Modern Linguistics 现代语言学, 2025, 13(11), 622-629
Published Online November 2025 in Hans. https://www.hanspub.org/journal/ml
https://doi.org/10.12677/ml.2025.13111205

Hans汉斯

基于语料库的小说文本对比分析

——以《杀死一只知更鸟》和《布谷鸟的呼唤》为例

刘怡欣

江西理工大学外国语学院，江西 赣州

收稿日期：2025年10月11日；录用日期：2025年11月7日；发布日期：2025年11月21日

摘　要

本文采用语料库研究方法，对哈珀·李的《杀死一只知更鸟》和J.K.罗琳的《布谷鸟的呼唤》进行了对比分析。通过构建语料库并应用文本分析工具，本文从词汇、句法和语篇三个层面揭示两部作品的文体特征及其社会意义。《杀死一只知更鸟》通过孩童的视角探讨种族与道德议题，而《布谷鸟的呼唤》则通过侦探叙事形式揭示现代社会对名利的扭曲追求。结果表明，两者在词汇复杂性、句法结构和叙事视角上存在一定差异。本研究为语料库方法在比较文学研究中的具体应用提供了一个案例，尝试在经典文学与当代流行文本之间建立初步的对话关系，通过量化数据揭示了两者在文体策略上的差异，为理解它们在不同时代背景下的社会批判功能提供了一个新的经验性视角。

关键词

语料库研究，小说文本，文体特征

A Corpus-Based Comparative Analysis of Novel Texts

—With Special Reference to *To Kill a Mockingbird* and *The Cuckoo's Calling*

Yixin Liu

School of Foreign Languages, Jiangxi University of Science and Technology, Ganzhou Jiangxi

Received: October 11, 2025; accepted: November 7, 2025; published: November 21, 2025

Abstract

This study employs a corpus-based methodology to conduct a comparative analysis of Harper Lee's

文章引用: 刘怡欣. 基于语料库的小说文本对比分析[J]. 现代语言学, 2025, 13(11): 622-629.
DOI: 10.12677/ml.2025.13111205

To Kill a Mockingbird **and J.K. Rowling's** ***The Cuckoo's Calling*****. By constructing specialized corpora and utilizing text analysis tools, this research reveals the stylistic features and social significance of the two works across three dimensions: lexical, syntactic, and discursive.** ***To Kill a Mockingbird*** **explores issues of race and morality through the lens of a child narrator, whereas** ***The Cuckoo's Calling*** **exposes the distorted pursuit of fame and fortune in modern society through its detective narrative framework. The results indicate certain differences between the two novels in terms of lexical complexity, syntactic structures, and narrative perspectives. This study provides a practical case of applying corpus methods in comparative literature research, attempting to establish a preliminary dialogue between classical and contemporary popular texts. By revealing differences in their stylistic strategies through quantitative data, it offers a new empirical perspective for understanding their social-critical functions within different historical contexts.**

Keywords

Corpus-Based Study, Novel Texts, Stylistic Features

1. 引言

《杀死一只知更鸟》(*To Kill a Mockingbird*)是美国作家哈珀·李于 1960 年出版的小说，以 1930 年代美国南方为背景，通过 6 岁女孩斯库特·芬奇的视角，讲述其父亲阿蒂克斯为黑人汤姆·鲁滨逊辩护的种族歧视案件。汤姆被诬告强奸白人女性，尽管证据不足仍被判罪，最终惨死。故事同时穿插斯库特与哥哥对神秘邻居“怪人”拉德利的探索，最终拉德利在危急时刻拯救了兄妹俩。而《布谷鸟的呼唤》(*The Cuckoo's Calling*)是 J.K.罗琳以笔名罗伯特·加尔布雷思于 2013 年出版的小说。当代伦敦背景下，私家侦探科莫兰·斯特莱克调查超模卢拉·兰德里坠楼案。此案看似自杀，实则涉及家族秘密、毒品交易与身份欺诈。斯特莱克与助手罗宾抽丝剥茧，揭露卢拉的生母为掩盖早年抛弃女儿的罪行，联合他人制造谋杀的故事。

