《卧龙记》与金庸作品间有着较为明显的差异。用字量方面，金庸小说中呈现出随着小说总字数的增加，用字量则不断下降的规律，同时，总字数在一定区间内的作品文



本，其相应的用字量也相对稳定。疑似作品《卧龙记》符合金庸小说的用字量基本规律。在词汇层面，金庸小说的平均词长、词长离散度、词汇密度并无明显的规律；小说的词汇丰富度方面，金庸中短篇小说的词汇多样性高于长篇小说，但同等量级的小说之间的差异较小；就词类分布情况来看，金庸小说中实词分布基本遵循的顺序是：名词>动词>代词>量词>数词>形容词；虚词基本分布顺序是：副词>助词>介词>>连词>拟音词。

疑似作品《卧龙记》与金庸小说在平均词长、词长离散度上存在明显差异；就词长分布、类符形符比、词汇密度的情况而言，疑似作品《卧龙记》与金庸小说之间并没有显著的差异；独现词频率方面，疑似作品《卧龙记》并不符合金庸小说独现词频的规律；在词频分布情况的综合分析中可知，《卧龙记》与金庸小说存在较低的相似性；各小说频次排名前一百的词语来说，疑似作品《卧龙记》与金庸小说共现的词语基本重合；在各词类的分布中，除了动词、副词、连词的具体占比与金庸小说的具体情况有所差异之外，无其他明显区别。综上，字符与词汇层面语言特征的分析还不足以判别《卧龙记》的真伪。在下章中，将进一步从句子和段落层面来对16部小说的语言风格特征进行分析。



## 第四章 句子和段落层面的语言风格特征及证伪分析

句子与段落在文本语言特征分析中占据重要地位。其中，关于句子长度的计算与分析最早可追溯至 1888 年。(施建军, 2016)句长的分析常用于对不同作家作品、不同语域文本等领域的对比分析，分析主要围绕句子长度的平均数、中位数和标准差，各长度句子在文本中的分布情况等方面展开研究。段落分析则主要包括平均段落长度和段落长度变化程度等方面。本章将从平均句长、句长离散度、句长分布以及平均段落长度来对文本语言风格进行研究分析。

### 4.1 句子层面的特征及证伪分析

对于句子的考察，主要包括平均句长、句长离散度及句长分布的基本分析和对句类与特定句式的考察。句子的长度信息能够反映文本语言风格。句子长度指的是一个句子中所包含的汉字数。本文在进行统计时，将一个句子定义为以标点符号“。”、“！”、“？”为结束符号的字段。本节将从平均句长、句长离散度及句长分布来考察小说中句长基本情况。

#### (一) 平均句长与句长离散度

平均句长是指文本的总字数除以总句数所得之商。句长离散度指的是文本中所有句子的长度偏离平均句长的程度。句长的离散度与前文中词长离散度一样，使用标准差公式来进行计算，公式不变，在此不再赘述。金庸 15 部小说及疑似作品《卧龙记》的平均句长与句长离散度计算结果如表 4.1 所示：



表 4.1 平均句长与句长离散度统计表

作品	项目	总字数	总句数	平均句长	句长离散度
越女剑		13761	578	23.81	13.86
鸳鸯刀		29344	1125	26.08	17.64
白马啸西风		57500	2316	24.83	15.64
雪山飞狐		110159	4319	25.51	15.70
连城诀		192732	7684	25.08	16.40
侠客行		309020	11135	27.75	19.02
碧血剑		350411	14348	24.42	16.20
卧龙记		361062	15093	23.92	15.01
飞狐外传		373676	13672	27.33	17.55
书剑恩仇录		431862	17302	24.96	16.36
射雕英雄传		758449	28308	26.79	18.40
神雕侠侣		810681	28948	28.00	18.67
倚天屠龙记		817568	29161	28.04	18.21
笑傲江湖		827171	31127	26.57	17.73
鹿鼎记		1021369	41493	24.62	16.65
天龙八部		1021443	37654	27.13	18.20

在文本中，句子的长度越长一般意味着句子的结构越复杂，所承载的语义越丰富。从表 4.1 中可观察到，金庸小说的平均句长范围在 23.81-28.04 之间，跨幅较大。从作品篇幅长短角度来看，中短篇小说《越女剑》、《鸳鸯刀》、《白马啸西风》的平均句长范围为 23.81-26.08，差异明显。其中《越女剑》平均句长最短，而《鸳鸯刀》最长。出现这样的差异，与故事的主题和作家选择的行文风格有关。《越女剑》讲述春秋时期，越女阿青助越王成功复仇的故事，行文爽快简炼，武戏干净利落，文戏点到为止，有“笔法纯熟，举重若轻”之誉（孔庆东，2009）；而《鸳鸯刀》则是围绕“鸳鸯宝刀”的秘密展开的喜剧故事，与前者相比在细节描写、人物刻画方面笔触更为细腻繁密：

范蠡走上几步，接过了金漆长匣，只觉轻飘飘地，匣中有如无物，当下打开了匣盖。旁边众人没见到匣中装有何物，却见范蠡的脸上陡然间罩上了一层青色薄雾，都是“哦”的一声，甚感惊讶。当真是剑气映面，发眉俱碧。（《越女剑》）

他一生经历过不少大风大浪，风头出过，钉板滚过，英雄充过，狗熊做过，砍过别人的脑袋，就差自己的脑袋没给人砍下来过，算得是见多识广的老江湖了，但从未像这一次走镖那样又惊又喜，心神不宁。如果护送宝刀平安抵京，刘大人曾亲口许下重赏，自然是“君子一言，



快马一鞭”，说不定皇上一喜欢，竟然赏下一官半职，从此光宗耀祖，飞黄腾达，周大镖头变成了周大老爷周大人。（《鸳鸯刀》）

小长篇《雪山飞狐》等平均句长则在 24.42-27.75 之间，其中，《侠客行》与《飞狐外传》远高于其他同等量级的作品。而超级长篇《射雕英雄传》等六部作品的平均句长约在 24.62-28.04 之间，除《鹿鼎记》外，其他作品之间的差异较小。从数据可知，《鹿鼎记》的篇幅在金庸众作品中排行前列，但平均句长却仅有 24.63，这实际上与《鹿鼎记》一书的语言风格密切相关，其文语言平实有力，嬉笑怒骂，雅俗共赏，句子长短相间，并不一味追求繁复的长句：

韦小宝要强好胜，吹牛道：“我骑过好几十次马，怎么不会骑？”从马背上跳了下来，走到另一匹马左侧，一抬右足，踏入了马镫，脚上使劲，翻身上了马背。不料上马须得先以左足踏镫，他以右足上镫，这一上马背，竟是脸孔朝着马屁股。（《鹿鼎记》）

就作家的写作时间而言，作家在同一时期的作品，其平均句长并不尽然处于稳定状态，创作时间存在交集的作品，有的平均句长之间差异小，如 1955 年至 1956 年之间创作的《书剑恩仇录》和《碧血剑》，1969 年至 1972 年间的《鹿鼎记》与《越女剑》；而有的平均句长之间差异较大，如 1961 年至 1963 年间创作的《倚天屠龙记》、《鸳鸯刀》与《白马啸西风》。

综上所述，文本的平均句长与文本长度和作家创作时间之间并无显著的因果联系，而与小说的主题和作家的风格选择之间关系较为密切。疑似作品《卧龙记》的平均句长虽落在金庸小说平均句长区间之内，但金庸小说内部的平均句长并无明显规律可循，难以直接以这一量化指标来判断其真伪。

句长离散度越大，意味着文本中句子变化越多样，句子结构越丰富。由表 4.1 可知，金庸小说的句长离散度介于 13.86-19.02 之间，金庸小说中，“射雕三部曲”之间的句长离散度具有非常高的相似性。文本长短变化最为灵活的是《侠客行》，而《越女剑》的句子风格则更为统一。疑似作品《卧龙记》的离散度也落在金庸小说句长离散度区间之内，但与《连城诀》等同等量级的作品相比，其句长离散度最小，而与短篇小说《越女剑》则较为相似。

## （二）句长分布

句长分布是指不同长度的句子在文本中所占的比例。本文以 10 字为组距，统计 16 部小说中的句长分布情况。由于随着句中字数的不断增加，句子的数量不断减少，占比微小，计算结果并不利于数据观察，因此在统计时，大于 60 字的句子归为一组，不再往下细分。具体统计数据如表 4.2 所示：



表 4.2 句长分布情况统计表

项目 作品	1-10 个 字	11-20 字	21-30 字	31-40 字	41-50 字	51-60 字	>60 字
书剑恩仇录	13.29%	35.04%	23.35%	13.52%	7.08%	3.63%	4.10%
碧血剑	16.00%	34.17%	23.59%	12.56%	6.31%	3.69%	3.68%
射雕英雄传	14.59%	31.27%	22.52%	13.58%	7.86%	4.44%	5.74%
雪山飞狐	13.38%	32.32%	24.87%	15.24%	6.92%	3.87%	3.40%
神雕侠侣	13.17%	29.14%	22.92%	14.62%	8.78%	5.03%	6.33%
飞狐外传	12.62%	30.00%	23.83%	14.74%	8.54%	5.21%	5.05%
倚天屠龙记	11.93%	29.55%	23.60%	15.07%	9.03%	4.94%	5.88%
鸳鸯刀	14.67%	33.07%	22.22%	12.89%	6.49%	4.89%	5.78%
白马啸西风	15.50%	31.35%	25.39%	13.73%	7.34%	3.45%	3.24%
天龙八部	14.52%	29.66%	22.84%	14.27%	8.26%	4.79%	5.66%
连城诀	16.93%	31.35%	22.64%	13.48%	7.51%	4.05%	4.03%
侠客行	14.28%	29.44%	22.07%	14.23%	8.68%	5.16%	6.15%
笑傲江湖	14.42%	30.56%	23.53%	13.95%	8.15%	4.42%	4.97%
鹿鼎记	16.27%	33.52%	23.44%	12.83%	6.74%	3.34%	3.86%
越女剑	13.15%	34.60%	25.78%	16.09%	6.23%	2.42%	1.73%
卧龙记	15.08%	33.23%	25.78%	14.31%	6.26%	2.76%	2.59%

