doi:10.3772/j.issn.2095-915x.2025.04.005

开放科学（资源服务）标识码（OSID）

数字人文视域下文学作品风格计量分析与比较研究
——以金庸古龙小说为例

任东海[1,2] 刘忠宝[1] 张博文[1]

1. 北京语言大学信息科学学院 北京 100083；
2. 河南经贸职业学院人工智能学院 郑州 450000

摘要：[目的 / 意义] 解决文学作品风格计量分析中缺乏系统性指标体系、指标测度方法适用性差等问题，为数字人文视域下文学作品风格的科学计量分析提供支撑。[方法 / 过程] 首先，基于文学作品内容构成粒度，从微观到宏观选取词汇、句子、段落、篇章 4 个层面的 10 项核心指标，构建两级指标体系；其次，对指标测度方法进行优化，包括采用变异系数替代标准差测度离散度以解决均值偏差问题，定义词频熵并结合外部词频表测度作品通俗性，设计分组匹配法对比作品词汇丰富度以控制篇幅干扰；最后，基于金庸古龙武侠小说语料库，通过统计分析与机器学习验证该指标体系的有效性。[局限] 当前研究指标体系仅以文本语言特征为主，未来可考虑整合情节网络、人物关系等非语言特征，构建全景式文学作品风格分析指标体系。[结果 / 结论] 研究发现古龙与金庸作品风格存在多方面显著差异，且发现了古龙创作风格的演变规律，证实了所构建的指标体系在风格分析、作者识别与作品聚类中的应用有效性。

关键词：作品风格分析；指标测度体系；机器学习；假设检验；数字人文

中图分类号：G35

A Quantitative Analysis and Comparative Study on the Styles of Literary Works from the Perspective of Digital Humanities: A Case Study of Jin Yong’s and Gu Long’s Novels

REN Donghai[1,2] LIU Zhongbao[1] ZHANG Bowen[1]

1. School of Information Science, Beijing Language and Culture University, Beijing 100083, China;
2. School of Artificial Intelligence, Henan Institute of Economic and Trade, Zhengzhou 450000, China

基金项目 研究阐释党的二十大精神国家社科基金重点项目“大数据时代古籍活化赋能文化自信自强的理论、方法与路径研究”(23AZD047)。

作者简介 任东海（1988-），博士研究生，主要研究方向为自然语言处理和数字人文；刘忠宝（1981-），博士，教授，主要研究方向为数字人文、文化数字化、语言大数据技术，E-mail: zbliu@blcu.edu.cn；张博文（2000-），硕士研究生，主要研究方向为自然语言处理和数字人文。

引用格式 任东海 , 刘忠宝 , 张博文 . 数字人文视域下文学作品风格计量分析与比较研究——以金庸古龙小说为例 [J]. 情报工程 , 2025, 11(4): 50-64.

Abstract: [Objective/Significance] This study aims to address issues such as the lack of a systematic indicator system and poor applicability of indicator measurement methods in the quantitative analysis of literary work styles, providing support for the scientific quantitative analysis of literary work styles from the perspective of digital humanities. [Methods/Processes] Firstly, based on the content composition granularity of literary works, ten core indicators at four levels (vocabulary, sentence, paragraph, and text) from micro to macro are selected to construct a two-level indicator system. Secondly, the indicator measurement methods are optimized, including using the coefficient of variation instead of the standard deviation to measure dispersion to solve the problem of mean deviation; defining word frequency entropy and combining it with external word frequency tables to measure the popularity of works; and designing a group matching method to compare the vocabulary richness of works to control the interference of length. Finally, based on the corpus of Jin Yong's and Gu Long's martial arts novels, the effectiveness of the indicator system is verified through statistical analysis and machine learning. [Limitations] The current research indicator system only focuses on text language features. In the future, non-language features such as plot networks and character relationships can be integrated to construct a panoramic indicator system for literary work style analysis. [Results/Conclusions] The study found that there are significant differences in the styles of Gu Long's and Jin Yong's works in multiple aspects, and discovered the evolution law of Gu Long's creative style, confirming the application effectiveness of the constructed indicator system in style analysis, author identification, and work clustering.

Keywords: Literary Style Analysis; Metric System; Machine Learning; Hypothesis Testing; Digital Humanities

引言

数字人文研究通过量化分析与计算模型揭示文学作品深层结构特征，推动文学批评从经验阐释向数据驱动范式转型。在文学风格分析领域，传统研究多依赖主观文本细读，难以系统刻画作家创作风格的历时演变与流派差异，数字人文范畴下的计算风格学，在用统计方法提取言语计量特征分析文学作品风格方面已取得一定进展。概言之，过往计算风格学着重围绕两个方面展开研究：一是计量特征的测度方法研究，二是计量特征在风格分析及作者识别等方面的应用研究，前者如 Guiraud[1]、陆芸 [2]、王泓等 [3] 对词汇丰富度的测度方法研究，张京楣 [4] 对文本通俗性（字频熵）、文本从众性（聚类度）、文本节奏（句子离散度）等的测度方法研究等；后者如 Burrows[5] 通过词汇频率及其出现模式分析文章叙事风格，Grieve[6] 对作者身份归属中常用的 39 种文本计量特征的比较研究，肖天久等 [7] 对金庸古龙小说在文本从众性、句子破碎度、n 元文法等方面的风格对比研究，涂梦纯等 [8] 选用段落、句长、标点、词类、词汇、颜色词等特征对余华和莫言作品风格差异的研究等。

尽管计算风格学已引入诸多计量指标辅助文学作品风格分析，但仍存在两大问题：其一，缺乏一个体系性的计量特征选取框架，以系统整合词汇、句法、篇章等多粒度特征的协同作用；其二，一些关键指标的测度方法有待完善——句长离散度的测度方法（标准差）易受均值干扰 [9]、文本通俗性测度依赖复杂字频熵 [4,10]、词汇丰富度测度受文本篇幅影响严重 [2] 等，制约跨体裁可比性和应用广泛性。

针对计算风格学既有研究存在的上述问题，本文立足数字人文视角，尝试四方面工作突破：（1）构建体系：按文学作品内容粒度（词汇、

doi:10.3772/j.issn.2095-915x.2025.04.005

句子、段落、篇章）构建两级指标体系，揭示多维度计量特征的协同效应；（2）方法革新：引入变异系数替代标准差测度离散度，提出词频熵整合外部语料库测度通俗性，设计分组匹配对比法消除词汇丰富度测度的篇幅偏差；（3）实例验证：基于超过 1400 万字的武侠小说语料库，量化对比揭示金庸、古龙创作风格差异，为文学风格计量及流派演化研究提供可复用的分析框架；（4）应用推广：结合机器学习方法验证指标体系在作品作者识别与风格聚类中的有效应用。

