或問　WAKUMON　175
No.47,（2025）pp.175-180

武侠小说与浪漫小说词汇特征对比分析[1]

王丽娟/孙曦宇

摘要：文体风格研究中，计量语言学的指标已成为一种有效分析和区别作者、文体和语体特征的方法。因此，本文从计量语言学角度对武侠小说与浪漫小说进行不同层面的加工和分析，通过计量指标的考察，揭示两种小说在语言表达上的内在规律与风格差异。

关键词：计量　武侠小说　浪漫小说

1 引言

词汇特征分析逐渐成为洞悉文本风格与类型差异的关键途径。小说作为一种以语言为载体的叙事艺术，其语言运用不仅承载着情节，更在人物塑造、情感传达以及叙事节奏等方面发挥着至关重要的作用。小说语言的选择与组织方式，通常反应作家的审美倾向，同时也映射出特定文本类型的功能需求与读者的期待。因此，对小说文体进行细致的词汇特征分析，有助于深入探究文学作品在内容与形式之间的互动关系。

在中文文学领域，不同类型的叙事作品在语言风格上展现出显著的差异。以武侠小说与浪漫小说为例，这两种极具代表性的文学体裁，在题材设定、叙事焦点以及词汇风格上均呈现出鲜明的对比。武侠小说以江湖世界为背景，塑造了一系列英雄形象，叙述了错综复杂的恩怨情仇。其语言风格通常充满动感和节奏感，频繁使用动词和武术术语，强调场景的紧张感和情节的推动力。此外，武侠小说中常见的四字成语、古典风格的词汇以及书面化的表达方式，共同构成了其独特的语言特色。相比之下，浪漫小说则更多地聚焦于人物情感地描绘，注重内心世界地深入刻画和心理变化地细腻描述。其语言运用倾向于抒情和感性表达，形容词、心理动词和情态副词地使用较为频繁，强调情感的层次和语言的精致性。这两种小说在词汇选用、句法结构以及语言节奏上的不同，反映了它们在情节发展、人物塑造以及整体审美追求上的特点。

在文体风格研究领域，计量语言学的指标已经成为分析和区分作者、文体以及语体特征的有效工具。在对文本风格进行分析的过程中，计量语言学运用了一系列可量化的技术手段，对文本进行多层次的处理和分析。其中，基于词频统计的词汇丰富度指标，已成为文本分析中广泛采用的重要工具。词汇丰富度涵盖了若干常用指标，例如型例比、单现词、h 点、信息熵、重复率等。然而，这些基于词频统计的词汇测量指标存在一个固有的问题，即过度依赖于文本的长度。为了降低文本长度对分析结果的影响，当前的研究更倾向于采用多个指标综合评估文本的词汇特征，以期获得更为全面和可靠的研究成果。在这样的研究背景下，本文选取了具有代表性的武侠小说

[1] 本文为浙江省教育厅课题（Y202353747）的最终研究成果。

和浪漫小说文本作为研究语料，基于高频词统计、MATTR、α指数以及单现词词频等多个维度，对这两种小说的词汇特征进行了系统的对比分析。通过这些定量指标的深入考察，本文旨在揭示武侠小说与浪漫小说在语言表达上的内在规律和风格差异，并进一步探讨小说语言如何满足不同类型文本的审美和功能需求。

2 语料与方法

2.1 语料来源

为达成对语言现象的量化分析，首要任务是搜集一定数量的自然语言素材，并对这些素材进行加工以提取所需的语言数据。本项研究采用的语料来自兰卡斯特中文语料库（以下简称LCMC）。该语料库由英国兰卡斯特大学语言学系与中国北京外国语大学共同构建，旨在为现代汉语的计算语言学研究及语体分析提供标准化、可量化的数据支持。LCMC是一个平衡型、标注型的现代汉语文本语料库，其设计原则借鉴了英国LOB语料库（LOBCorpus），内容覆盖新闻、科普、社论、小说等十五种不同体裁的书面文本，每种体裁包含五万词，总计达到百万级词汇。在体裁分类方面，LCMC进一步将小说文本细分为不同类型，以便研究者对特定文本类型进行深入比较分析。本研究选取了具有代表性的“武侠小说”与“浪漫小说”文本作为研究对象，以确保语料在来源和体裁上的一致性和可比性。由于LCMC的文本已经完成了词性标注和句子划分，这使得提取高频词、计算类型-标记比（TTR）、MATTR、α指数以及单现词频率等词汇统计指标变得简便，因此，它成为进行词汇特征对比研究的理想语料基础。所选武侠小说文本为多部武侠小说的片段合集，内容涉及古龙、金庸、还珠楼主、卧龙生等著名作家的作品。而浪漫小说文本则包括留学生题材小说、民间故事新编、中篇小说以及独立短篇等多种类型。