通过对上表 4.2 的观察可知，金庸小说中的句子长度分布集中于 11-20 字与 21-30 字两个区间，约占小说各类句长的二分之一。其中 11-20 字的句子长度在金庸的各小说中占据首要位置。除此之外，1-10 字与 31-40 字的句长在小说的占比不相上下，当句长超过 31 个字时，随着句子长度的增加，其在文本中的比例也在不断下降。金庸小说句长分布的总体趋势，可通过下面的折线图 4.1 得到更为清晰的展示：

图 4.1 句长分布情况折线图

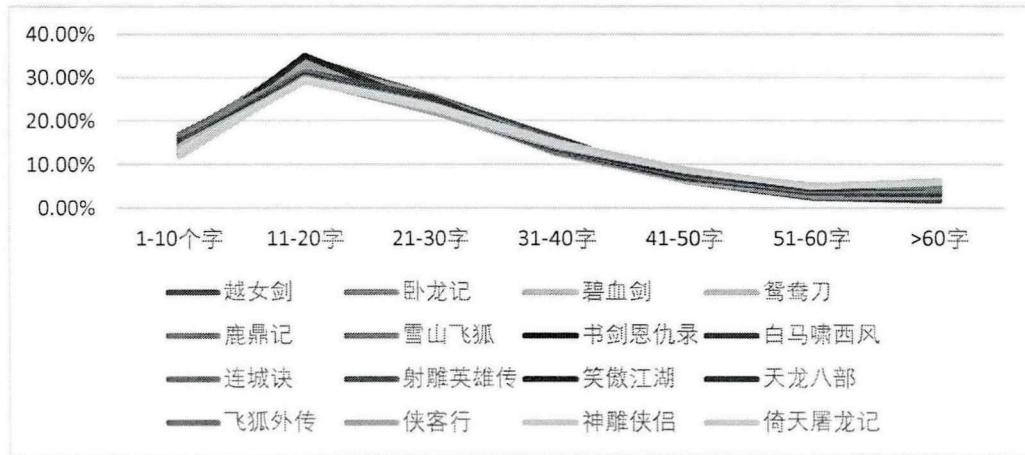




图 4.1 中横轴为句子长度区间，纵轴为各句子长度区间的出现频率。由图 4.1 可知，虽然各小说在不同句长区间的占比存在差异，但彼此之间差距较小。而对于整个句长分布的趋势来说，15 部小说间存在很大的相似性。将疑似作品《卧龙记》与金庸小说相比，其句长分布情况基本上符合统计所得的句长分布规律图式。

## 4.2 段落层面的特征及证伪分析

段落的长短取决于作家的写作习惯与偏好。统计段落的相关情况有助于更好地把握文本的整体气质。段落的平均长度是常用于揭示文本段落特征的指标之一。段落平均长度为文本的总字数除以总段落数所得之商。16 部小说的平均段落长度具体如表 4.3 所示：

表 4.3 平均段落长度统计表

项目 作品	总字数	段落数	平均段落长度
卧龙记	361062	13390	26.97
越女剑	13761	193	71.30
连城诀	192732	2373	81.22
侠客行	309020	3548	87.10
鹿鼎记	1021369	11078	92.20
碧血剑	350411	3787	92.53
天龙八部	1021443	10847	94.17
飞狐外传	373676	3867	96.63
笑傲江湖	827171	8512	97.18
倚天屠龙记	817568	8395	97.39
雪山飞狐	110159	1122	98.18
白马啸西风	57500	584	98.46
射雕英雄传	758449	6904	109.86
神雕侠侣	810681	6872	117.97
书剑恩仇录	431862	3564	121.17
鸳鸯刀	29344	204	143.84

表 4.3 中的平均段落长度可用折线图来表示，如图 4.2 所示：



图 4.2 16 部小说平均段落长度折线图

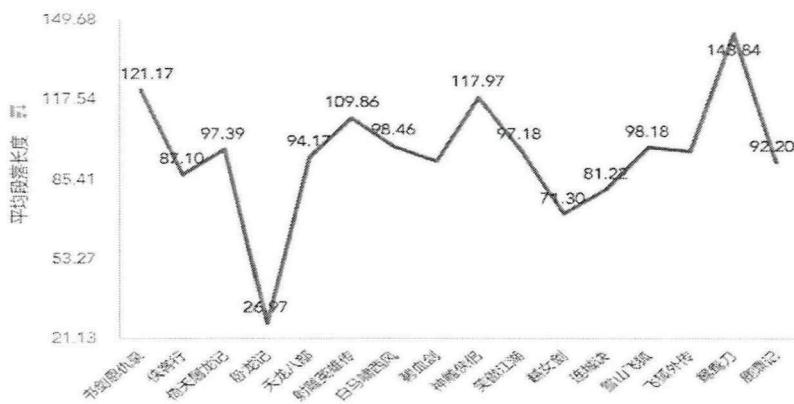


图 4.2 中横轴为小说名，纵轴为各小说的平均段落长度。一般而言，平均段落的长度越长，意味着文本整体上多由长段落构成。观察表 4.3 和图 4.2 可知，金庸小说的平均段落长度约在 71.30-143.84 之间，并无明显规律可循。其中平均段落最长的为《鸳鸯刀》，最短为《越女剑》。而疑似作品《卧龙记》的平均段落长度与其他金庸小说作品差异巨大，仅为 26.97。对比文本会发现，《卧龙记》中，人物对话往来多是独立成段，而在金庸的已知作品中，人物对话与心理活动多是在一个自然段中完成，如下所示：

盖一鸣道：“不错，我们太岳四侠，是江湖上铁铮铮的好汉，决不能难为妇孺之辈。你只须留下坐骑，我们不碰你一根毫毛。想我八步赶蟾、赛专诸、踏雪无痕……”那少女伸手掩住双耳，忙道：“别说，别说。你们不知道我是谁，我也不知道你们是谁，是不是？”盖一鸣奇道：“是啊！不知道那便如何？”那少女微笑道：“咱们既然互不相识，若有得罪，爹爹便不能怪我。呔！好大胆的毛贼，四个儿一齐上吧！”（《鸳鸯刀》）

公孙咳道：“一千几百两，绝对不会有问题。”

岳小玉道：“你有把握可以成功？”

公孙咳道：“只要没有差池，咱们就一定可以功成身退，然后一起去大快朵颐可也！”

岳小玉大是兴奋，道：“好，就照这么办。”

公孙咳望了他一眼，道：“你骑在我的肩膀上，咱们先潜进镖局里再说！”（《卧龙记》）

《卧龙记》中的段落多为上述例子所示。综上可知，《卧龙记》平均段落长度与金庸其他作品的平均段落长度之间存在明显差异。



### 4.3 本章小结

本章主要从句子和段落两个层面对 16 部小说进行了统计，句子层面主要包括平均句长、句长离散度、句长分布情况等三个方面，段落层面则是从平均段落长度方面来进行分析。

在句子层面，金庸小说的平均句长、句长离散度并没有明显可循的规律，其平均句长、句长离散度与文本长度和作家创作时间之间并无显著的因果联系，而与小说的主题和作家的风格选择之间关系较为密切。疑似作品《卧龙记》的平均句长与金庸小说之间并无显著差别，但句长离散度与金庸作品存在差异。就句长分布而言，金庸各小说在不同句长区间的占比存在差异，但彼此之间差距较小。而对于整个句长分布的趋势来说，15 部小说间存在很大的相似性。疑似作品《卧龙记》句长分布情况基本上符合金庸小说的句长分布规律图式。在段落层面，疑似作品《卧龙记》的平均段落长度远远短于金庸小说的平均段落长度。

综上可知，仅从句子层面难以判别《卧龙记》的作者是否为金庸，而在段落层面，则可以观察到《卧龙记》与金庸其他已知小说的显著差异。下章将进一步从计算文本相似度的角度对 16 部小说的语言风格进行分析。



## 第五章 基于文本相似度计算的语言风格特征及证伪分析

文本的相似度计算，是指利用统计学原理与数学模型来对文本中语言特征进行选取、赋值、计算，并最终通过具体的计算结果来判断文本之间的相似性。本章将基于改进 TF-IDF 算法与基于 LSI 算法来对文本的特征项进行选取与赋值，利用余弦公式计算文本相似度，并进行详细分析。除此之外，本章还将基于情感分析理论对小说文本的情感倾向进行统计，并利用假设检验对文本之间的相似度进行检验与分析。

### 5.1 基于改进 TF-IDF 算法的文本相似度分析

向量空间模型（Vector Space Model, VSM）是常用的一种文本分类方法。向量空间模型的基本思想是将文本看作是由若干相互独立的特征项组成的集合，根据重要程度，赋予特征项不同的权重，由此构建出一个由特征项为横坐标，其权重为对应坐标的 N 维向量空间模型。TF-IDF 是常用的基于向量空间模型的特征项权重计算方法。其中 TF (term frequency) 是指词频，IDF(inverse document frequency)指逆文本频率，两者之积就是文本特征项的权重值。由于传统的 TF-IDF 的公式中 IDF 部分的计算忽视了特征项在某一类别和不同类别中的具体分布情况，因此，张保富（2011）等人提出了一种改进的 TF-IDF 计算公式，这一公式结合了特征项的类内和类间信息分布熵，有效提高了计算结果。改进后的 TF-IDF 计算公式具体如下：

$$W_{ik}(d) = \frac{tf_{ik}(d) \times \log\left(\frac{N}{n_k} + 0.01\right)}{\sqrt{\sum_{i=1}^n (tf_{ik}(d))^2 \times \left[\log\left(\frac{N}{n_k} + 0.01\right)\right]^2}} \times \alpha(H_{ac}) \times H_{ic} \quad (\text{公式 6})$$

其中  $tf_{ik}(d) \times \log\left(\frac{N}{n_k} + 0.01\right)$  为传统 TF-IDF 公式， $\frac{tf_{ik}(d) \times \log\left(\frac{N}{n_k} + 0.01\right)}{\sqrt{\sum_i^n (tf_{ik}(d))^2 \times \left[\log\left(\frac{N}{n_k} + 0.01\right)\right]^2}}$  为 TF-IDF