本文不仅深化了武侠作品亚类型内部风格分异认知，更探索数字人文方法在经典文学阐释中的实践路径，为文化遗产的智能解析提供方法论参考。

1　文学作品风格分析指标体系

1.1　指标体系构建

指标体系包含两级指标，其中一级指标依据文学作品内容构成的粒度大小，从微观到宏观依次选取词汇、句子、段落、篇章 4 个层面特征。词汇特征关注词汇外在结构（“词有多长”）和内在分布（“词有多杂”）两个方面，体现为词汇长度和词汇丰富度 2 个二级指标，分别反映了词汇的形式复杂度和多样性。句子特征关注语法结构、信息密度、节奏控制和表达多样性等 4 个维度，体现为句法复杂性、句子长度、句长离散度、句中停顿次数等 4 项二级指标，系统解析作品的句子风格特征。段落特征关注作品叙事情感风格及节奏稳定性等，通过段落长度和段长离散度 2 个二级指标量化分析。篇章特征关注作品的“大众化—个性化”表达风格，体现为通俗性和从众性 2 个二级指标，分别反映作品的词汇常用度与规约化语言使用倾向。以上共计 10 项二级指标，各级指标对应关系及测度方法如表 1 所示。

表 1　文学作品风格分析指标体系

	一级指标	二级指标	测度方法
文学作品风格分析特征体系	词汇特征	词汇丰富度	类符形符比
		词汇长度	平均词长
	句子特征	句法复杂性	平均依存距离
		句子长度	平均句长
		句长离散度	句长变异系数
		句中停顿次数	句子破碎度
	段落特征	段落长度	平均段长
		段长离散度	段长变异系数
	篇章特征	通俗性	词频熵
		从众性	聚类度

1.2　指标定义与计算

1.2.1　词汇特征

（1）词汇长度

词汇长度用平均词长来测度，平均词长（ALW）是指文本中出现的所有词 v 长度之和与所有词的数量之和的比率 [4]，如式（1）所示。

$$ALW = \frac{\sum_{v} length(v)}{\sum_{type} count(type(v))} \qquad (1)$$

（2）词汇丰富度

目前，文本词汇丰富度测度方法 [3, 6] 众多，但均不同程度地受文本长度影响。类符形符比

（TTR）作为测度词汇丰富度的经典方法，计算简洁且含义直观，其计算方法为：

$$TTR = V / \sqrt{N} \tag{2}$$

式（2）中，类符数（V）表示作品中不重复词汇的总数，形符数（N）为作品总词数。

为了消除文本长度对词汇丰富度测度的影响，本文后续对比实验设计了一种分组匹配对比方法，确保同组内作品平均字数均衡，以控制作品篇幅偏差。

1.2.2　句子特征

（1）句法复杂性

句子句法复杂性反映的是句子的结构层次与语法多样性，参考梁君英等[11]研究，本文采用平均依存距离来测度句法复杂性，在依存句法分析方法里，一个句子被看作一个线性的词串，句子中词与词之间是有联系的，依存距离是指句子中两个有依存句法关系的词之间的线性距离，依存距离可以判别一个句子的理解难度。句子的平均依存距离可以消除句子长度带来的影响，一个句子的平均依存距离越大，理解难度越大[11]。

（2）句子中停顿次数

句子中停顿次数统计句中点号（逗号、顿号、冒号、分号）或自然停顿的位置，反映句子的信息组织方式和表达流畅度。借鉴肖天久等[7]研究，用句子破碎度来测度句中停顿次数。一个文学作品的句子破碎度是指所有句子的平均破碎度，计算方法如式（3）所示。

$$\text{句子破碎度} = \frac{\text{句中点号总数}}{\text{总句数}} \tag{3}$$

（3）句子长度

句子长度反映单句的信息密度和阅读难度，采用平均句子长度（简称平均句长）作为测度方法，其计算公式如式（4）所示。

$$\bar{L} = \frac{1}{n}\sum_{i=1}^{n} l_i \tag{4}$$

其中，l_i 表示第 i 个句子的长度，n 为作品句子总数，$\bar{L}$ 为平均句长。

（4）句长离散度

句长离散度通过量化句子长度的分布差异，反映作品行文节奏的变化幅度与情感张力强度。其数值越低，表明句长分布越趋近均质化，行文节奏平稳且具有规律性；其数值越高，则句长波动剧烈，作品表现力更为丰富[4]。

①测度方法优化

过往研究[4, 7-8]多采用句长标准差测度离散度，如式（5）所示。

$$S = \sqrt{\frac{1}{n}\sum_{i}\left(l_i - \bar{l}\right)^2} \tag{5}$$

其中，l_i 为第 i 句的句长，$\bar{l}$ 为平均句长，S 为标准差。然而，标准差的数值受数据均值与量纲影响[9]，在比较不同尺度或量纲的数据集时易产生偏差（如小说与短诗间的句长对比）。为此，本文引入变异系数（CV）作为句长离散度的稳健测度方法，其通过计算标准差与均值的比消除尺度影响，如式（6）所示。

$$CV = \frac{S}{\bar{l}} \times 100\% \tag{6}$$

②方法对比验证

为验证变异系数的有效性，选取两段典型文本进行对比分析：

文本 1（李清照《如梦令》节选）：“试问卷帘人 | 却道海棠依旧 | 知否 | 知否 | 应是绿肥红瘦”

计算得：$\bar{l}_1 = 4.2$，S_1=1.83，CV_1=44%。

doi:10.3772/j.issn.2095-915x.2025.04.005

文本 2（朱自清《荷塘月色》节选）：“月光如流水一般 | 静静地泻在这一片叶子和花上 | 薄薄的青雾浮起在荷塘里 | 叶子和花仿佛在牛乳中洗过一样 | 又像笼着轻纱的梦”

计算得：$\bar{l}_2=10.4$，S_2=2.94，CV_2=28%。

尽管文本 2 的标准差（S_2=2.94）高于文本 1（S_1=1.83），但其变异系数（CV_2=28%）明显低于文本 1（CV_1=44%）。此结果与文本实际呈现节奏感受一致：文本 1 因短长句交替（如“知否 | 知否”与“应是绿肥红瘦”）呈现强烈跌宕感，而文本 2 长句为主且长度相近，节奏更为平缓。标准差因受均值干扰（$\bar{l}_2 \gg \bar{l}_1$）导致误判，而变异系数通过标准化处理准确捕捉节奏差异。

相较于传统标准差测度，变异系数能够消除平均句长对离散度评估的干扰，可用于跨文学作品类型比较（如诗歌与散文）。

1.2.3 段落特征

（1）段落长度

段落长度用平均段落长度（简称平均段长）来测度，平均段长定义为作品所有段落的长度均值，其计算公式如式（7）所示。

$$\bar{P}=\frac{1}{m}\sum_{i=1}^{m}p_i \tag{7}$$

其中，p_i 表示第 i 个段落的长度，m 为作品段落总数，$\bar{P}$ 为平均段长。该指标反映作品段落的信息承载密度与结构化特征。

（2）段长离散度

这里延续式（6）句长离散度的测度逻辑，采用段落长度的变异系数测度段长离散度，计算公式如式（8）所示。

$$CV_p=\frac{\sqrt{\frac{1}{n}\sum_i\left(p_i-\bar{p}\right)^2}}{\bar{p}} \tag{8}$$