两部小说均对社会中的结构性压迫进行了深刻的批判。《杀死一只知更鸟》通过孩童纯真的视角聚焦种族主义，探讨“正义”的定义，并凸显同理心与社会责任的重要性；《布谷鸟的呼唤》则采用冷硬派侦探小说的形式，探讨“真相的代价”，批判现代社会对名利的扭曲追求。基于语料库的文学研究是指采用语料库方法，在分析文学文本的语言特征并进行相关数据统计的基础之上，深入探讨文学文本特征、作家风格以及文学与社会的关系等问题[1]。自 20 世纪 80 年代以来，随着计算机的跨学科应用和语料库语言学的不断发展，对语料库文体学的研究已逐渐发展成为一种重要的研究范式[2]。对比分析这两部作品，有助于观察文学进行社会批判的不同叙事手法。本文为比较文学研究提供新的案例，通过细致的文本分析，揭示文学如何以特定叙事策略介入社会议题。两部作品分别出自 20 世纪与 21 世纪，在主题上具有互补性：前者聚焦正义，后者探讨真相，展现了不同时期文学回应社会问题的多元方式。通过将其置于各自的社会语境中考察，可以更清晰地把握文学在不同历史条件下实现社会批判的具体路径，从而深化对文学反映和探讨社会问题的理解。

2. 研究方法

2.1. 语料库的构建

语料库构建的第一步是语料的选择，根据要求，笔者选择的是哈珀·李的经典小说《杀死一只知更

鸟》的英文原版 *To Kill A Mockingbird*，该书的字数约为十万字。这部作品自首版以来便在全球范围内引起了广泛的关注与讨论，不仅因其引人入胜的故事性，更因为其深刻的社会和道德哲学探讨，成为了文学经典。作为参考语料，笔者选择了由 J.K.罗琳创作的侦探小说《布谷鸟的呼唤》(*The Cuckoo's Calling*)。这部小说同样采用了鸟类隐喻，同样提供了一个关于调查和揭示真相的引人入胜的故事，尽管其主题和风格与《杀死一只知更鸟》有所不同。选择这两本小说的原因在于它们在叙事结构和人物描绘上的共通性，同时又能够展示出不同文化和时代背景下的写作特色。

语料库的文本处理主要包括计算机自动处理和人工属性标注两个步骤[3]。语料选取完毕后，开始降噪清洗。首先删除文本中所有空行，打开 Word 自带的“查找与替换”功能，选择“替换”，查找内容为“^p^p”，替换为“^p”，然后选择“全部替换”，这样文本中所有的空行都删除完毕。由于小说中存在大量对话，因此要将所有的引号(包括单引号和双引号)全部由全角变为半角格式，此功能仍然是由“查找与替换”中的“替换”功能来完成。降噪完成，将其另存为 TXT 格式。接下来是词性标注，打开兰卡斯特大学在线免费标注网站，因为本网站只能一次性标注十万字英文文本，所以笔者分两次进行文本粘贴，即可得到全部的词汇标注。另一本书操作同上。初步准备工作完成后，打开 Wordsmith，选择 Word list 功能，选中已生成的 TXT 格式文本文件，点击 Make a word list now，即可生成一个表格，其中包含 frequency 词频表、alphabetical 按字母排序表格、statistics 相关数据表格等五个子表格。而 frequency 词频表又包括 freq 词频、per million 标准频数、dispersion 离散度等。

2.2. 数据分析工具与技术

本研究主要采用了 Wordsmith 7.0 和 AntConc 4.2.4.0 这两款文本分析软件，以深入探索文本间的差异和共性。这两个工具各具特色，能够在词汇、句法和语篇三个层面提供全面的分析功能。Wordsmith 是一款功能强大的文本分析软件，能够帮助研究者进行词汇频率统计、关键词提取、搭配分析等。通过该软件，我们可以直观地了解文本中使用的高频词汇及其上下文，进而识别出文本的主题和潜在语义。同时，Wordsmith 提供的词汇搭配功能能够揭示出词与词之间的搭配关系，帮助我们理解语言的使用习惯和风格特征。AntConc 是一款开放源代码的文本分析工具，尤其适用于语料库语言学研究。通过使用 AntConc，我们能够执行语料库检索、共现分析和词云生成等操作，这有助于我们从更宏观的角度把握文本的结构和模式。此外，AntConc 还提供丰富的可视化选项，使得分析结果一目了然，更容易进行直观的比较和解释。为了进一步支持和验证分析结果，我们还运用了统计工具 Chi-Square Calculator，以计算卡方值(χ^2)和 p 值。这一统计方法通过对不同文本中词汇使用的显著性差异进行量化，为我们的分析提供了科学依据。通过统计检验，我们可以判断观察到的差异是否具有统计学上的显著性，从而增强结论的可靠性和说服力。