的归一化计算公式。 $W_{ik}$  指特征项  $t_i$  在  $d_k$  类文档中的权重， $tf_{ik}$  表示特征项  $t_i$  在  $d_k$  类文档中出现的频率， $N$  表示包含特征  $t_i$  的所有文档总数， $n_k$  表示  $d_k$  类文档中出现特征项  $t_i$  的文本数。 $\alpha(H_{ac})$  表示特征项的类间分布信息熵，当含有特征项的文本在各类文档中分布越均匀时， $\alpha(H_{ac})$  值越小，反之越大； $H_{ic}$  则是特征项的类内信息分布熵，当特征项在一类文档的各文本中分布越均匀时， $H_{ic}$  值越大，反之越小。

即便确定是金庸的 15 部作品，利用传统的 TF-IDF 算法，作品间的相似度也不高。为了对疑似作品《卧龙记》进行归属的判定，本文采用改进的 TF-IDF 算法，并



利用向量夹角余弦值（cosine）来计算文本之间的相似度。向量夹角余弦相似度（cosine similarity）具体计算公式如下：

$$sim(d_1, d_2) = \frac{\sum_{t=1}^n w_{d_1t} \times w_{d_2t}}{\sqrt{\sum_{t=1}^n (w_{d_1t})^2} \sqrt{\sum_{t=1}^n (w_{d_2t})^2}} \quad (\text{公式 7})$$

其中  $sim$  即指  $d_1$  和  $d_2$  两个文档的余弦相似度。文本中特征项向量构成了文本向量，可用于计算文本向量的相似度。因此，可将文档  $d_1$  和  $d_2$  的向量分别表示为：

$d_1(w_{d_11}, w_{d_12}, w_{d_13}, \dots w_{d_1t}), d_2(w_{d_21}, w_{d_22}, w_{d_23}, \dots w_{d_2t})$ 。

由上文可知，计算文本的相似度，首先需要使用 TF-IDF 算法公式来对文本中的词项进行赋值，得到文本中所有特征项向量。但如果将文本中所有词项向量来进行相似度的计算，会导致文本向量过于稀疏，产生过高的计算代价。因此，文本将利用 TF-IDF 算法公式所得的特征项根据权重值进行排序，选取 983 个重要特征项向量来代表文本向量。具体操作步骤如下：

- ① 将 16 部小说的编码格式调整为 utf-8 格式；
- ② 利用分词软件 ICTCLAS 进行分词并对分词结果进行校验；
- ③ 利用公式 6，计算 16 部小说中词项的 TF-IDF 值，选取重要特征项，构建文本向量，共 16 个向量；
- ④ 利用公式 7，计算 16 个向量之间的余弦相似度；
- ⑤ 得到各部小说之间的相似度。所得的余弦相似度的值介于 0 和 1 之间，两个向量越接近 1，说明这两个向量所代表的小说之间越相似，两个向量越接近 0，说明这两个向量所代表的小说之间越不相似。

具体统计结果如下表 5.1：



表 5.1 基于改进的 TF-IDF 算法的相似度统计表

作品 相似度 作品	①	②	③	④	⑤	⑥	⑦	⑧
①书剑恩仇录	1	0.852	0.686	0.712	0.548	0.799	0.587	0.652
②碧血剑	0.852	1	0.687	0.794	0.522	0.830	0.637	0.585
③射雕英雄传	0.686	0.687	1	0.712	0.748	0.727	0.761	0.420
④雪山飞狐	0.712	0.794	0.712	1	0.600	0.784	0.662	0.501
⑤神雕侠侣	0.548	0.522	0.748	0.600	1	0.622	0.755	0.369
⑥飞狐外传	0.799	0.830	0.727	0.784	0.622	1	0.690	0.606
⑦倚天屠龙记	0.587	0.637	0.761	0.662	0.755	0.690	1	0.350
⑧鸳鸯刀	0.652	0.585	0.420	0.501	0.369	0.606	0.350	1
⑨白马啸西风	0.597	0.544	0.452	0.461	0.369	0.516	0.387	0.418
⑩天龙八部	0.452	0.622	0.624	0.576	0.576	0.525	0.787	0.262
⑪连城诀	0.550	0.526	0.683	0.507	0.615	0.589	0.706	0.375
⑫侠客行	0.400	0.600	0.531	0.567	0.471	0.450	0.698	0.249
⑬笑傲江湖	0.424	0.471	0.561	0.412	0.541	0.498	0.698	0.486
⑭鹿鼎记	0.602	0.650	0.750	0.614	0.653	0.654	0.782	0.343
⑮越女剑	0.090	0.237	0.132	0.204	0.164	0.097	0.236	0.021
⑯卧龙记	0.144	0.136	0.161	0.112	0.135	0.176	0.195	0.081



续表

作品 相似度 作品	⑨	⑩	⑪	⑫	⑬	⑭	⑮	⑯
①书剑恩仇录	0.597	0.452	0.550	0.400	0.424	0.602	0.090	0.144
②碧血剑	0.544	0.622	0.526	0.600	0.471	0.650	0.237	0.136
③射雕英雄传	0.452	0.624	0.683	0.531	0.561	0.750	0.132	0.161
④雪山飞狐	0.461	0.576	0.507	0.567	0.412	0.614	0.204	0.112
⑤神雕侠侣	0.369	0.576	0.615	0.471	0.541	0.653	0.164	0.135
⑥飞狐外传	0.516	0.525	0.589	0.450	0.498	0.654	0.097	0.176
⑦倚天屠龙记	0.387	0.787	0.706	0.698	0.698	0.782	0.236	0.195
⑧鸳鸯刀	0.418	0.262	0.375	0.249	0.486	0.342	0.021	0.081
⑨白马啸西风	1	0.316	0.425	0.268	0.277	0.423	0.095	0.119
⑩天龙八部	0.316	1	0.652	0.863	0.635	0.725	0.410	0.187
⑪连城诀	0.425	0.652	1	0.552	0.623	0.762	0.120	0.159
⑫侠客行	0.268	0.863	0.552	1	0.564	0.656	0.452	0.198
⑬笑傲江湖	0.277	0.635	0.623	0.564	1	0.672	0.156	0.176
⑭鹿鼎记	0.423	0.725	0.762	0.656	0.672	1	0.158	0.244
⑮越女剑	0.095	0.410	0.120	0.452	0.156	0.158	1	0.052
⑯卧龙记	0.119	0.187	0.159	0.198	0.176	0.244	0.052	1

根据表 5.1 所得数据可知，金庸小说中，各小说之间的相似度情况复杂，最高可达 0.863（《天龙八部》与《侠客行》之间），最低则只有 0.021（《书剑恩仇录》与《越女剑》之间）。《越女剑》是金庸目前已知的唯一一部短篇小说，并且也是连载最晚的小说之一，它与金庸其他作品之间的相似度非常小，除与《侠客行》与《天龙八部》的相似度在 0.4 以上以外，与其他的作品的相似度都在 0.3 以下，甚至有 0.1 以下的相似度。疑似作品《卧龙记》与已知的 15 部金庸作品之间的相似度非常低，几乎都在 0.2 以下。

由前文所知，金庸的 15 部小说可分为：中短篇、小长篇、及超级长篇。以小说篇幅作为分类标准来对统计结果进行分析，可整理出下表 5.2、表 5.3 以及表 5.4，具体情况如下：



表 5.2 金庸中短篇小说之间的相似度统计表

作品 相似度 作品	①	②	③
①越女剑	1	0.021	0.095
②鸳鸯刀	0.021	1	0.418
③白马啸西风	0.095	0.418	1

由上表 5.2 可知，在金庸的中短篇小说中，《越女剑》为短篇小说，与其他两部小说之间的差异非常大，相似度都仅在 0.1 之下。《鸳鸯刀》与《白马啸西风》之间的相似度明显较高。

表 5.3 金庸小说小长篇之间的相似度统计表

作品 相似度 作品	①	②	③	④	⑤	⑥	⑦
①雪山飞狐	1	0.507	0.567	0.794	0.784	0.712	0.112
②连城诀	0.507	1	0.552	0.526	0.589	0.55	0.159
③侠客行	0.567	0.552	1	0.6	0.45	0.4	0.198
④碧血剑	0.794	0.526	0.6	1	0.83	0.852	0.136
⑤飞狐外传	0.784	0.589	0.45	0.83	1	0.799	0.176
⑥书剑恩仇录	0.712	0.55	0.4	0.852	0.799	1	0.144
⑦卧龙记	0.112	0.159	0.198	0.136	0.176	0.144	1

由表 5.3 可知，在金庸篇幅在 12 万字到 50 万字的小长篇中，《连城诀》和《侠客行》两部小说与其他小说之间的相似度都较低，都只在 0.4-0.6 之间，而《雪山飞狐》等其他四部小说之间的相似度都在 0.7 以上，其中，《碧血剑》与《书剑恩仇录》的相似度最高。从创作时间上看，从 1955 到 1956 年间陆续创作的《碧血剑》和《书剑恩仇录》有着最高的相似度，与 1959 到 1961 年间陆续创作的《雪山飞狐》与《飞狐外传》也有着较高的相似度。但这四部小说与 1963 年后创作的其他两部小说之间的差异较为明显。前文已知，《越女剑》与《卧龙记》和其他金庸小说之间的相似度都非常低，但由于《越女剑》篇幅长度在金庸小说中的特殊性，因此可将其视作特例。但已知金庸小说中，有多部作品与《卧龙记》的篇幅十分相似，因此将它们之间的相似度进行对比，可以得到较为可信的结论。由上表可见，《卧龙记》与金庸已知的六部小长篇之间的相似度远低于六部小说之间的相似度，仅在 0.200 之下。



表 5.4 金庸超级长篇小说之间的相似度统计表

作品 相似度 作品	①	②	③	④	⑤	⑥
①射雕英雄传	1	0.748	0.761	0.624	0.561	0.75
②神雕侠侣	0.748	1	0.755	0.576	0.541	0.653
③倚天屠龙记	0.761	0.755	1	0.787	0.698	0.782
④天龙八部	0.624	0.576	0.787	1	0.635	0.725
⑤笑傲江湖	0.561	0.541	0.698	0.635	1	0.672
⑥鹿鼎记	0.75	0.653	0.782	0.725	0.672	1