其中，CV_p 表示段长离散度，p_i 为单个段落长度，$\bar{p}$ 为平均段落长度。离散度低则作品段落长度趋近一致，结构规整；离散度高则段落长短交错，节奏多变。

1.2.4 篇章特征

（1）通俗性（Popularity）

通俗性表征作品词汇与日常高频词的契合程度，基于张景祥等[10] 字频熵理论，本文提出词频熵测度方法，其计算公式如式（9）所示。

$$Z=-k\sum p_v\log p_v' \tag{9}$$

其中，p_v 为词汇 v 在作品中的频率，p_v' 为外部参照语料库（本文采用 *BCC* 汉语词频表[12]）中 v 的频率（冷僻词默认频数为 1），k>0 为一修正常数（调整尺度）。Z 值低，作品高频词占比高，通俗性强（如通俗小说）；Z 值高，作品冷僻词或专业术语密集，通俗性弱（如学术论文）。

词频熵通过引入外部参照词频表（跨文本空间对比），增强对作品词汇社会常用度[13] 的刻画能力。

（2）从众性（Conventionality）

从众性衡量作品使用规约化语言单元（包括成语、谚语、诗词名言、固定搭配等）的倾向性，借鉴已有研究[4, 7]，采用聚类度（C）测度从众性，聚类度定义为成词率（WFR）与词汇平均长度（ALW）的乘积，计算公式如式（10）所示。

$$C=WFR\times ALW \tag{10}$$

其中，词汇平均长度（ALW）如式（1）所示，成词率（WFR）如式（11）所示。

$$WFR=\frac{\sum length(v)}{\sum length(c)} \tag{11}$$

式（11）中，分子为作品中所有词汇 v 长度之和，分母为作品中所有字符 c 长度之和。

高聚类度作品依赖规约化长词（如“刻舟求剑”“人工智能”），体现作者对固定表达的熟练运用；低聚类度作品偏好短词或自由组合新词（如网络新词“躺平”“绝绝子”），反映语言创新性或非正式性。

2 实验与分析

本文以实验验证文学作品风格分析指标体系的有效性，为更好体现指标体系的风格刻画能力，选取同类型的金庸与古龙武侠小说作对比分析，并进一步探索了指标体系在作品作者识别及风格聚类分析等方面的应用效果。

2.1 实验语料库

金庸与古龙均为新派武侠小说家的代表人物，两人的小说影响深远[7]，本文构建的语料库包括金庸全部 14 部中长篇小说及古龙不同创作时期、不同篇幅的 16 部代表作。语料库共计 30 部小说，字数规模超 1400 万，作品名称、大致创作年份及字数统计如表 2 所示。

表 2 语料库作品统计

金庸小说	创作年份	总字数	古龙小说	创作年份	总字数
书剑恩仇录	1955	441307	浣花洗剑录	1964	476412
碧血剑	1956	378230	大旗英雄传	1965	485686
射雕英雄传	1958	785527	名剑风流	1966	530774
雪山飞狐	1959	115605	武林外史	1966	573495
神雕侠侣	1960	850831	绝代双娇	1969	670436
飞狐外传	1961	383914	楚留香系列	1970	974122
白马啸西风	1961	58714	多情剑客无情剑	1970	372953
鸳鸯刀	1961	29881	欢乐英雄	1971	269933
倚天屠龙记	1962	837475	流星蝴蝶剑	1971	219891
连城诀	1963	195520	陆小凤系列	1973	808279
侠客行	1965	315349	七种武器	1973	455670
天龙八部	1966	1054089	三少爷的剑	1973	192960
笑傲江湖	1969	838303	萧十一郎	1973	351597
鹿鼎记	1972	1031865	天涯明月刀	1974	147602
			边城浪子	1975	181966
			圆月弯刀	1977	48795
总计		7316610	总计		6760571

2.2 指标体系有效性

本节对所选取的金庸古龙小说依据表 1 的指标体系进行作品风格多层面对比分析，使用 Python 进行文本处理及指标计算，通过箱型图展示对比指标数据分布情况，采用统计假设检验方法进行指标均值差异显著性判断，显著性水平取值 $\alpha=0.05$[9]。

2.2.1 词汇特征对比分析

小说词汇特征对比词汇丰富度与词汇长度。

金庸古龙小说词汇特征对比结果如图 1 所示：

doi:10.3772/j.issn.2095-915x.2025.04.005

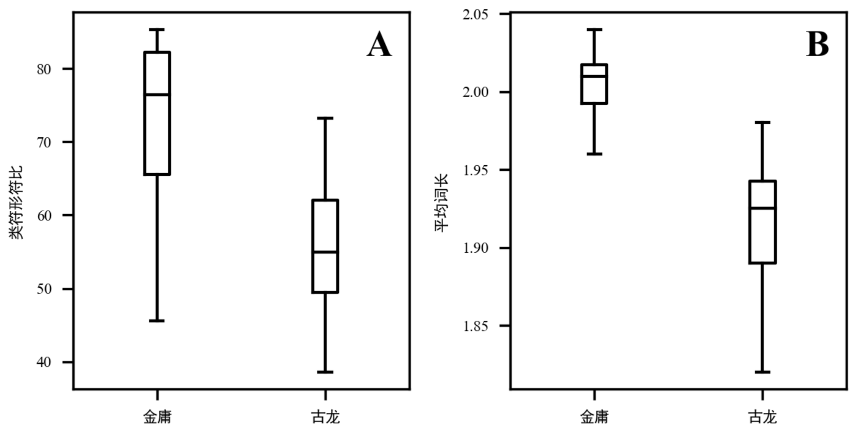

图 1　金庸古龙小说词汇特征对比

图 1 初步显示金庸整体的词汇丰富度比古龙更高，平均词长更长。下面通过分组匹配对比与统计假设检验，进一步系统对比两者词汇特征差异，结果如下：

（1）词汇丰富度（TTR）对比

控制篇幅偏差：鉴于 TTR 与作品总字数呈强正相关 [6]（在本次研究语料库上，两者皮尔逊相关系数 [9] r=0.74），不同作品词汇丰富度测度会受其篇幅干扰，故本文按篇幅字数对作品分组匹配对比，如表 3 所示，确保同组内作品平均字数均衡，消除作品篇幅干扰，增强结果可信度。