3. 文本对比分析

本文基于自建的小说语料库，从词汇、句法和语篇三个层面，对《杀死一只知更鸟》和《布谷鸟的呼唤》英文原本进行文体特征的比较与分析。

3.1. 词汇层面

3.1.1. 词汇长度

平均词长指文本中所使用词语的平均长度，以词语的字母数为计算标准。词长标准差指文本中每个单词长度与平均词长的差异[4]。一般而言，平均词长的数值越高，表明文本中使用的复杂词就越多。将两本小说导入 Wordsmith 7.0 的 Word list 功能，即可得出两个语料字库的平均词长与词长标准差。对比

表 1 数据可知，《杀死一只知更鸟》的平均词长和词长标准差略小于《布谷鸟的呼唤》，说明后者使用的复杂词更多，正式程度也越高。

Table 1. Mean word length and word length standard deviation in the two subcorpora
表 1. 两个语料子库的平均词长及词长标准差

参数类型	*To Kill a Mockingbird*	*The Cuckoo's Calling*
平均词长	4.18	4.26
词长标准差	2.11	2.25

Table 2. Total number and percentage of words (≥6 letters) in the two subcorpora
表 2. 两个语料子库的长词总数与占比

参数类型	*To Kill a Mockingbird*	*The Cuckoo's Calling*
长词	23,371	38,581
总词数	100,607	149,613
长词占比	23.23%	25.79%

此外，英语中通常将 6 个字母以上的词称为长词。通过 Word list 得出的数据，可以绘制出两个语料子库的长词总数及其占比表。对比表 2 可知，《杀死一只知更鸟》的长词占比略小于《布谷鸟的呼唤》一书，可见其文本活泼性更高，而后者用词较严谨正式。

3.1.2. 词汇密度

词汇密度由 Ure 提出，指文本中的词项数量(即实词数量)与该文本的单词总量之比，计算方法为：词汇密度 = 实词数/词汇总数 × 100% [5]。英语中的实词指名词、实义动词、形容词和副词。利用 CLAWS 软件对两个语料子库进行词性标注后，运行 AntConc 的 Word list 可得两个语料子库的实词总数，随后根据计算公式得出词汇密度。此外，为区别两个语料子库数据间的差异是否具有统计学意义，引入了卡方检验。卡方检验可以检测两个或多个分类变量之间相关性是否显著，若卡方值越大，二者偏差程度越大；反之，二者偏差越小。Fisher 提出的 p 值(p-value)也常用于检验组间的数据是否具有显著性差异。由于表 3 中 p 值小于 0.05，说明两个语料子库数据差异显著。《杀死一只知更鸟》的词汇密度小于《布谷鸟的呼唤》一书，可见其实词占比较低，信息负载量与文本难度也较小，此特征符合小说文本的基本特点，可读性较强，易于读者接受。

Table 3. Total lexical words and lexical density in the two subcorpora
表 3. 两个语料子库的实词总数及词汇密度

参数类型	*To Kill a Mockingbird*	*The Cuckoo's Calling*	χ^2	p
名词	15,821	26,738	196.2413	0.000
实义动词	15,603	21,629	52.5859	0.000
形容词	4368	9049	345.2655	0.000
副词	5132	8016	7.9675	0.005
实词总数	40,924	65,432	230.0479	0.000
总词数	100,607	149,613		
词汇密度	40.68%	43.73%		

3.1.3. 类符形符比

语料库语言学中，类符是文本中所使用的不同词汇的种类，形符是文本中所有词汇的总数量。类符/形符比(type/token ratio, TTR)是指文本中所使用的不同词语的数量与词语总数量间的比值，由于常用的不同词汇数量有限，文本长度可能会有较大的差异，因此 Scott 提出采用标准化类符/形符比(standardized type/token ratio, STTR)作为计量标准，标准化类符/形符比值越大，表示该文本词汇重复率低，使用不同词汇的数量越多，词汇变化性和多样性程度高。

Table 4. Type-Token Ratio (TTR) of the two subcorpora
表 4. 两个语料子库的类符与形符比

参数类型	*To Kill a Mockingbird*	*The Cuckoo's Calling*
类符	8305	11,531
形符	100,607	149,613
标准化类/形符比率(%)	42.67	44.79

词语的丰富性是指相同长度的语料中不同词语的数量大小[6]。根据 Wordsmith 7.0 中的 Word list 功能，可以得出两个语料子库的类符与形符数及标准类符与形符比率，即表 4。《杀死一只知更鸟》一书的标准化类符形符比较低，说明不同词汇量相对较小，词汇重复率较高，多样化程度较低。