表1 语料基本信息

文本风格	文本数量	词例数
武侠	30	47162
浪漫	28	47681

2.2 词汇测量指标

词汇的出现频率是其特征中最容易量化的属性之一。本研究采用的词汇度量指标均与词频相关，包括平均型例比（MATTR）、α 指数以及单现词的词频。在分析过程中，将利用 AntConc3.5.8 软件对高频词汇进行计算与分析，并使用计量文本分析软件 QUITA 来计算 MATTR，同时统计语料库中的 h 点和单现词。

高频词的分析可以为观察文本频次分布和特点带来一定的参考意义，它通常能够反映出文本突出讨论的重点话题。高频词大多为只具有功能和语法意义的功能词，也包括部分实词。高频实词的统计对我们分析文本主要围绕哪些话题展开具有重要意义。

词的型例比（TTR）是文本中不同词型数和整体文本词例数的比例，是一种经典的词汇复杂度的测量指标。TTR虽然能够衡量文本中不同词项与总词数之间的比例，但其结果容易受到文本

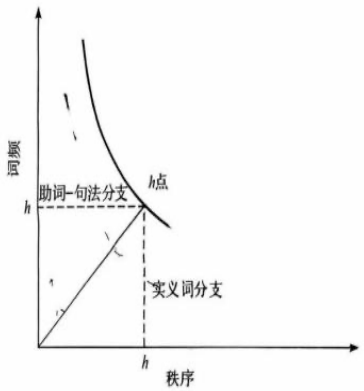
长度的显著影响，不利于不同文本之间的横向比较，为了克服这一局限，本文采用MATTR这一更加稳健的指标。MATTR是一种滑动窗口下的类型-标记比率计算方法，具体做法是在文本中设定一个固定长度的滑动窗口（如500词），并逐词滑动地计算该窗口内的TTR，最后取所有窗口TTR的平均值。该方法能够有效平衡局部与整体的信息，减少文本长度对TTR的影响，从而更稳定、

$$MATTR = \frac{1}{N-w+1}\sum_{i=1}^{N-w+1}\frac{V_i}{w}$$

准确地反映文本的词汇多样性。MATTR计算公式如上。

其中，Vi是窗口中的型符数，W是窗口大小，N为文本长度，窗口大小须小于文本长度（W＜N）。在文本长度相近或控制变量一致的情况下，MATTR越高，说明该文本使用的不同型符越多，意味着文本具有较高的词汇丰富度。

α指数是基于h点提出的一个指标。对于文体研究来说，h点是反映文本风格特征的重要指标之一。h点，即在秩频分布曲线中，频次与频序相等的点。h点的计算分为两种情况，如果词频分布曲线上能够找到频次f（r）与频序（r）相等的点，则h=r。如果不存在这样的观测点，则需要通过公式计算h的值。h点的一个重要的语言学意义在于它将词频曲线分成了两个部分，粗略划分了语义词和句法功能词。如图1，在词频曲线中，h点之前的词通常为虚词或者助动词，数量不多但是使用频次很高，并且h点之前出现的实词往往与文本主题具有强相关性。h点之后的词多为实词，频次不高，但是对词汇的丰富性具有重要意义。h点取值越小，则高频词尤其是句法功能词所占比重较小；h点取值越大，则实词、低频词和单现词所占比重越小。h点指数虽然对于文本特征分析有重要意义，但是容易受文本大小的影响。在此基础上，不易受文本大小影响的α指数成为另一个有力的测量指标。α指数与h点呈反向关系，即h点数值越大，α值越低，文本中句法语义词越多，实词等语义词比重就越小。