表 3　作品词汇丰富度分组匹配对比

分组	金庸小说	TTR	字数	组平均TTR	组平均字数	古龙小说	TTR	字数	组平均TTR	组平均字数
一	鸳鸯刀	45.61	29881	46.96	44298	圆月弯刀	38.58	48795	38.58	48795
	白马啸西风	46.3	58714							
二	雪山飞狐	63.07	115605	63.07	115605	天涯明月刀	50.26	147602	50.26	147602
三	连城诀	64.18	195520	64.18	195520	边城浪子	42.35	181966	44.98	198272
						三少爷的剑	45.44	192960		
						流星蝴蝶剑	47.15	219891		
四	侠客行	69.56	315349	69.56	315349	欢乐英雄	56.45	269933	54.28	310765
						萧十一郎	52.1	351597		
五	碧血剑	78.13	378230	78.13	378230	多情剑客无情剑	52.13	372953	52.13	372953
六	飞狐外传	75.79	383914	75.79	383914	七种武器	53.41	455670	53.41	455670
七	书剑恩仇录	80.31	441307	80.31	441307	浣花洗剑录	71.27	476412	64.59	574361
						大旗英雄传	66.95	485686		
						名剑风流	61.76	530774		
						武林外史	62.75	573495		
						绝代双娇	60.19	670436		
八	射雕英雄传	85.34	785527	85.34	785527	陆小凤系列	61.01	808279	61.01	808279
九	倚天屠龙记	82.86	837475	80.31	922513	楚留香系列	73.2	974122	73.2	974122
	笑傲江湖	73.16	838303							
	神雕侠侣	84.88	850831							
	鹿鼎记	77.01	1031865							
	天龙八部	83.64	1054089							

进行配对样本 t 检验，组内对比显示，金庸小说 TTR 均值显著高于古龙（$\bar{X}_{金庸}=71.5\pm11.3$, $\bar{X}_{古龙}=54.7\pm9.8$，p<0.01），表明金庸作品词汇多样性优势独立于篇幅效应。

（2）平均词长对比

进行独立样本 t 检验，结果显示金庸小说平均词长显著长于古龙（$\bar{X}_{金庸}=2.01\pm0.02$, $\bar{X}_{古龙}=1.92\pm0.04$，p<0.01），且数据分布更集中（图 1B 箱体高度差异）。

综合以上，金庸作品在词汇丰富度与词汇形态复杂度上均显著高于古龙作品，其创作语言风格倾向于使用更多样、更长的词汇。

2.2.2　句子特征对比分析

小说句子特征对比句法复杂度、句子长度、句长离散度、句中停顿次数 4 项指标，各指标测度值箱线图如图 2 所示，结合统计假设检验，核心结果如下：

（1）均值差异显著性检验（独立样本 t 检验）

金庸小说在平均句子长度、平均句中停顿次数显著更高，古龙小说则在平均句法复杂度显著更高，两者在句长离散度无显著差异。

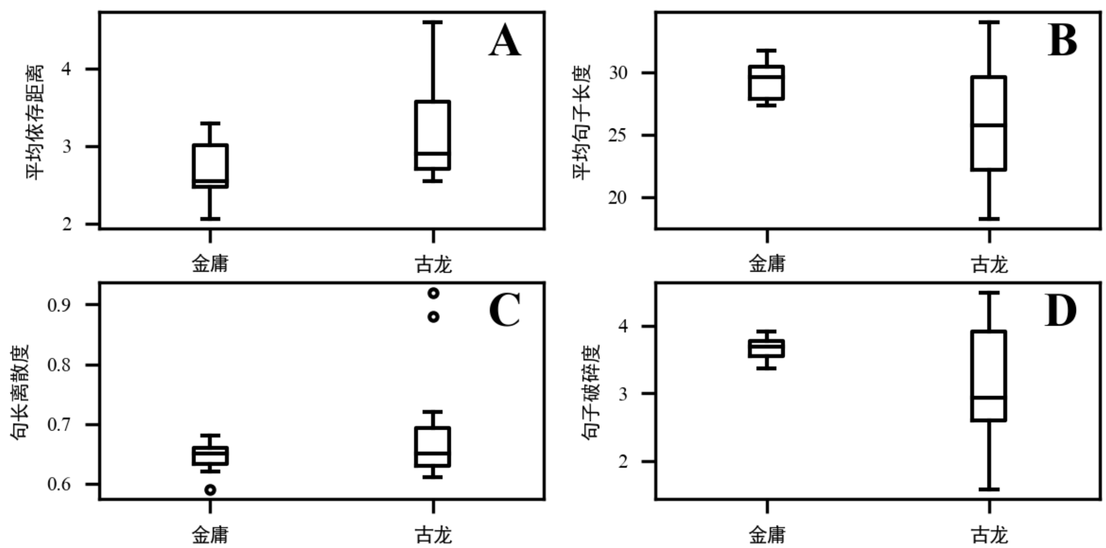

图 2　金庸古龙小说句子特征对比

（2）各自作品风格稳定性分析

古龙小说在句子特征 4 项指标上的箱线图高度均明显大于金庸，如图 2 所示，表明其创作风格动态多变，不同作品间句子风格差异大，体现为句法结构灵活性与节奏调控的多样性。

相较而言，金庸小说句子特征各指标分布集中（箱线图高度矮），反映其风格稳定性与创作一致性。

综合而言，古龙小说通过高句法复杂度与强句子风格异质性突显语言创新性，而金庸小说以长句高频停顿与风格统一性强化叙事密度。二者差异体现了武侠文学流派内部表达范式分化。

2.2.3　段落特征对比分析

小说段落特征对比段落长度与段长离散度两项指标，指标测度值如图 3 所示。

通过箱线图可视化与统计假设检验，对比

doi:10.3772/j.issn.2095-915x.2025.04.005

两者段落风格差异，核心结果如下：

（1）均值差异显著性检验（独立样本 t 检验）

金庸小说平均段长显著长于古龙（$\bar{X}_{金庸}=123.8\pm17.04$, $\bar{X}_{古龙}=32.4\pm5.22$），表明金庸偏好信息密集型段落结构，强化叙事连贯性。