3.1.4. 高频词参数

高频词指反复出现的一定数目的相同词汇，即语料库中出现频率较高的词。Laviosa 将高频词界定为“一个词项出现频率至少占库容的 0.10%的词”[7]。参照此定义，设定高频词所占比例的数值 ≥0.10%。借助 Wordsmith 制作词表，查询 Frequency 列表，得出两个文本的高频词数据。

Table 5. High-frequency words in the two subcorpora
表 5. 两个语料子库的高频词数据

参数类型	*To Kill a Mockingbird*	*The Cuckoo's Calling*
高频词数目	158	144
累计比例	60.67%	58.23%
高频词重复率	386.49	605.60
高频词与低频词之比	0.1924	0.1614

根据表 5 数据可知，《杀死一只知更鸟》一书中的高频词重复率远低于《布谷鸟的呼唤》，但其高频词与低频词之比较高，可见其词汇变化较大，用语较丰富。

3.2. 句法层面

3.2.1. 平均句长

平均句长指一个篇章中的句子含有词语数量的平均值，计算公式为：平均句长 = 形符数/句子数。句长标准差指句子的长度在平均句长左右浮动的程度，标准差值越高，表明文本中句子长短变化越大，句式更为灵活，可读性也就越强。借助 Wordsmith 7.0 的统计功能，可以得出两个语料子库的平均句长。

Table 6. Mean sentence length of the two subcorpora
表 6. 两个语料子库的平均句长数据

参数类型	*To Kill a Mockingbird*	*The Cuckoo's Calling*
句子个数	8573	10,615
平均句长	11.74	14.09
句长标准差	9.84	11.93

根据表 6 可知，《杀死一只知更鸟》的平均句长与句长标准差均低于《布谷鸟的呼唤》，说明前者句子较短，变化幅度较小，可读性较强。而《布谷鸟的呼唤》一书用语较正式严谨。

3.2.2. 句子结构类型

英语句子按其结构可分为简单句、并列句和复合句。简单句只有一个主谓结构，并列句用并列连词将两个或两个以上的简单句连在一起，复合句由从属连词将两个或两个以上的简单句连接在一起。英语连词具有结构连接和语义连接两种主要功能。在故事撰写层面，使用连词可适当缩小阅读难度，让文章更连贯、更清晰。将经过词性赋码的两个语料子库导入 AntConc，运行其 Word list 功能，检索并列连词(CC、CCB)和从属连词(CS、CSA、CSN、CST、CSW)的数量。

Table 7. Connecting words in the two subcorpora
表 7. 两个语料子库的相关连词数据

参数类型	*To Kill a Mockingbird*	*The Cuckoo's Calling*	χ^2	p
从属连词	2234	3557	6.5541	0.010
并列连词	3281	4356	24.8605	0.000
总词数	100,607	149,613		
从属连词占比	2.22%	2.38%		
并列连词占比	3.26%	2.91%		

表 7 数据表明，《杀死一只知更鸟》的从属连词占比略低，但并列连词占比较高。说明两本小说在句子衔接方面各有侧重。且连词 p 值均小于 0.05，说明其差异性显著。

3.3. 语篇层面

3.3.1. 语篇衔接

语篇语言成分之间的语义联系，利用衔接手段可实现语篇上下文的逻辑连贯。逻辑联系语包括词、短语或分句。词包括连词和连接副词，分句包括非限定分句和限定分句。本文选择以词语为衔接机制的逻辑联系语为检索项，将词性赋码的两个语料子库导入 AntConc，运行其 Word list 功能，检索以连词(并列连词与从属连词)和连接副词(RGQ、RGQV、RL、RRQ、RRQV)为衔接形式的逻辑联系语数量。

Table 8. Logical markers in the two subcorpora
表 8. 两个语料子库的逻辑联系语数据

参数类型	*To Kill a Mockingbird*	*The Cuckoo's Calling*	χ^2	p
从属连词	2234	3557	6.5541	0.010

续表

并列连词	3281	4356	24.8605	0.000
连接副词	1085	1456	6.6327	0.010
总词数	100,607	149,613		
逻辑联系语占比	6.56%	6.26%		