单现词是指文本中只出现一次的词。在词频分布当中，单现词占据了其末端部分。单现词比例可以体现文本的词汇丰富度，比例越高，说明有更多复杂或不常见的词，反之，则有更多重复的词。

3 结果

3.1 高频词

高频词，尤其是实词（如名词、动词和形容词）的分布能够在一定程度上体现文本的风格和主题。本文使用AntConc对武侠小说和浪漫小说的文本分别进行分析，分别统计这两个文本的高频词和高频实词。

表2 两文本的高频词（前20位）

文本类型	高频词
武侠	的，了，一，是，他，人，你，不，这，我，那，看，她，说，就，来，在，也，道，有，
浪漫	的，了，我，她，是，一，你，不，地，说，着，也，就，和，上，人

从两个文本的高频词统计可以看出，不论是武侠小说还是浪漫小说，出现频率最高的词基本上都是句法功能词。此外，两个小说类型的文本当中，你、我、他等人称代词出现频率较高，这说明不管是武侠小说还是浪漫小说，其情节始终都是围绕着人的行为展开的，人与人之间的互动都是这两类小说的核心主题。

表3 武侠小说和浪漫小说的高频实词（前10位）

文本类型	高频实词
武侠	少女，剑，打，姑娘，无忌，说道，船，江湖，在下
浪漫	姑娘，琉璃，笑，柯比，家，仙女，爱，生活，妻子，丈夫，贝贝

通过分析高频实词（参见表3），可以明显观察到武侠小说文本中对人物身份、动作行为以及对话内容的依赖性。例如，“少女”“剑”“打”“姑娘”“无忌”“说道”等词汇的高频出现，揭示了武侠小说特有的叙事特点。这些词汇中，“剑”“打”“船”“江湖”等词的使用，反映了武侠小说通常以武力冲突和江湖游历为核心，展现了其叙事的动态性和场景的流动性。而“在下”“无忌”等词汇则体现了武侠小说中特有的语言风格，强调了人物间的礼仪关系和江湖世界的规则。相对而言，浪漫小说的高频词汇则集中在“姑娘”“琉璃”“笑”“柯比”“家”“仙女”“爱”“生活”“妻子”“丈夫”等，这些词汇突出了家庭、情感和日常生活作为主要叙事内容的重要性。特别是“爱”“生活”“妻子”“丈夫”等词汇，清晰地描绘了浪漫小说对亲密关系的重视，而“笑”“琉璃”“仙女”等词汇则增强了文本的抒情性和浪漫氛围。词云图（参见图1）进一步揭示了两类文本在关键词使用上的差异。

图1 武侠小说和浪漫小说关键词词云

武侠小说的核心关键词包括“道”“人”“他”“这”“你”等，这些词汇构成了以“人物—对话—行为”为轴心的叙事架构。在这些文本中，动作性或言语性动词“说道”“听”“来”等频繁出现，这反映了武侠小说在人物互动和情节推进方面对对话和行为描写的依赖。同时，地名、人名和专有名词如“剑”“江南”“无忌”“杜”等的高频使用，揭示了武侠文本特有的江湖地理布局和人物谱系构建。在浪漫小说中，词云以“我”“她”“你”“也”“地”等代词为高频词汇，突显了叙述者的主观情感和人物间情感的张力。此外，关键词如“琉璃”“柯比”“宏”“安伦”“仙女”“喜欢”“真心”等，进一步展现了浪漫小说在命名、意象选择和情感表达方面的柔美、私密和诗意特质。

3.2 MATTR

MATTR是研究和分析语料词汇丰富度的重要指标。通过使用WordList软件计算得出，武侠小说文本和浪漫小说文本的MATTR值分别为0.604和0.589，这说明武侠小说语料词汇丰富程度高于浪漫小说语料。同时WordList计算得出的词汇密度当中，武侠小说语料为0.796，远高于浪漫小说语料的0.729，这都证明了武侠小说语料的词汇丰富程度高于浪漫小说语料。此一现象可理解为文本风格上的差异：武侠小说往往包含复杂的动作描写、地理描写及大量人物与事件之间的交错关系，对具象词汇的依赖性更强；而浪漫小说更关注心理活动与情感细腻度，语体相对主观化、抽象化，导致实词密度相对较低。