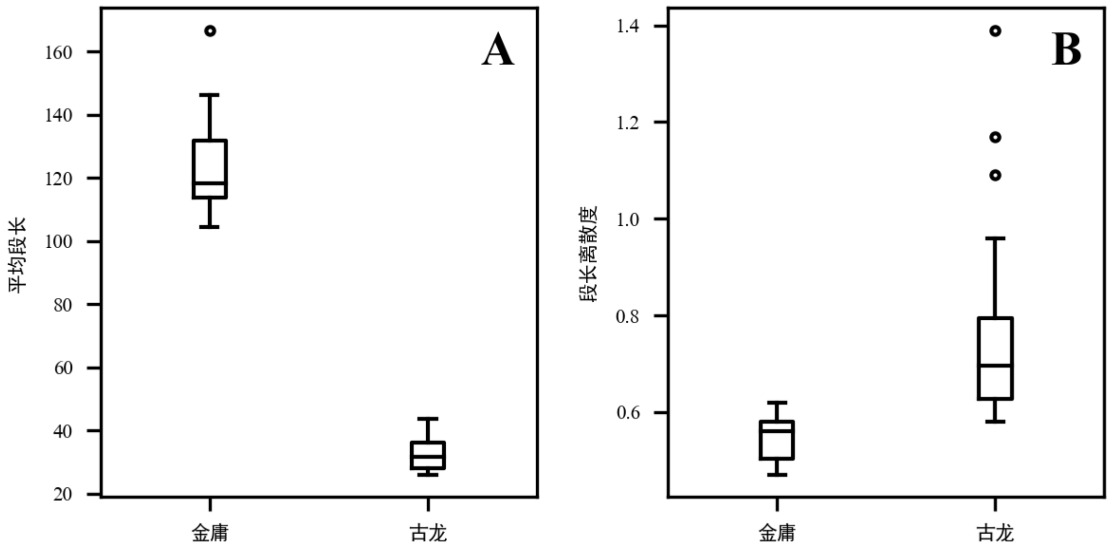

图 3　金庸古龙小说段落特征对比

古龙小说段长离散度显著高于金庸（$\bar{X}_{古龙}=0.79\pm0.23$, $\bar{X}_{金庸}=0.55\pm0.04$），反映其行文段落节奏动态调控倾向。

（2）各自作品风格稳定性分析

金庸不同小说间平均段长差异更大，表明其不同作品间段落密度策略差异。

古龙不同小说间段长离散度差异更大，体现其跨作品段落节奏灵活多变。

综合而言，金庸小说以长段落稳定性构建叙事纵深，而古龙通过高段长离散度实现节奏张力，二者差异体现武侠文学中“宏大叙事”与“碎片化表达”的范式分野。

2.2.4　篇章特征对比分析

小说篇章特征对比通俗性与从众性两项指标，指标测度值如图 4 所示。

通过箱线图可视化与统计假设检验，对比二者篇章风格差异，核心结果如下：

（1）均值差异显著性检验（独立样本 t 检验）

通俗性（词频熵）方面，古龙小说词频熵显著低于金庸（$\bar{X}_{金庸}=18.22\pm0.52$, $\bar{X}_{古龙}=17.00\pm0.65$），表明古龙创作更倾向使用高频日常词汇，作品通俗性更强。

从众性（聚类度）方面，金庸小说显著高于古龙（$\bar{X}_{金庸}=1.71\pm0.02$, $\bar{X}_{古龙}=1.66\pm0.04$），反映其更多运用规约化语言单元，源于其作品大量引用古典诗词、典故等。

（2）各自作品风格稳定性分析

通俗性方面，图 4A 显示金庸词频熵中存在明显异常值（《白马啸西风》），《白马啸西风》呈现出另类风姿 [14]，体现其创作中偶现的突破性尝试，但整体维持高词频熵特征。

从众性方面，图 4B 显示古龙小说箱线

图数据分布更离散（极差：古龙为 0.14，金庸为 0.07），表明其作品间从众性策略分化多变。

综合而言，古龙作品以强通俗性与灵活从众性贴近大众阅读偏好，而金庸作品以弱通俗性（低频词运用）与高从众性（传统语言继承）构建典雅叙事风格，二者差异映射武侠文学中"通俗化创新"与"经典化重构"的创作路径分野。

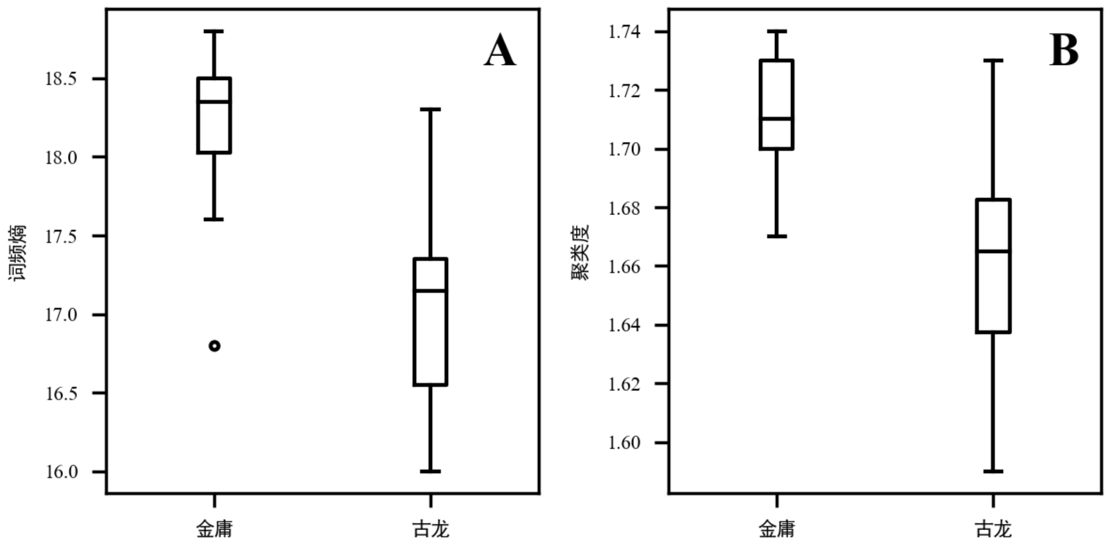

图 4　金庸古龙小说篇章特征对比

以上基于词汇、句子、段落、篇章四维特征的量化分析与统计检验，表明金庸与古龙小说在语言风格与叙事范式上都呈现较大差异性，金庸以系统性风格稳定延续武侠文学经典化传统，而古龙不断突破，开辟通俗化创新路径。二者差异不仅映射武侠亚类型的内部多样性，也为计算风格学提供了量化解析文学风格的典型范例。

2.3　指标体系应用

2.3.1　作者识别与特征筛选

将表 2 中 30 部小说按照作者标签分为两组，两组数据集均随机以 2:1 划分为训练集、测试集，以 10 个二级指标为特征向量输入，采用 logistic 分类器进行作者识别，为了避免结果的偶然性，借鉴王阳阳 [15] 的做法，本文在用 logistic 分类估计的时候一共循环抽取训练集训练并测试 500 次，最终结果评价取 500 次结果的平均值。

参考现有研究 [16-17]，结果评价用 $F1$ 值作为衡量标准，$F1$ 值是精确率（Precision）和召回率（Recall）的调和平均数，其计算如式（12）至（14）所示。

$$\text{Precision}=\frac{TP}{TP+FP} \tag{12}$$

$$\text{Recall}=\frac{TP}{TP+FN} \tag{13}$$

$$F1=\frac{2\times\text{Precision}\times\text{Recall}}{\text{Precision}+\text{Recall}} \tag{14}$$