由表 5.4 中数据可知，在金庸的超级长篇作品中，六部作品之间的相似度在 0.5-0.8 之间。其中，《倚天屠龙记》与《天龙八部》之间相似度最高，《笑傲江湖》与《射雕英雄传》以及《神雕侠侣》之间，《天龙八部》与《神雕侠侣》之间的相似度相对较低。六部作品的最初连载时间基本是连续的，除了《天龙八部》与《笑傲江湖》之间的连载时间相隔较久。被誉为“射雕三部曲”的《射雕英雄传》、《神雕侠侣》以及《倚天屠龙记》，故事发生背景南宋到明朝建立，内容具有相对连续性，数据也表明三者之间具有较高的相似度。另外，连载时间最晚的《鹿鼎记》与其他小说之间的相似度都相对较高。

## 5.2 基于 LSI 算法的文本相似度分析

潜在语义索引 (latent semantic indexing, LSI)，又称为潜在语义分析 (latent semantic analysis)，是一种无监督的数据挖掘技术，它充分地考虑了关键词的上下文信息，因而能够较为有效地消除文本中同义词、多义词等语义问题带来的影响。潜在语义索引模型，首先是将文本集通过特定的算法转化为特征项向量空间，然后通过奇异值分解法 (Singular Value Decomposition, SVD) 来对特征向量空间进行分解，以达到降维的目的，最终得到一个简化后的向量空间。

前文中改进过的 TF-IDF 算法依然受到诸如同义词、多义词带来的影响。为了对疑似作品《卧龙记》进行归属的判定，我们先利用 TF-IDF 算法来为特征项赋值，构建向量空间，再采用基于上下文语境的 LSI 算法来进行降维，最后计算向量的余弦相似度。对金庸确认的 15 部作品和疑似作品《卧龙记》共计 16 部小说进行计算的主要步骤如下：

- ① 将 16 部小说的编码格式调整为 utf-8 格式；



- ②利用分词软件 ICTCLAS 进行分词并对分词结果进行校验；  
③利用公式 6，计算 16 部小说中词项的 TF-IDF 值，选取重要特征项，组成词袋模型并构建文本向量空间；  
④利用 LSI 算法对文本向量空间进行降维处理，得到简化后的文本向量，共 16 个向量；  
⑤利用公式 7，计算所得的 16 个向量之间的余弦相似度。  
⑥得到各部小说之间的相似度。所得的余弦相似度的值介于 0 和 1 之间，两个向量越接近 1，说明这两个向量所代表的小说之间越相似，两个向量越接近 0，说明这两个向量所代表的小说之间越不相似。具体统计结果如表 5.5：

表 5.5 基于 LSI 算法的相似度统计表

作品 相似度 作品	①	②	③	④	⑤	⑥	⑦	⑧
①书剑恩仇录	1	0.963	0.861	0.950	0.770	0.994	0.722	0.753
②碧血剑	0.963	1	0.861	0.983	0.777	0.974	0.774	0.702
③射雕英雄传	0.861	0.861	1	0.905	0.985	0.914	0.940	0.475
④雪山飞狐	0.950	0.983	0.905	1	0.869	0.968	0.849	0.591
⑤神雕侠侣	0.770	0.777	0.985	0.869	1	0.830	0.953	0.352
⑥飞狐外传	0.994	0.974	0.914	0.968	0.830	1	0.823	0.701
⑦倚天屠龙记	0.722	0.774	0.940	0.849	0.953	0.823	1	0.422
⑧鸳鸯刀	0.753	0.702	0.475	0.591	0.352	0.701	0.422	1
⑨白马啸西风	0.960	0.931	0.761	0.895	0.642	0.914	0.635	0.729
⑩天龙八部	0.558	0.687	0.770	0.743	0.789	0.658	0.895	0.389
⑪连城诀	0.718	0.725	0.935	0.796	0.948	0.821	0.972	0.573
⑫侠客行	0.507	0.661	0.674	0.697	0.683	0.591	0.816	0.362
⑬笑傲江湖	0.521	0.545	0.661	0.553	0.663	0.608	0.766	0.663
⑭鹿鼎记	0.741	0.769	0.946	0.838	0.951	0.837	0.985	0.531
⑮越女剑	0.097	0.319	0.086	0.296	0.077	0.106	0.236	0.006
⑯卧龙记	0.150	0.142	0.134	0.102	0.074	0.146	0.154	0.062



续表

作品 相似度 作品	⑨	⑩	⑪	⑫	⑬	⑭	⑮	⑯
①书剑恩仇录	0.960	0.558	0.718	0.507	0.521	0.741	0.097	0.150
②碧血剑	0.931	0.687	0.725	0.661	0.545	0.769	0.319	0.142
③射雕英雄传	0.761	0.770	0.935	0.674	0.661	0.946	0.086	0.134
④雪山飞狐	0.895	0.743	0.796	0.697	0.553	0.838	0.296	0.102
⑤神雕侠侣	0.642	0.789	0.948	0.683	0.663	0.951	0.077	0.074
⑥飞狐外传	0.914	0.658	0.821	0.591	0.608	0.837	0.106	0.146
⑦倚天屠龙记	0.635	0.895	0.972	0.816	0.766	0.985	0.236	0.154
⑧鸳鸯刀	0.729	0.389	0.573	0.362	0.663	0.531	0.006	0.062
⑨白马啸西风	1	0.466	0.615	0.422	0.344	0.666	0.082	0.230
⑩天龙八部	0.466	1	0.882	0.960	0.756	0.924	0.532	0.175
⑪连城诀	0.615	0.882	1	0.730	0.815	0.984	0.076	0.185
⑫侠客行	0.422	0.960	0.730	1	0.718	0.855	0.661	0.186
⑬笑傲江湖	0.344	0.756	0.815	0.718	1	0.862	0.108	0.157
⑭鹿鼎记	0.666	0.924	0.984	0.855	0.862	1	0.161	0.225
⑮越女剑	0.082	0.532	0.076	0.661	0.108	0.161	1	0.046
⑯卧龙记	0.230	0.175	0.185	0.186	0.157	0.225	0.046	1

根据表 5.5 所得数据可知,与利用改进的 IF-IDF 算法得到的各小说相似度相比,基于 LSI 算法所得的相似度基本上有明显的升高,并且各小说之间相似度的情况也发生了变化。金庸 15 小说中,《书剑恩仇录》与《飞狐外传》之间的相似度最高,为 0.994;而《鸳鸯刀》与《越女剑》之间的相似度则最低,只有 0.006。《越女剑》仍是金庸小说中基本上与其他小说相似度最低的作品,但它与《侠客行》和《天龙八部》之间的相似度分别达到了 0.661 和 0.532。疑似作品《卧龙记》与已知金庸作品之间的相似度仍然非常低,相似度范围仅在 0.046-0.230 之间。

与上文一致,以小说篇幅作为分类标准来对统计结果进行分析,可整理出表 5.6、表 5.7 以及表 5.8,具体情况如下:

表 5.6 金庸中短篇小说之间的相似度统计表

作品 相似度 作品	①	②	③
①越女剑	1	0.006	0.082
②鸳鸯刀	0.006	1	0.729
③白马啸西风	0.082	0.729	1



由表 5.6 可知，短篇小说《越女剑》与其他两部小说的相似度仍旧十分低，而另外两部小说之间的相似度则较高。与上文所得的计算结果相比，《越女剑》与另外两部小说的相似度更低，《鸳鸯刀》与《白马啸西风》的相似度则明显升高。

表 5.7 金庸小说小长篇之间的相似度统计表

作品 相似度 作品	①	②	③	④	⑤	⑥	⑦
①雪山飞狐	1	0.796	0.697	0.983	0.968	0.95	0.102
②连城诀	0.796	1	0.73	0.725	0.821	0.718	0.185
③侠客行	0.697	0.73	1	0.661	0.591	0.507	0.186
④碧血剑	0.983	0.725	0.661	1	0.974	0.963	0.142
⑤飞狐外传	0.968	0.821	0.591	0.974	1	0.994	0.146
⑥书剑恩仇录	0.95	0.718	0.507	0.963	0.994	1	0.15
⑦卧龙记	0.102	0.185	0.186	0.142	0.146	0.15	1

通过观察表 5.7 中的数据可知，在金庸 12 万字到 50 万字的小长篇中，《侠客行》与四部小说之间的相似度较低，都在 0.750 以下，与《飞狐外传》和《书剑恩仇录》的相似度都仅在 0.600 以下。《连城诀》与其他小说之间的相似度也仅在 0.718-0.821 之间。而《雪山飞狐》等其他四部小说之间则有着非常高的相似度，它们之间的相似度均在 0.950 之上。《雪山飞狐》等四部作品的连载时间较为接近，与《连城诀》与《侠客行》两部小说之间的连载时间有着约两年以上的间隔。疑似作品《卧龙记》与其他已知金庸小长篇有着非常显著的差异，它们之间的相似度都只在 0.200 以下。

相比基于改进的 IF-IDF 算法计算所得的文本相似度，金庸六部小说之间的相似度有显著上升，《雪山飞狐》等作品之间的相似度接近于 1。然而，它们与《卧龙记》之间的相似度并未随着算法的改变而发生明显变化，其结果都只在 0.200 之下。

表 5.8 金庸超级长篇小说之间的相似度统计表

作品 相似度 作品	①	②	③	④	⑤	⑥
①射雕英雄传	1	0.985	0.94	0.77	0.661	0.946
②神雕侠侣	0.985	1	0.953	0.789	0.663	0.951
③倚天屠龙记	0.94	0.953	1	0.895	0.766	0.985
④天龙八部	0.77	0.789	0.895	1	0.756	0.924
⑤笑傲江湖	0.661	0.663	0.766	0.756	1	0.862
⑥鹿鼎记	0.946	0.951	0.985	0.924	0.862	1



由表 5.8 可见，金庸的六部超级长篇中，各小说的相似度皆在 0.600 以上。其中，《鹿鼎记》与所有其他五部小说之间都有着较高的相似度，都在 0.850 以上；“射雕三部曲”之间的相似度则在 0.940 以上；《天龙八部》与《笑傲江湖》两部小说与“射雕三部曲”之间的相似度相对较低，在 0.661-0.895 之间。“射雕三部曲”的创作连载时间早于《天龙八部》与《笑傲江湖》，基本上呈现时间相隔越久，它们与后两者之间的相似度越低的规律。《倚天屠龙记》完成连载之后紧接着便是《天龙八部》的连载，因此两者间的相似度明显高于其他两部“射雕”作品与《天龙八部》之间的相似度。与上文计算所得的相似度相比，六部小说间的相似度同样明显升高，多在 0.700 以上。但各小说之间相似度的规律并没有发生变化。