表 8 数据表明，两个语料子库在从属连词、并列连词和连接副词等方面存在显著差异。虽然《杀死一只知更鸟》在库容上略小于《布谷鸟的呼唤》，但其逻辑联系语占比却略高，说明该书衔接词较多，其信息单元之间的逻辑和语义关系更加紧密和清晰，也有利于读者更有效地理解句段和语篇。

3.3.2. 叙事视角

叙事视角是语篇构建的基础，文本的撰写会基于不同的叙事角度，实现其故事主题与情感的传递。人称代词是语篇连贯的一种手段，一般而言，第一人称视角有助于增强故事的真实性，搭建真实性与主观情感融合的叙事结构。第二人称视角使篇章的叙事更具有互动性。第三人称视角则是以旁观者的视角进行观察与叙述，具备客观性的同时，也融合了经历者本身的声音[8]。选择通过检索人称代词在两个语料子库的数量，判断各自所属叙事视角的特点，以人称代词为检索项，将词性赋码的两个语料子库导入 AntConc，运行其 Concordance 功能。检索项目包括第一人称(I, we)、第二人称(you)和第三人称(he, she, they)。

Table 9. Personal pronouns in the two subcorpora
表 9. 两个语料子库的人称代词

参数类型	*To Kill a Mockingbird*	*The Cuckoo's Calling*	χ^2	p
第一人称	3329	1777	1185.1093	0.000
第二人称	1703	1669	56.1625	0.000
第三人称	3253	5373	307.7802	0.000
总句数	8573	10,615		

表 9 数据显示，三类人称的 p 值都小于 0.05，可见其差异显著。《杀死一只知更鸟》采用较多的第一与第三人称视角，而《布谷鸟的呼唤》更多的是从第三人称视角出发，说明后者更为客观，而前者互动性更强，易于读者感同身受。

4. 结论

本研究通过对比分析《杀死一只知更鸟》与《布谷鸟的呼唤》的文体特征与社会批判模式，揭示出文学语言与时代语境之间的深刻关联。哈珀·李的经典文本以孩童视角为叙事核心，采用简洁句式与低词汇密度的文体策略，通过“法庭”、“知更鸟”等隐喻体系及高频出现的“同理心”、“责任”等伦理词汇，构建出对美国南方种族压迫的双重批判——既呈现制度性歧视的荒诞性，又呼唤人性良知的社会重建。与之形成鲜明对照的是，J.K.罗琳在《布谷鸟的呼唤》中运用冷硬派侦探小说的复杂句法与高词汇密度，借助“资本操纵”、“媒体异化”等消费社会术语与多线索嵌套叙事，层层剖解当代都市中名利对人性的系统性扭曲。定量分析显示，两者在平均句长与词汇密度上分别存在一定的差异，这种文体分野恰与其社会批判路径形成镜像：前者以情感化的质朴叙事追问普世正义，后者以理性化的精密结构质疑

真相本质。研究进一步指出，文本的叙事选择折射出各自时代的文化焦虑——20 世纪中期对种族伦理的启蒙诉求与 21 世纪全球资本主义的生存困境，共同构成文学介入社会议题的两种范式，为理解不同时期文学作品的叙事特征与社会内涵提供了参考。

参考文献

[1] 胡开宝, 杨枫. 基于语料库的文学研究: 内涵与意义[J]. 浙江大学学报(人文社会科学版), 2019, 49(5): 143-156.

[2] 诸葛晓初, 吴世雄. 国外语料库文学文体学研究: 回顾与前瞻[J]. 外语学刊, 2022(4): 11-18.

[3] 刘淼, 邵青. 基于多译本平行语料库的翻译语言特征研究——对契诃夫小说三译本的对比分析[J]. 解放军外国语学院学报, 2015, 38(5): 126-133.

[4] 蔡强, 黄婷婷. 中西方媒体“人物故事”文本的文体特征对比分析[J]. 沈阳工程学院学报(社会科学版), 2021, 17(2): 71-77.

[5] 蔡强, 张建平. 基于语料库的科技论文摘要英译语言特征研究[J]. 江西理工大学学报, 2014, 35(6): 60-65.

[6] 张继光, 张政. 基于语料库的当代英语散文汉译规范研究[J]. 外语教学理论与实践, 2014(4): 83-91+95.

[7] Laviosa, S. (2002) Core Patterns of Lexical Use in a Comparable Corpus of English Narrative Prose. *Meta*, **43**, 557-570. https://doi.org/10.7202/003425ar

[8] 申丹. 对叙事视角分类的再认识[J]. 国外文学, 1994(2): 74.