其中，TP 表示作品作者被识别为金庸，实际也为金庸的数量；FP 表示作品作者被识别为金庸，实际为古龙的数量；FN 表示作品作者被识别为古龙，实际为金庸的作品数量。

doi:10.3772/j.issn.2095-915x.2025.04.005

（1）筛选代表性特征（一级指标）

五组对比实验，对比词汇、句子、段落、篇章4个层面特征对分类器识别准确率的作用，第一组输入全部特征指标，其余四组每组轮流丢弃一维特征指标，最终结果如表 4 所示。

表 4 中，第 0 组 logistic 分类器的 *F*1 值为 0.968，表明本文构建的文学作品风格分析指标体系对作者识别有良好的应用效果。

与第 0 组结果对比，第 1 组（丢弃篇章特征）效果更好，即篇章特征对分类器有一定的干扰性，表明金庸和古龙的部分小说在从众性和通俗性上相近；

表 4　输入不同特征指标分类器作者识别结果

实验组号	指标输入	作者识别结果（*F*1 值）
0	全部特征指标	0.968
1（— 篇章）	段落 + 句子 + 词汇	0.986
2（— 段落）	篇章 + 句子 + 词汇	**0.807**
3（— 句子）	篇章 + 段落 + 词汇	0.979
4（— 词汇）	篇章 + 段落 + 句子	0.978

同理，第 3 组和第 4 组结果好于第 0 组，即在句子和词汇特征上，金庸和古龙的部分小说之间风格也有交叉。这一方面体现了通俗武侠小说所具有的共同风格特点，另一方面体现了作者作品创作风格的发展变化。

但是，第 2 组（丢弃段落特征）分类器识别能力明显下降，这表明段落特征是金庸和古龙小说风格的最主要区分特征，即金庸与古龙小说在段落特征上都具有独特的自我风格。

（2）筛选代表性指标（二级指标）

进一步考察段落特征的 2 个二级指标（段落长度与段长离散度），在上文第 2 组实验基础上增加两组对比实验，第 5 组加入段长离散度指标，第 6 组加入平均段长指标，分类器结果如表 5 所示。

表 5　段落特征指标作用检验结果

实验组号	指标输入	作者识别结果（*F*1 值）
0	全部特征指标	0.968
2	篇章 + 句子 + 词汇	0.807
5	篇章 + 句子 + 词汇 + 段长离散度	0.848
6	篇章 + 句子 + 词汇 + 平均段长	0.962

与第 2 组对比，第 5、6 组识别效果都有提升，尤其是第 6 组效果明显提升；与第 0 组对比，第 5、6 组识别效果都下降，尤其是第 5 组效果明显下降。表明段落长度和段长离散度均为金庸、古龙小说风格的代表性指标，且段落长度对作者识别作用更大。

2.3.2　聚类与可视化分析

本文采用层次聚类法[18-21]，按照作品风格指标测度值，基于欧式距离以及 Ward 方差最小化算法，对语料库中 30 部作品进行聚类，分析不同作品风格相似性。

结合上节研究结果，段落特征作为金庸古龙作品的代表性特征，分别以全部特征指标、非段落特征指标、段落特征指标进行三次聚类，分析作品不同层面风格特点，进一步挖掘两者作品风格差异性与相似性，并对金庸、古龙小说段落特征分布进行可视化呈现和多视角分析，增强结论可信度。

（1）全部特征指标聚类分析

小说全部特征聚类结果如图 5 所示。

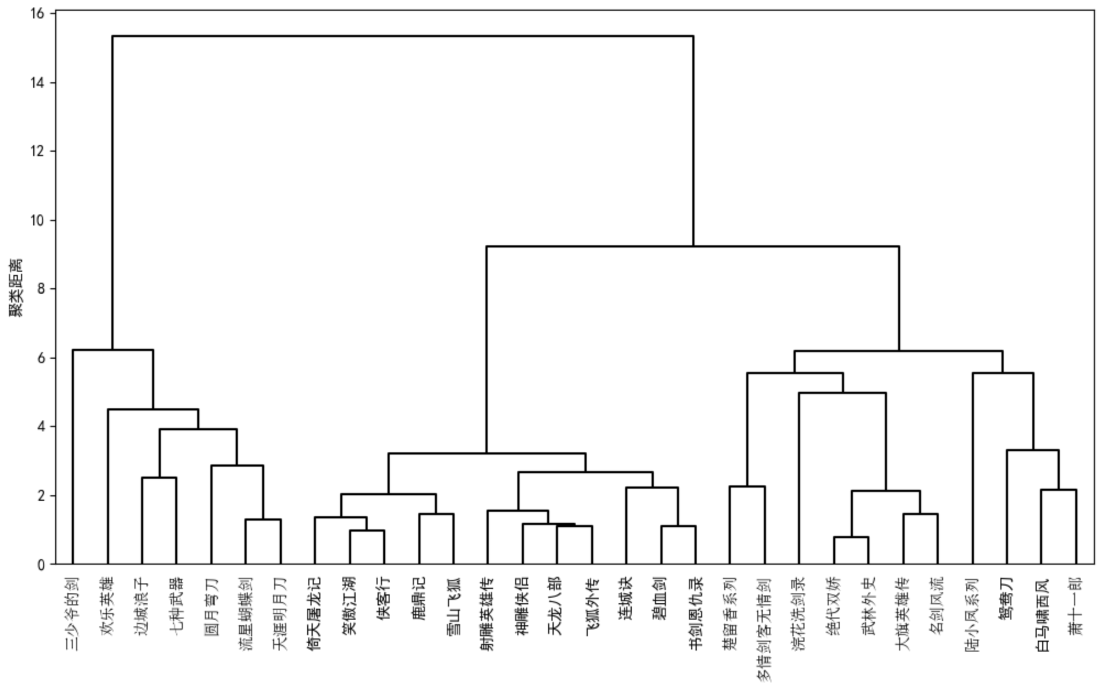

图 5　小说全部特征指标聚类结果

图 5 横轴是小说作品名称，其中金庸作品名称黑体加粗显示（下同），纵轴是作品聚类距离，距离越小，特征相似度越高。

30 部小说可聚为 3 类。图上左侧第一类是属于古龙的 7 部小说，中间第二类是金庸的 12 部小说，第三类是古龙的另外 9 部和金庸的 2 部（《鸳鸯刀》和《白马啸西风》）。

从纵轴聚类距离看作品相似度，第二类金庸主要作品间的高相似度，再次体现金庸创作系统风格稳定与连贯性，其余两类中古龙作品间相似度相对低，尤其是《三少爷的剑》《陆小凤系列》与其他作品距离大，再次体现了古龙的作品创作风格独特多变，他不断突破传统武侠小说的框架风格，为读者带来独特的阅读体验。

（2）非段落特征指标聚类

小说非段落特征（篇章 + 句子 + 词汇）指标聚类结果如图 6 所示。

图 6 中 30 部小说被聚成两大类：第一类包括 5 部古龙早期作品和 13 部金庸作品；第二类包括 11 部古龙后期作品和金庸的《白马啸西风》（进一步印证《白马啸西风》是金庸所有小说中叙事风格独特的一部作品 [14]）。