### 5.3 基于情感分析的文本相似度分析

情感分析（sentiment analysis）是指利用计算机对文本、图像视频、语音等信息中的观点、情感、态度等的分析挖掘和计算建模，在舆情管理、商业决策、反恐侦查、大数据分析等方面有着重要意义。（黄民烈，2016）本文将利用情感分析相关理论与方法对金庸 15 部小说及其疑似作品《卧龙记》进行探究，并采用假设检验中的卡方检验对情感分析的结果进行检验，从而进一步探究《卧龙记》的归属问题。

#### （一）情感倾向统计

情感分析主要有以规则为主的方法、基于传统机器学习的方法和基于深度学习的方法等。情感词典就是基于以规则为主的方法所构建的情感资源。基于情感词典进行情感分析的原理是将文本内容与情感词典中的词项进行参照匹配，通过寻找重合的情感词来确定文本的情感极性。情感极性一般有积极（正向）、消极（负向）与中性。常见的情感词典有台湾大学中文情感极性词典（NTUSD）、清华大学李军中文褒贬义词典（TSING）、大连理工大学中文情感词汇本体库、知网情感词典（HOWNET）。

本文采用知网情感词典（HOWNET）来对文本进行情感分析。知网情感词典中包括有程度级别词语、正面情感词语、负面情感词语、正面评价词语、负面评价词语以及主张词语等几个部分。文本中存在着大量不承载语义信息的词语，如代词“你”、“他”，助词“的”、“得”等，这类词语被称作停用词（Stop Words）。由于它们不具有任何情感意义，但却会增加分析工作量，因此在文本加工过程中需将其进行删除。除此之外，在进行情感分析的过程中，否定词往往会改变词语的情感倾向，导致文本的情感值发生变化，因此在分析中需加入否定词表。

本文在对小说文本进行分析时，是以句子为基本单位输出情感极性，即对单个



文本中的所有的句子进行情感分析。具体流程是，参照知网情感词典及否定词表，逐一匹配句子中的词语，然后对匹配的词语进行赋值后，最终确定句子的情感值。16 部小说中不同情感倾向的句子数详见附录 4。由于每部小说的字数不同，因此计算各部小说中每类情感倾向的句子在文本的比重，具体结果如下表 5.9 所示：

表 5.9 各类情感倾向比重统计表

作品 \ 情感倾向	积极	消极	中性
书剑恩仇录	36.13%	11.58%	52.29%
碧血剑	37.69%	11.58%	50.73%
射雕英雄传	40.76%	10.57%	48.67%
雪山飞狐	39.36%	11.14%	49.50%
神雕侠侣	41.97%	11.78%	46.25%
飞狐外传	40.74%	11.52%	47.74%
倚天屠龙记	41.64%	11.62%	46.74%
鸳鸯刀	38.40%	11.11%	50.49%
白马啸西风	43.44%	13.21%	43.35%
天龙八部	41.92%	11.91%	46.19%
连城诀	43.30%	10.98%	45.72%
侠客行	41.51%	11.51%	46.98%
笑傲江湖	39.21%	11.93%	48.86%
鹿鼎记	38.65%	12.95%	48.39%
越女剑	32.70%	13.15%	54.15%
卧龙记	42.97%	17.72%	39.31%

将表 5.9 中的不同情感倾向句子所占百分比绘制成图，以便更清晰地展示各小说在不同情感倾向方面的情况，具体如图 5.1、图 5.2 及图 5.3 所示：



图 5.1 积极情感倾向句子占比柱状图

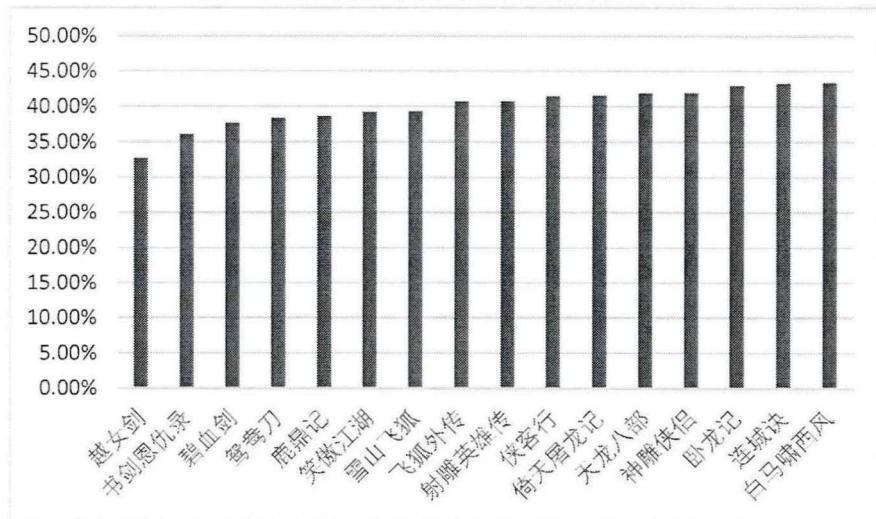


图 5.1 中横轴为各小说名，纵轴为积极情感倾向句子的比重。结合图表可知，在金庸各部小说中，积极情感倾向在整体文本中的占比范围约在 32.70%-43.44% 之间，均值约为 39.8%，标准差为 0.029。其中，《越女剑》的占比明显最低，与其他金庸小说之间的差距较显著。金庸小说的积极情感倾向与小说的篇幅和创作时间无明显关联。疑似作品《卧龙记》中积极情感倾向的占比约为 42.97%，落在已知的金庸小说范围之内，高于平均值，在 16 部小说中排位较高。

图 5.2 消极情感倾向句子占比柱状图

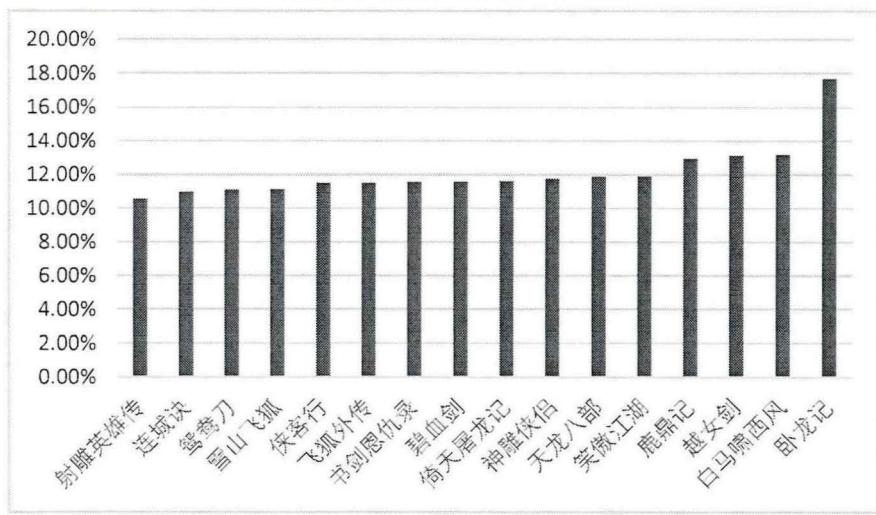


图 5.2 中横轴为各小说名，纵轴为消极情感倾向句子的比重。观察图表会发现，金庸小说在消极情感倾向方面的占比范围为 10.6%-13.2%，均值为 11.77%，标准差



为 0.008。其中，《射雕英雄传》消极情感倾向的句子占比最低，《白马啸西风》则最高。消极情感倾向也与金庸小说的篇幅和创作时间无明显联系。疑似作品《卧龙记》的消极情感占比远高于金庸小说作品，与《白马啸西风》的差值约 4.52%。

图 5.3 中性情感倾向句子占比柱状图

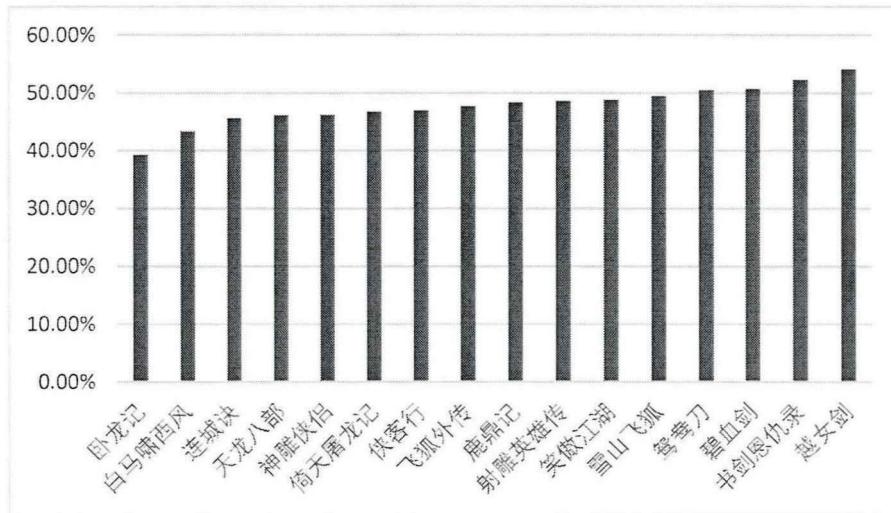


图 5.3 中横轴为各小说名，纵轴为中性情感倾向句子的占比。由图表可知，金庸的 15 部小说中性情感倾向的句子在文本中的占比范围为 43.35%-54.15%，均值为 48.40%，标准差为 0.028。其中，《白马啸西风》中性情感倾向的句子占比最高，《越女剑》最低。同样地，中性情感倾向与小说长短和写作时间之间无明显关联。疑似作品《卧龙记》的占比约为 39.31%，未落在金庸小说已知的范围内，且与金庸小说之间的差异较大。

某一类情感倾向的句子数量是用于判断文本情感倾向的数据支持。综上可知，金庸小说中积极情感倾向的句子占比明显高于消极情感，持平或略低于中性情感倾向。总体而言，金庸小说体现出中性的感情倾向。

## （二）情感倾向的卡方检验

为了进一步对金庸的 15 部小说与疑似作品《卧龙记》的情感倾向进行对比分析，我们将金庸 15 部小说中相应情感倾向的句子进行合并，再与疑似作品《卧龙记》的情感倾向进行卡方检验。由于 16 部小说的字数差异较大，为提高检验的准确性，需对金庸小说和《卧龙记》统计所得的不同情感倾向的句子进行标准化处理，即统计每一千句中，金庸小说与《卧龙记》不同情感倾向的句子可能出现的频数。金庸小说不同情感倾向的句子合并方法为计算每一类情感倾向的句子数的平均值，具体数值如表 5.10 所示：