3.3 h点分析

h点对于间接分析文本中词汇的丰富程度以及实词所占比重具有重要意义。使用QUITA软件计算得出武侠小说文本和浪漫小说文本的h点分别为71.5和72.5，两者差异较小。这说明这两个文本在词汇分布情况以及实词占比情况上极为相似。此外，R1（词汇丰富度指标）是基于h点对文本中的实词（例符）的比例的估计。因为h点之前可能存在实词，h点之后也可能存在功能词，所以只看h点这一单一衡量指标无法准确估量文本的实词使用的丰富程度。所以在h点的基础上引入了新的指标F（h），并通过F（h）计算出R1，使这一指标能够更加客观的反映文本中的实词分布情况。

通过使用QUITA软件，分别计算出武侠小说文本和浪漫小说文本的R1值分别为0.714688和0.6567。在两个文本的h点值十分近似的情况下，引入R1值分析可以得出，与浪漫小说相比，武侠小说的实词使用程度更加丰富。值得注意的是，R1值不仅仅反映了实词在总体词汇中的占比，更隐含了文本对信息密度与语义表达依赖程度的差异。在本研究中，武侠小说文本的R1值显著高于浪漫小说，说明其在语言构建中更依赖具有语义承载功能的词汇。具体而言，武侠小说常涉及人物动作、武功描写、场景调度等复杂叙述内容，需大量使用动词、名词等实词来构建叙事空间与推动情节发展；而浪漫小说则可能更偏向于心理描写和情感表达，使用较多抽象性表达与功能性词语。由此可见，R1指标在h点的基础上，有效弥补了仅以词频分布判断词汇丰富度可能存在的误差，使得对不同小说文本之间语言特征的比较更具客观性与解释力。

3.4 单现词词频

单现词词频（Hapax percentage）指的是单现词的占N的比例，能够在一定程度上体现词汇的丰富程度。通过AntConc软件对武侠小说文本和浪漫小说文本进行词频分析，得出武侠小说文本中单现词为4415项，浪漫小说文本中单现词为4476项。使用QUITA对两者的单现词词频进行计算分析，得出武侠小说文本和浪漫小说文本的单现词词频分别为0.093635和0.09444。在对武侠小说与浪漫小说语料进行词频统计后发现，两类文本的单现词数量及其占总体词项的比例十分接近。这一结果表明，尽管武侠小说与浪漫小说在叙事结构与情节内容方面存在明显差异，但在词汇使用的丰富度和稀有词的运用程度上却具有相似性。换言之，两类小说在语言表达上都表现出一定程度的多样化倾向，即均倾向于使用较多只出现一次的词汇，以增强描写的细腻性与个体化特征。此现象可能与小说体裁本身对文学表达的要求密切相关：无论是武侠小说中的动作描写与情境构建，还是浪漫小说中的情感刻画与心理描写，都需要借助较为丰富、灵活的词汇资源，从而形成具有风格化和表现力的语言文本。

4 总结

本文基于兰卡斯特中文语料库（LCMC）的武侠小说与浪漫小说文本，通过高频词统计、MATTR、α指数（R1值）及单现词词频等量化指标，系统对比了两类小说在词汇特征上的差异。

研究发现，武侠小说与浪漫小说在词汇选择、语义重心及语言功能上存在显著差异，具体体现为以下核心结论：

1、 在武侠小说中，词汇的丰富性较高（MATTR=0.604），且实词所占比例显著（R1=0.715）。高频实词如“剑”“打”“江湖”以及专有名词（“无忌”“江南”）的频繁使用，突显了其以动作驱动的叙事特点：依赖具体的物理空间（地理、兵器）、动态行为（武打、游历）以及人物关系的构建，通过语义密度高的语言推动情节冲突和江湖格局的展开。相比之下，浪漫小说的词汇丰富性相对较低（MATTR=0.589），实词比例有所减弱（R1=0.657）。其高频词“爱”“妻子”“笑”“仙女”等主要集中在情感状态和家庭关系的描述上，反映了心

（下转192页）