结合作品创作时间分析，第一类中古龙的 5 部作品（《大旗英雄传》《浣花洗剑录》《名剑风流》《绝代双骄》《武林外史》）大致创作于 1964—1969 年间，同属于古龙早期作品，金庸的大部分作品创作时间要早于古龙。古龙早期作品与金庸作品在非段落特征上的高相似度，表明古龙早期作品创作很可能受到金庸作品风格的影响；第二类古龙后期作品（1970 年之后创作）间距离较大，表明其后期创作风格不断突破、发展变化，逐渐形成了独特个人风格。

以上结果体现了古龙作品创作的风格演变路径：从早期承袭到后期不断创新突破，逐渐形成独特的个人风格。

doi:10.3772/j.issn.2095-915x.2025.04.005

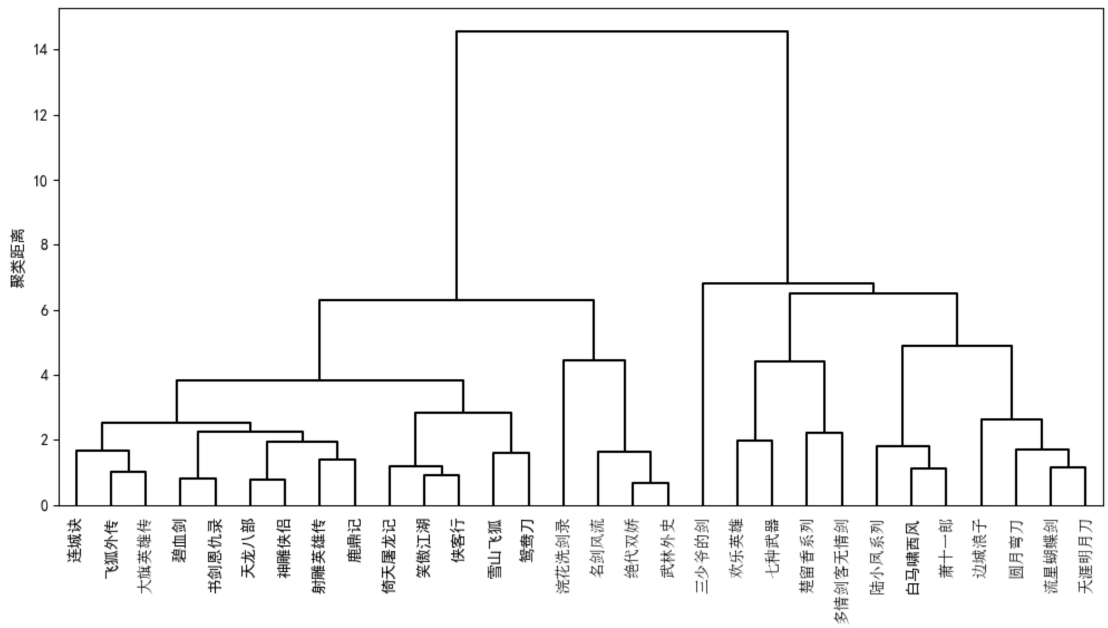

图 6　小说非段落特征指标聚类结果

（3）段落特征指标聚类

段落特征指标聚类结果如图 7 所示，结果显示，金庸小说紧密聚为一类，古龙小说聚为另一类，进一步说明段落特征确实是两位作者作品创作的代表性特征。

从图上纵轴聚类距离来看，金庸作品之间段落特征相似度非常高；古龙作品又可分成两个小类，小类内部段落特征相似度很高，两个小类之间作品差异性高，这仍体现出古龙不同作品的风格变化。