表 5.10 金庸小说与《卧龙记》每一千句中各类情感倾向句子数统计表

作品 \ 情感倾向	积极	消极	中性
金庸小说	398.28	117.70	484.04
卧龙记	429.73	177.21	393.07

由于在词频分布的部分已对 R×C 卡方检验进行过相关介绍，这里不再赘述。为检验《卧龙记》与金庸小说在积极情感、消极情感以及中性情感上是否存在差异，提出原假设：它们在三类情感倾向上没有明显差异。卡方检验结果如表 5.11：

表 5.11 卡方检验表

卡方检验			
	值	自由度	渐进显著性 (双侧)
皮尔逊卡方	22.479 <sup>a</sup>	2	.000
似然比	22.576	2	.000
线性关联	8.875	1	.003
有效个案数	2000		

a. 0 个单元格 (0.0%) 的期望计数小于 5。最小期望计数为 147.50。

由表 5.11 可知，皮尔逊卡方值为 22.479，自由度为 2，P 值<0.05，则原假设不成立，《卧龙记》与金庸小说的各情感倾向有显著差异，即两者在积极、消极和中性三类情感倾向上不完全一致。

利用 Post hoc testing 检验来进一步考察《卧龙记》与金庸小说之间在三类情感倾向上的差异，可得到情感类别与作品交叉表，具体如表 5.12 所示：



表 5.12 情感类别与作品交叉表

情感类别 \* 作品 交叉表

情感类别	积极	计数	作品		
			金庸小说	卧龙记	总计
情感类别	积极	计数	398a	430a	828
		占作品的百分比	39.8%	43.0%	41.4%
情感类别	消极	计数	118a	177b	295
		占作品的百分比	11.8%	17.7%	14.8%
情感类别	中性	计数	484a	393b	877
		占作品的百分比	48.4%	39.3%	43.9%
总计		计数	1000	1000	2000
		占作品的百分比	100.0%	100.0%	100.0%

每个下标字母都指示作品类别的子集，在 .05 级别，这些类别的列比例相互之间无显著差异。

观察表 5.12 可知，金庸小说与《卧龙记》在积极情感方面的计数值后都是 a 的字母标记，这说明两者在这一情感倾向上并无明显差异。在消极情感上，金庸小说与《卧龙记》之间存在明显差异，两者的计数后字母标记分别为 a 和 b。从百分比上可以看出，金庸小说的消极情感倾向占比低于《卧龙记》。就中性情感倾向而言，两者同样存在差异，金庸小说的中性情感倾向明显高于《卧龙记》。

综上可知，金庸小说的积极情感倾向与《卧龙记》之间没有明显差异，而在消极情感与中性情感方面差异较大，这进一步佐证了上文中通过分析句子情感倾向占比所得的结论。因此，总的来说，金庸小说与《卧龙记》在情感倾向上是存在差异的。

#### 5.4 本章小结

本章从基于改进 TF-IDF 算法、基于 LSI 算法以及基于情感分析的角度对 16 部小说的相似度进行了计算与检验。在前两节中，用余弦相似度分别与改进 TF-IDF 算法、LSI 算法相结合对 16 部小说之间的相似度进行了计算，所得结果较为复杂无序，为了更加有效地对文本之间的相似度进行分析，研究中将金庸的 15 部小说分为中短篇、小长篇、超级长篇三类，并对其中的小说之间的相似度进行了比较。在基于改进 TF-IDF 算法和基于 LSI 算法的两部分中，短篇小说《越女剑》与其他金庸小说相似度最低，且远远低于其他小说间的相似度。在六部小长篇中，《连城诀》和《侠客行》两部小说与四部其他小说之间相似度较低，而《雪山飞狐》等其他四



部小说之间的相似度都相对较高；小说间的相似度与创作时间的先后有着一定程度上的一致性。在六部超级长篇中“射雕三部曲”之间具有较高的相似度，而连载时间最晚的《鹿鼎记》与其他小说之间的相似度都相对较高。对比第一小节与第二小节计算所得的数据会发现，基于 LSI 算法所得的金庸小说文本间相似度明显高于仅基于改进 TF-IDF 算法所得的结果。而不论是基于改进 TF-IDF 算法，还是基于 LSI 算法，疑似作品《卧龙记》与金庸已知的 15 小说之间的相似度都非常低，远远低于金庸已知小说之间的平均相似度，与同等量级的作品之间的差异也非常显著。在基于情感分析的部分，金庸小说中积极情感倾向的句子占比明显高于消极情感，持平或略低于中性情感倾向。总体而言，金庸小说体现出中性的情感倾向。对金庸小说与疑似作品《卧龙记》中各类情感倾向进行卡方检验，检验结果表明，金庸小说积极情感与《卧龙记》无明显差异，但消极情感与中性情感倾向方面则存在明显差别。综上，疑似作品《卧龙记》与金庸小说之间存在着不可忽视的差异。下一章将对全文进行总结，并反思研究中存在的不足和进一步研究的展望。



## 第六章 结语

### 6.1 本文总结

当前，多角度、多方法对金庸小说文本进行比较全面的计量统计与分析的相关研究数量极少。以往对金庸语言风格进行探索的研究多是从定性分析的角度出发，对其语言风格进行较为主观的概括与描述，这一方面可以更为细致地刻画金庸语言的风格特点，但在另一方面却又无法展现出较为客观抽象的语言整体风貌。因此，本文基于语料库语言学、统计语言学、计算语言学等领域的原理与方法，从字符、词汇、句子、段落等多个角度对金庸小说的语言特征进行统计与分析，结合定量分析与定性分析，从宏观与微观两个角度探究金庸小说语言风格。除此之外，本文还加入了对金庸疑似作品《卧龙记》的判别研究。

研究中分别从字符、词汇、句子、段落等各层面选取不同的语言特征来进行频率统计，在此基础上，利用相关算法模型对文本相似度进行计算和检验。在字符与词汇层面选用了标点符号、用字量、平均词长、词长离散度、词长分布、类符形符比、独现词频率、词汇密度、词类分布等语言特征进行统计。在句子与段落层面选用了平均句长、句长离散度、句长分布、平均段落长度等语言特征进行统计。最后是利用基于改进 TF-IDF 算法和 LSI 算法与余弦相似度相结合的方法来计算文本之间的相似度，另外，还采用情感分析相关理论和方法对文本进行统计分析，并在此基础上利用卡方检验对文本相似度进行假设检验。

金庸各小说语言风格之间存在差异，但在某种程度上来说，也存在相似性，存在可观察的规律。统计结果表明，金庸 15 部小说的平均词长、词长离散度、词汇密度、平均句长、句长离散度、平均段落长度之间存在着差异，并呈现出无序的状态。而在标点符号、用字量、词汇丰富度、词类分布、句长分布等方面具有某些可循的规律。在标点符号的使用上，金庸的 15 部小说之间具有较高的一致性；用字量方面，金庸小说小长篇内部与超级长篇内部的用字量相对稳定；就词汇丰富度而言，金庸中短篇小说的词汇多样性高于长篇小说，同等量级的小说之间的差异较小；就词类分布情况来看，金庸 15 部小说中实词内部分布顺序具有一致性，虚词内部分布顺序也相一致；在词长分布与句长分布上，各小说的具体分布图式具有相似性。另外，文本相似度计算结果表明，短篇小说《越女剑》与其他金庸小说相似度最低，且远远低于其他小说间的相似度。在六部小长篇中，《连城诀》和《侠客行》两部小说与四部其他小说之间相似度较低，而《雪山飞狐》等其他四部小说之间的相似度



都相对较高；小说间的相似度与创作时间的先后有着一定程度上的一致性。在六部超级长篇中“射雕三部曲”之间具有较高的相似度，而连载时间最晚的《鹿鼎记》与其他小说之间的相似度都相对较高。情感分析中，总的来说，金庸小说呈现出中性的感情倾向。

而从判别伪作方面来看，将《卧龙记》的统计结果与金庸小说相对比，发现在用字量、词长分布、类形符比、词汇密度、频次排名前一百的词语、词类分布、平均句长、句长分布等语言特征上，《卧龙记》与金庸小说之间并无显著差异。而在标点符号的使用、平均词长、词长离散度、独现词频率、词频分布、句长离散度、平均段落长度、相似度计算与情感分析方面，两者有着明显不同。以上语言特征的频率统计对于辨别文本的归属问题较为粗糙，有着非常大的局限性。而基于算法模型的相似度计算与基于情感分析的相似度分析可以对单一的频率统计进行较为有益的补充。因此，对文本进行真伪判别时需要利用多种方法来进行多维度的考察，才可能得到比较可靠的结论。综合各部分的统计结果和分析来看，疑似作品《卧龙记》是金庸作品的可能性极低。

## 6.2 不足与展望

本研究中存在许多明显的不足，主要有以下几点：

一，语料的分词还不够完善。虽然我们使用了目前中国最优秀的由北京理工大学海量语言信息处理与云计算工程研究中心开发的分词软件 ICTCLAS，但是 ICTCLAS 的切分效果仍然不尽如人意。尽管我们对语料分词后的人名和地名进行了人工校对，但是仍然还有许多其他类别的未登录词发生切分错误，这在一定程度上会影响我们的统计结果。

二，基于情感词典的分析模块存在一些缺陷。本文在判断小说的积极、消极和中性情感倾向时使用的是基于情感词典的算法。情感词典在带来稳定性、便捷性的同时，也有一些天生的、不可避免的缺陷。比如有的词在不同的语境下，情感倾向会有所有不同。但是在情感词典里的情感倾向却是固定的。因此，难免会使相应的统计结果产生偏差。

三，文章的定性分析稍显不足。限于篇幅以及为了突出定量方法在作家语言风格研究上的作用，本文的定量分析方法运用得较多，但是定性分析稍显不足。文中对某一语言特征进行定量分析之后，往往没有再结合语料进一步进行解释，而是根据所得到的统计结果直接就给出了相应的结论。

四，文章中一些部分的考察过于单薄。由于文章篇幅限制和具体操作方面的



困难，本文在各个语言单位的考察方面存在着内容不够丰富全面的问题。例如，在句子层面，缺少对金庸小说作品中的句子在语序上的特殊性和在句法结构选取的倾向性等方面的探究、统计和分析。

计量语言学在作家语言风格与作家作品的证伪研究等领域，的确能解决以往单纯靠定性分析方法无法解决的问题，但同时也存在许多不足。在未来的研究中有以下几点值得进一步探索：

首先，提高分词软件的精度。现有的分词软件为了求全，其词库往往缺乏精细化的设置。为了在某一方面的分词精度有所提高，可以在现有效果较好的分词软件的基础上进行二次开发。

其次，为避免基于情感词典在分析情感不考虑上下文语境的缺陷，可以在传统情感词典分析的基础上把情感词左右各 X 个窗口的内容考虑进去，或者直接采用基于机器学习的方法，提高情感分析的精度。

利用统计分析方法对作家的作品进行分析，不仅能够揭示某一作家独有的语言风格，还能对一门语言结构进行更为深入地了解。与目前基于统计学、数学等领域的理论和方法对文本语言风格进行研究中，一方面需要在数学方法与理论上有所改进与突破，一方面也需要在语言风格特征的选取问题上进行优化。目前统计中所使用的语言特征具有一定的有效性，但许多特征并没有显著的代表性。语言特征的选取与语言学研究密切相关，如何能够让统计的结果更有效地反映文本的本质特征，进而用于区别不同文本，这是语言学研究中能够有所作为的领域，也是非常值得探索的领域。



## 参考文献

- [1]陈大康.从数理语言学看后四十回的作者——与陈炳藻先生商榷[J].红楼梦学刊,1987(01):293-318.
- [2]陈玲.《哈里·波特系列(I—VI)》的计量文体学分析[D].大连海事大学,2007.
- [3]陈芯莹,李雯雯,王燕.计量特征在语言风格比较及作家判定中的应用——以韩寒《三重门》与郭敬明《梦里花落知多少》为例[J].计算机工程与应用,2012,48(03):137-139+208.
- [4]陈岳红.从计算文体学的角度分析《莫里斯·格斯特》和《乔治一家的妻子》[D].西北大学,2008.
- [5]范亚超,罗天健,周昌乐.基于降噪自编码器特征学习的作者识别及其在《西游记》诗词上的应用[J].厦门大学学报(自然科学版),2018,57(06):884-889.
- [6]黄晖.从计算风格学角度考察《源氏物语》中译本[D].浙江工商大学,2017.
- [7]黄伟,刘海涛.汉语语体的计量特征在文本聚类中的应用[J].计算机工程与应用,2009,45(29):25-27+33.
- [8]洪巍,李敏.文本情感分析方法研究综述[J].计算机工程与科学,2019,41(04):750-757.
- [9]贺湘情,刘颖.基于文本聚类的语言韵律和节奏风格特征挖掘[J].中文信息学报,2014,28(06):194-200+207.
- [10]黄永新,张黎黎.基于语料库的莫言小说英译语言特征考察[J].河北广播电视台学报,2016,21(02):54-58.
- [11]金迪.基于语料库的格非、余华小说计量风格学研究[D].南京师范大学,2018.
- [12]姜晓艳.基于语料库的《简·爱》语言特点及主题词表征分析[J].江苏科技大学学报(社会科学版),2016,16(02):77-84.
- [13]蒋跃,张英贤,韩纪建.英语被动句人机翻译语言计量特征对比——以《傲慢与偏见》译本为例[J].外语电化教学,2016(03):46-51+63.
- [14]吉志薇.改进的TF-IDF 算法在作品抄袭判定中的应用——以《梦里花落知多少》和《圈里圈外》为例[J].文教资料,2014(31):120-124.
- [15] Kennedy, Graeme. An Introduction to Corpus Linguistics[M].北京:外语教学与研究出版社,2000.
- [16]李惠,刘颖.基于语言模型和特征分类的抄袭判定[J].计算机工程,2013,39(05):230-234.
- [17]林鸿飞,战学刚,姚天顺.基于潜在语义索引的文本分析方法[J].模式识别与人工智能,2013,26(05):101-106.



能,2000,13(01):47-51.

- [18]刘海涛,潘夏星.汉语新诗的计量特征[J].山西大学学报(哲学社会科学版),2015,38(02):40-47.
- [19]刘海涛.计量语言学导论[M].北京:商务印书馆,2017.
- [20]梁军,柴玉梅,原慧斌,昝红英,刘铭.基于深度学习的微博情感分析[J].中文信息学报,2014,28(05):155-161.
- [21]刘健,张维明.基于互信息的文本特征选择方法研究与改进[J].计算机工程与应用,2008,44 (10): 135-137.
- [22]林敏.查尔斯·狄更斯和弗吉尼亚·沃尔夫作品的计量文体学分析[D].广西师范大学,2014.
- [23]刘旭鹏.《平凡的世界》中动词重叠式的考察——基于统计的方法分析作品语言风格[J].现代语文(语言研究版),2013(06):60-62.
- [24]陆芸.词汇丰富性测量方法及计算机程序开发:回顾与展望[J].南京工业大学学报(社会科学版),2012,11(02):104-108.
- [25]刘颖.统计语言学[M].北京:清华大学出版社,2014.
- [26]李以建.金庸小说研究的前沿进展与体系构建[J].西南大学学报(社会科学版),2012,38(02):88-93.
- [27]李媛媛,马永强.基于潜在语义索引的文本特征词权重计算方法[J].计算机应用,2008(06):1460-1462+1466.
- [28]刘座簪.当代汉语文本言语特征系统提取研究[D].清华大学,2004.
- [29]Mahlberg M. Corpus Stylistics and Dickens's Fiction[M]. New York: Routledge, 2013.
- [30]年洪东,陈小荷,王东波.现当代文学作品的作者身份识别研究[J].计算机工程与应用,2010,46(04):226-229.
- [31]Oakes, M.P. Statistics for Corpus Linguistics[M]. Edinburgh: Edinburgh University Press, 1998.
- [32]彭炫.当代西方文体学研究方法论略[J].修辞学习,2003(06):11-12.
- [33]Sebastian M. Rasinger. Quantitative Research in Linguistics.[M]. London: Bloomsbury Academic, 2013
- [34]时季.聚类分析方法在文学作品风格比较中的应用——以毕飞宇、苏童小说的比较分析为例[J].文教资料,2017(Z1):19-22.
- [35]施建军.基于支持向量机技术的《红楼梦》作者研究[J].红楼梦学刊,2011(05):35-52.



- [36]施建军.计量文体学导论[M].北京:北京大学出版社, 2016.
- [37]宋明媚,林丽清.基于卡方齐性检验的网络小说抄袭的判定研究[J].统计与管理,2017(07):61-63.
- [38]宋沁潞.金庸小说语言研究[D].山东大学,2008.
- [39]尚永亮.数据库、计量分析与古代文学研究的现代化进程[J].文学评论,2007(06):187-190.
- [40]王景丹.从句频分析看八位剧作家的风格异同[J].修辞学习,2003(04):28-29.
- [41]吴礼权.从统计分析看“简约”与“繁丰”的修辞特征及其风格建构的原则[J].修辞学习,2003(02):18-20.
- [42]吴礼权.平淡风格与绚烂风格的计算统计研究[J].云南师范大学学报(哲学社会科学版),2004(02):42-46.
- [43]吴礼权.庄重风格与幽默风格的计算统计研究[J].渤海大学学报(哲学社会科学版),2004(05):100-104.
- [44]王少康,董科军,阎保平.基于语句节奏特征的作者身份识别研究[J].计算机工程,2011,37(09):4-5+8.
- [45]魏巍,向阳,陈千.中文文本情感分析综述[J].计算机应用,2011,31(12):3321-3323.
- [46]肖天久,刘颖.基于聚类和分类的金庸与古龙小说风格分析[J].中文信息学报,2015,29(05):167-177.
- [47]肖天久,刘颖.《红楼梦》词和 N 元文法分析[J].现代图书情报技术,2015(04):50-57.
- [48]杨基栋.通过统计虚词使用频率检验作品的时期和作者[D].华东师范大学,2013.
- [49]杨建军.定量分析法在中国现当代文学研究中的运用[J].厦门大学学报(哲学社会科学版),2016(04):35-42.
- [50]杨群英.统计分析法对文书作者推定研究中的文体特征提取[J].中国司法鉴定,2006(02):36-39.
- [51]余韵.巴金前后期小说的计量风格学研究[D].华中师范大学,2017.
- [52]张保富,施化吉,马素琴.基于 TFIDF 文本特征加权方法的改进研究[J].计算机应用与软件,2011,28(02):17-20.
- [53]张继东.金庸小说艺术研究[D].南京师范大学,2003.
- [54]詹菊红,蒋跃.语言计量特征在译者身份判定中的应用——以《傲慢与偏见》的两个译本为例[J].外语学刊,2016(03):95-101.
- [55]张京楣.基于统计方法的文本风格分析研究[D].山东大学,2012.
- [56]周立柱,贺宇凯,王建勇.情感分析研究综述[J].计算机应用,2008(11):2725-2728.



- [57]张威.莎士比亚戏剧汉译定量分析研究[D].上海外国语大学,2014.
- [58]张优.《德伯家的苔丝》艺术效果的鉴赏新模式——基于语料库检索技术的文本分析[J].小说评论,2009(S2):160-162.
- [59]Zhao Y, Zobel J. Searching with Style: Authorship Attribution in Classic Literature[C]. In: Proceedings of the 30th Australasian Computer Science Conference (ACSC'07). Darlinghurst: Australian Computer Society, 2007: 59-68.
- [60]张运良,朱礼军,乔晓东,张全.基于句类特征的作者写作风格分类研究[J].计算机工程与应用,2009,45(22):129-131+223.
- [61]赵妍妍,秦兵,刘挺.文本情感分析[J].软件学报,2010,21(08):1834-1848.
- [62]曾毅平,朱晓文.计算方法在汉语风格学研究中的应用[J].福建师范大学学报(哲学社会科学版),2006(01):14-17.
- .....