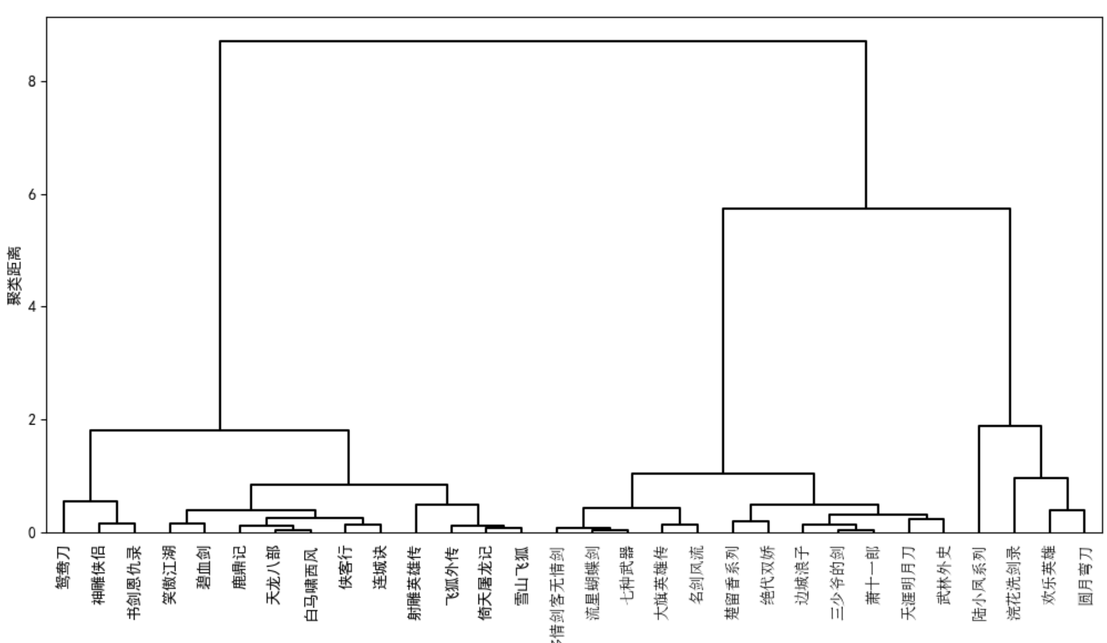

图 7　小说段落特征聚类结果

（4）段落特征可视化

前面不同研究方法均体现了金庸、古龙作品风格在段落特征上的各自独特性，在此以二维散点图可视化呈现二者作品段落特征的差异性与独特性，如图 8 所示。

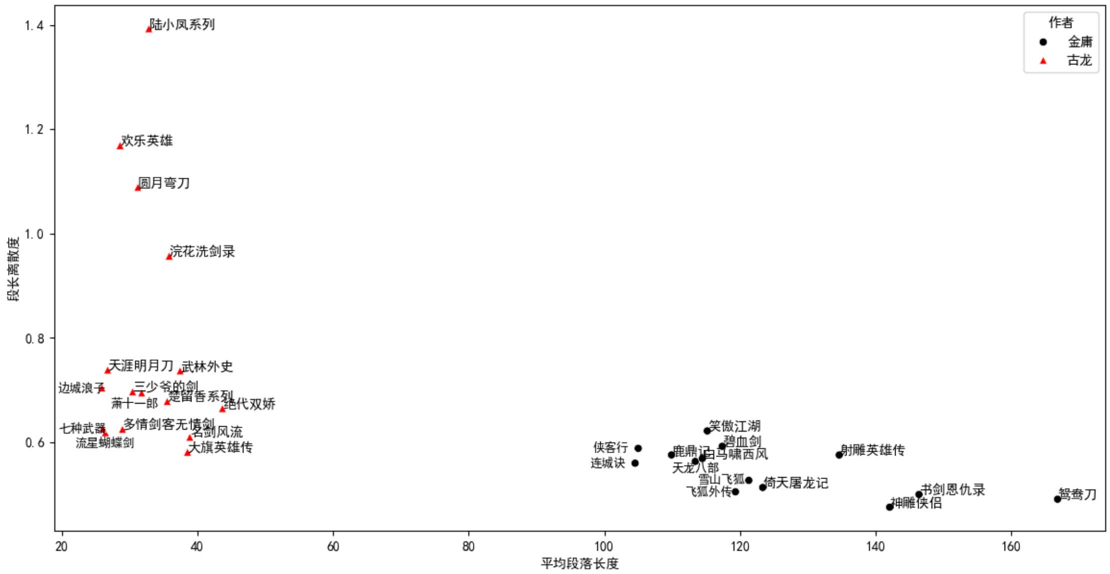

图 8　小说段落特征二维散点图

图 8 显示古龙小说分布在左侧，金庸小说分布在右下侧，各自分散互不交叉。

古龙作品在左侧上下分布，因其平均段落长度都较短，稳定在 40 字左右，但是不同小说的段长离散度却富于变化。段落特征综合反映了古龙小说叙事简洁明快，节奏紧凑，扣人心弦，情节推进速度较快，追求高效信息传递，能够吸引读者注意力的特点。

金庸小说相反，金庸小说平均段落长度长，偏向于大段落叙事，且每部小说段落长度离散度小，同小说内部段落长度一致性较高，行文节奏较稳，这综合反映了金庸小说舒缓、从容的叙事节奏，其擅长通过细腻描写和复杂句式，以高超的语言驾驭能力深度刻画人物心理、环境细节，让读者沉浸其中，深入人物角色的内心世界。

3　总结与讨论

3.1　研究贡献

本文面向中文文学作品，构建的“词汇—句子—段落—篇章”层次化风格分析指标体系，具有一定的普适性。虽然本文仅对武侠小说进行了比较研究，但其他体裁文学作品（如现代散文、诗歌）同样具有“词汇—句子—段落—篇章”结构，本文研究依然为其提供了作品风格计量分析的框架参考。

3.2　核心发现

通过多指标协同量化分析揭示了金庸与古龙创作范式的差异——整体上金庸以系统性风格稳定延续武侠经典化传统，古龙以动态风格多变性推动通俗化创新，并且一定程度上反映

doi:10.3772/j.issn.2095-915x.2025.04.005

了通俗性—从众性悖论：金庸作品词汇更丰富，低通俗性与高从众性结合，体现传统文化转译策略；古龙作品偏好口语化日常高频词汇，高通俗性与灵活从众性体现大众化创新。

尽管两者不同时期不同作品的创作风格不断变化，但段落特征（段落长度与离散度）却是金庸、古龙作品区分的核心代表性特征；结合作品创作年代分析，进一步发现古龙早期作品（1964—1969 年）在非段落特征聚类中与金庸高度相似，表明其早期风格承袭金庸，其后期作品（1970 年后）不断突破，形成独特叙事风格。

3.3 未来方向

当前研究指标体系以文本语言特征为主，未来可考虑整合情节网络、人物关系等非语言特征，构建全景式文学作品风格分析指标体系，并进行跨体裁扩展研究，验证指标体系在诗歌、散文等非虚构文学中的普遍适用性。

参 考 文 献

[1] GUIRAUD P. Problemes et methodes de la statistique linguistique [M]. Dordrecht: D. Reidel, 1960.

[2] 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(2): 104-108.

[3] 王泓 , 方艳梅 , 黄方军 . 基于信息熵的语言风格分析方法初探 [J]. 中山大学学报（自然科学版）, 2020, 59(6): 113-125.

[4] 张京楣 . 基于统计方法的文本风格分析研究 [D]. 济南：山东大学 , 2012.

[5] BURROWS J F. Word patterns and story shapes: the statistical analysis of narrative style[J]. Literary and Linguistic Computing, 1987, 2(2): 61-67.

[6] GRIEVE J. Quantitative authorship attribution: an evaluation of techniques[J]. Literary and Linguistic Computing, 2007, 22(3): 251-270.

[7] 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(5): 167-177.

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

[9] 贾俊平 . 统计学（第 7 版）[M]. 北京：中国人民大学出版社 , 2018.

[10] 张景祥 , 安培壮 , 苏娜 . 唐诗字频熵分析与通俗性定级 [J]. 科技资讯 , 2009(6): 241-243.

[11] 梁君英 , 刘海涛 . 语言学的交叉学科研究：语言普遍性、人类认知、大数据 [J]. 浙江大学学报（人文社会科学版）, 2016, 46(1): 108-118.

[12] 荀恩东 , 饶高琦 , 肖晓悦 , 等 . 大数据背景下 BCC 语料库的研制 [J]. 语料库语言学 , 2016(1): 93-109.

[13] 朱君辉 , 王梦焰 , 杨尔弘 , 等 . 大模型生成回答与人类回答文本的语言特征比较研究 [J]. 中文信息学报 , 2024, 38(4): 17-27.

[14] 黄大军 , 伊彩霞 . 塞外迷情荒漠孤女——论金庸武侠小说《白马啸西风》的叙事张力 [J]. 名作欣赏 , 2011(27): 50-52.

[15] 王阳阳 . 基于朴素贝叶斯与 BP 网络神经分类方法的《红楼梦》文本特征差异研究 [J]. 统计与决策 , 2018, 34(13): 121-125.

[16] 吴树芳 , 常欢 , 刘畅 , 等 . 引入有向联合图的谣言传播中关键节点群组检测研究 [J/OL]. 现代情报 , 1-15[2025-07-21]. http://kns.cnki.net/kcms/detail/22.1182.G3.20250102.1010.002.html.

[17] 唐朝 , 陈波 , 谭泽霖 , 等 . 基于知识蒸馏的半监督古籍实体抽取 [J/OL]. 数据分析与知识发现 , 1-17[2025-07-21]. http://kns.cnki.net/kcms/detail/10.1478.G2.20250205.1415.004.html.

[18] 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(6): 194-200.

[19] KERNER Y H, MARGALIOT O. Authorship attribution of responsa using clustering[J]. Cybernetics and Systems, 2014, 45(6): 530-545.

[20] 叶怡雄 . 企业股权分配与内部控制要素的多重共线性聚类分析 [J]. 统计与决策 , 2020, 36(4): 181-185.

[21] 陈曦 . 基于层次分析法和聚类分析法的高等教育自学考试多元评价模型探析 [J]. 中国考试 , 2021(3): 15-21.