## 附 录

### 附 1.各标点符号比重

项目 作品	顿号	逗号	冒号	双引号	单引号	破折号	分号	句号	感叹号	问号	省略号
书剑恩仇录	1.48%	53.61%	9.24%	9.96%	0.21%	0.03%	0.03%	20.12%	3.29%	1.60%	0.44%
碧血剑	1.32%	52.57%	9.39%	9.91%	0.29%	0.00%	0.10%	20.20%	3.77%	1.83%	0.62%
射雕英雄传	1.55%	54.58%	9.24%	9.92%	0.38%	0.00%	0.08%	18.07%	4.05%	1.58%	0.55%
雪山飞狐	1.22%	54.10%	8.22%	8.26%	2.42%	0.00%	0.09%	20.46%	3.50%	1.32%	0.41%
神雕侠侣	1.48%	56.20%	8.73%	9.58%	0.23%	0.00%	0.07%	17.29%	4.20%	1.55%	0.68%
飞狐外传	1.46%	54.95%	8.64%	9.96%	0.44%	0.00%	0.09%	17.85%	3.89%	1.99%	0.72%
倚天屠龙记	1.97%	54.53%	8.76%	9.63%	0.51%	0.00%	0.07%	18.03%	4.02%	1.47%	1.00%
鸳鸯刀	1.28%	51.85%	9.91%	11.23%	0.88%	0.17%	0.08%	16.36%	4.44%	2.77%	1.03%
白马啸西风	1.06%	51.47%	9.38%	10.26%	0.20%	0.01%	0.07%	19.61%	4.21%	2.38%	1.36%
天龙八部	1.98%	53.01%	9.24%	10.33%	0.81%	0.00%	0.07%	16.71%	4.33%	1.97%	1.54%
连城诀	1.65%	53.43%	8.00%	9.50%	0.75%	0.04%	0.05%	17.57%	4.55%	2.70%	1.77%
侠客行	1.94%	52.91%	9.16%	10.11%	0.73%	0.01%	0.11%	16.40%	4.73%	1.94%	1.96%
笑傲江湖	1.74%	51.32%	10.16%	10.71%	0.88%	0.01%	0.11%	17.05%	4.58%	1.82%	1.61%
鹿鼎记	1.78%	49.71%	10.81%	11.26%	0.56%	0.00%	0.10%	18.56%	3.98%	1.64%	1.59%
越女剑	1.38%	52.00%	9.41%	9.64%	0.09%	0.00%	0.05%	20.74%	3.35%	2.43%	0.92%
卧龙记	0.57%	39.00%	16.64%	18.71%	0.33%	0.05%	0.07%	13.70%	6.93%	3.56%	0.43%



## 附 2.中科院计算所汉语词性标注集

代码	名称	代码	名称	代码	名称
n	名词	bl	区别词性惯用语	udeng	等 等等 云云
nr	人名	z	状态词	uyy	一样 一般 似的 般
nr1	汉语姓氏	r	代词	udh	的话
nr2	汉语名字	rr	人称代词	uls	来讲 来说 而言 说来
nrj	日语人名	rz	指示代词	uzhi	之
nrf	音译人名	rzt	时间指示代词	ulian	连 (“连小学生都会”)
ns	地名	rzs	处所指示代词	e	叹词
nsf	音译地名	rvz	谓词性指示代词	y	语气词(delete yg)
nt	机构团体名	ry	疑问代词	o	拟声词
nz	其它专名	ryt	时间疑问代词	h	前缀
nl	名词性惯用语	rys	处所疑问代词	k	后缀
ng	名词性语素	ryv	谓词性疑问代词	x	字符串
t	时间词	rg	代词性语素	xx	非语素字
tg	时间词性语素	m	数词	xu	网址 URL
s	处所词	mq	数量词	w	标点符号
f	方位词	q	量词	wkz	左括号, 全角: ( ( [ { 《 【 〔 < 半 角: ([ { <
v	动词	qv	动量词	wky	右括号, 全 角: ) ) ] } 》 】 〗 > 半 角: ) ] { >
vd	副动词	qt	时量词	wyz	左引号, 全角: “ ‘ 『
vn	名动词	d	副词	wyy	右引号, 全角: ” ’ 』
vshi	动词“是”	p	介词	wj	句号, 全角: 。
vyou	动词“有”	pba	介词“把”	ww	问号, 全角: ? 半角: ?
vf	趋向动词	pbei	介词“被”	wt	叹号, 全角: ! 半角: !
vx	形式动词	c	连词	wd	逗号, 全角:, 半角: ,
vi	不及物动词(内 动词)	cc	并列连词	wf	分号, 全角: ; 半角: ;
vl	动词性惯用语	u	助词	wn	顿号, 全角: 、
vg	动词性语素	uzhe	着	wm	冒号, 全角: : 半角: :
a	形容词	ule	了喽	ws	省略号, 全角: ..... ...
ad	副形词	uguo	过	wp	破折号, 全角: —— — — —— 半角: --- ---
an	名形词	udel	的底	wb	百分号千分号, 全 角: % ‰ 半角: %
ag	形容词性语素	ude2	地	wh	单位符号, 全角: ¥ \$ £ ° ℃ 半角: \$
al	形容词性惯用语	ude3	得		
b	区别词	usuo	所		



### 附 3.各类实词比重

项目 作品	名词	动词	形容词	数词	量词	代词
书剑恩仇录	22.36%	21.12%	2.48%	4.35%	5.98%	6.85%
碧血剑	21.69%	20.74%	2.67%	4.24%	5.84%	7.55%
射雕英雄传	20.73%	20.14%	2.80%	3.85%	5.47%	7.46%
雪山飞狐	20.60%	19.71%	2.53%	4.30%	5.66%	8.05%
神雕侠侣	20.26%	19.39%	2.92%	3.78%	5.11%	7.61%
飞狐外传	20.93%	19.78%	2.77%	4.48%	5.79%	7.49%
倚天屠龙记	20.74%	19.06%	2.82%	3.88%	5.23%	7.59%
鸳鸯刀	21.53%	19.07%	2.49%	4.90%	6.71%	7.09%
白马啸西风	20.46%	19.80%	2.84%	4.40%	5.83%	9.09%
天龙八部	20.55%	19.64%	2.87%	3.67%	5.17%	8.38%
连城诀	19.41%	20.53%	2.93%	4.13%	5.35%	8.98%
侠客行	19.32%	19.79%	2.77%	3.94%	5.29%	8.59%
笑傲江湖	20.28%	19.33%	2.75%	3.96%	5.68%	8.44%
鹿鼎记	21.81%	20.68%	2.68%	3.83%	5.74%	8.32%
越女剑	23.64%	18.74%	2.48%	4.52%	5.97%	6.39%
卧龙记	20.32%	17.78%	3.36%	3.69%	6.92%	8.26%



#### 附 4. 各类虚词比重

项目 作品	副词	介词	连词	助词	拟音词
书剑恩仇录	6.94%	2.82%	1.85%	4.87%	0.12%
碧血剑	7.54%	2.76%	1.98%	5.40%	0.15%
射雕英雄传	6.16%	2.76%	2.11%	5.34%	0.14%
雪山飞狐	7.83%	2.78%	2.18%	5.56%	0.19%
神雕侠侣	7.88%	2.58%	2.53%	5.75%	0.16%
飞狐外传	7.52%	2.50%	2.42%	5.60%	0.16%
倚天屠龙记	7.35%	2.40%	2.53%	5.37%	0.14%
鸳鸯刀	7.15%	2.20%	2.15%	5.67%	0.26%
白马啸西风	7.79%	2.60%	2.32%	7.95%	0.19%
天龙八部	7.88%	2.37%	2.40%	5.66%	0.19%
连城诀	8.04%	2.47%	2.21%	6.58%	0.24%
侠客行	8.13%	2.34%	2.36%	5.66%	0.21%
笑傲江湖	7.96%	2.44%	2.52%	5.63%	0.17%
鹿鼎记	7.45%	2.35%	2.05%	5.94%	0.16%
越女剑	6.55%	2.56%	1.94%	5.55%	0.22%
卧龙记	9.78%	2.23%	3.72%	6.18%	0.19%

#### 附 5.16 部小说中各类情感倾向的句子数与句子总数

情感倾向 作品	积极	消极	中性	总句数
书剑恩仇录	6251	2003	9048	17302
碧血剑	5408	1662	7279	14349
射雕英雄传	11538	2992	13778	28308
雪山飞狐	1700	481	2138	4319
神雕侠侣	12150	3409	13389	28948
飞狐外传	5570	1575	6527	13672
倚天屠龙记	12142	3389	13630	29161
鸳鸯刀	432	125	568	1125
白马啸西风	1006	306	1004	2316
天龙八部	15786	4484	17384	37654
连城诀	3327	844	3513	7684
侠客行	4622	1282	5231	11135
笑傲江湖	12205	3713	15209	31127
鹿鼎记	16037	5375	20078	41490
越女剑	189	76	313	578
卧龙记	6482	2673	5929	15084



## 致 谢

时光匆匆，转眼就将和桂子山道别了。在这里，我度过了人生中最宝贵，也最难忘的时光。在每一条我走过的路上，都充满着无数的回忆。感谢桂子山的树与阳光，感谢每一位给与我帮助与支持的人。

感谢我的导师沈威，他是一位让人敬佩的老师。因为有了他悉心的指导和帮助，我才能顺利完成我的论文写作。还记得凌晨时总会收到来自他的反馈与指导，还有不论何时都能收到他及时的回复，等等这些都令我惭愧汗颜。沈老师不是闲人，他有繁重的科研与本科教学任务，也有自己的私人生活，他把时间掰开揉碎了，不放过一刻钟。他总对我们说，抓紧时间。我从没有这么深刻地意识到，这句话是多么的形象呀。

感谢我的室友们，她们是一群可爱的人。我们朝夕相对，也相伴而行。我们总在阳光下一起晒衣服，常纠结于一日三餐吃什么；也相约去图书馆刻苦学习，不时在寝室高谈阔论。感谢她们的陪伴与支持，也感谢她们让我成为更完整的人。

感谢我的父母与亲人，他们总是默默地支持我，最爱问我吃饭了吗。我爱说时，会倾听会鼓励；我沉默时，不忘问吃过饭了吗。他们从不给我压力，总说，想做什么就去做吧。谢谢他们，让我远行，也让我永远有回家的路。

最后，感谢每一位谆谆教导的师长，每一位耐心陪伴的朋友，每一位在我成长路上留下身影的人。感谢！