8万本书中的语义新颖性轨迹：跨语料库嵌入分析

W. Frederick Zimmerman Nimble
Books LLC，安娜堡，密歇根州
wfz@nimblebooks.com

2026年3月

arXiv:2603.01791v1 [cs.CL] 2 Mar 2026

摘要

我将 Schmidhuber 的有趣性压缩进展理论应用于语料库规模的分析，研究跨越两个世纪英语出版史的 8 万余本书中的语义新颖性轨迹。我使用句子变换器段落嵌入和运行质心新颖度度量，对比了 28,730 本 1920年以前的古腾堡计划图书（PG19）与 52,796 本现代英语书籍（Books3，约为 1990–2010）。主要发现有四点。第一，现代书籍的段落级新颖度均值约高出 10%（0.503 对 0.459）。第二，轨迹迂回度——即嵌入空间中累计路径长度与净位移的比率——在现代语料库中几乎翻倍（+67%）。第三，新颖度逐渐下降并趋向稳定语义语域的收敛叙事曲线，在 1920年以前的文学中常见 2.3× 。第四，新颖性与读者质量评分正交（$r = -0.002$），这表明，按照 Schmidhuber 的定义，趣味性在结构上独立于人们感知到的文学价值。通过 PAA-16 表示对段落级轨迹进行聚类，可识别出八种不同的叙事形态原型，其分布在不同时代之间有显著变化。所有分析代码和交互式探索工具包均已公开，网址为 https://bigfivekiller.online/novelty_hub。

关键词： 计算文学研究、语义新颖度、句子嵌入、远读、压缩进展、叙事轨迹

1 引言

是什么让一段文本有趣？Schmidhuber（2009年）基于算法信息论提出了一种形式化解释：如果学习算法随着时间推移能够更好地压缩数据流，观察者就会觉得它越有趣。新数据带来的瞬时“趣味性”，就是压缩进展——编码迄今所见内容所需比特数的减少。将这一框架转化为现代表征学习的语言后，它预测：当连续段落相对于文本不断累积的上下文，位于嵌入空间中语义新颖的区域时，读者会感受到更强的投入感。

尽管这一理论优雅简洁，Schmidhuber 的理论在自然语言文本领域仍缺乏大规模实证检验。现有计算文学研究考察了词汇新颖性（Moretti，2005；Jockers，2013）、主题分布（Blei 等，2003），以及每词熵和困惑度等信息论量度（Genzel 和 Charniak，2002）。这些方法着眼于词频或主题比例。密集句子嵌入的出现（Reimers 和 Gurevych，2019年）开辟了一条互补路径：在能够捕捉超越表层词汇的意义的高维语义空间中测量新颖性。

据我所知，本文首次在 80,000+ 本完整篇幅图书的规模上，开展了跨语料库、基于嵌入的新颖性分析。本文的贡献包括：

1. 提出一种**运行质心新颖度度量**，用于将句子嵌入空间中的压缩进展操作化（§3）。2. 对两个相隔约一个世纪的大型英语语料库进行**系统比较**：由 1920 年前的公版作品构成的 PG19 数据集，以及收录现代（约1990—2010年）出版物的 Books3 文集（§4）。3. 提供证据表明，**新颖性在结构上独立于感知质量**，从而动摇“越出人意料就越好”这一朴素假设（§4.5）。4. 确定**八种叙事轨迹原型**，其在不同文学时代中的占比发生了显著变化（§4.4）。5. 提供一个**公开的交互式工具包**，用于在单本图书和语料库层面探索新颖度轨迹（§6）。

本文其余部分安排如下。第 2 节综述相关研究。第 3 节介绍数据、嵌入流程和轨迹分析方法。第 4 节呈现定量结果。第 5 节解读研究发现并说明局限性。第 6 节介绍公开提供的工具包。

2 相关研究

远读与宏观分析。对大型文学语料库进行定量研究有着悠久的传统。Moretti（2005）创造了“远读”一词，并证明，出版元数据和体裁频率中的总体模式揭示了细读难以察觉的宏观动态。Jockers（2013）通过将计算主题建模和情感分析应用于十九世纪小说，拓展了这一研究方向，并发现不同体裁和性别的叙事形态存在系统性差异。HathiTrust 研究中心使人们能够对数百万卷文献进行词频和元数据分析（Underwood，2019年）。本研究延续了远读理念，但分析对象是密集的语义嵌入，而非词频或主题比例。

信息论方法。Shannon（1948）奠定了测量文本信息量的数学基础。此后的研究将熵、困惑度和意外度应用于文学分析。Genzel 和 Charniak（2002）表明，文本中不同位置的句子级熵大致保持恒定（即“熵率恒定性”假说）；Keller（2004）则考察了信息密度如何随叙事章节而变化。Schmidhuber（2009年）将这些观点推广为一种理论，提出以“压缩进展”作为审美趣味和好奇心的替代指标。新颖性度量可理解为意外度在嵌入空间中的类比：即新段落偏离当前语义上下文的程度。

句子嵌入与语义相似度。 基于 Transformer 的语言模型（De-vlin 等人，2019年；Vaswani 等人，2017）能够生成密集的向量表示，高度忠实地捕捉语义关系。Reimers 和 Gurevych（2019年）推出了 Sentence-BERT（SBERT），通过在自然语言推理数据上微调孪生网络和三元组网络架构，生成适用于余弦相似度比较、具有语义意义的句子嵌入。我们采用的 all-mpnet-base-v2 模型在 MTEB 基准测试中代表了当前最先进的通用句子编码器（Muennighoff 等人，2023年）。

叙事形态与情节结构。 Vonnegut（1981年）曾提出，故事遵循少数几种情感“形状”。 Reagan 等人（2016年）利用情感轨迹将这一假设转化为可操作的研究方法，并在包含 1,327 部古腾堡计划小说的语料库中识别出六种核心情感弧线。Jockers（2015年）也提取了基于情感的情节形状。本研究以语义新颖度取代情感，捕捉叙事结构的另一个维度：关注的不是人物的感受，而是文本如何探索其语义空间。

实践中的压缩进展。 Schmidhuber（2009年）及后续研究（Schmidhuber，2010年）对压缩进展、好奇心与创造力之间的关系进行了形式化。相关实证检验已应用于音乐感知（Gold 等人，2019年）、视觉艺术（Mayer 和 Landwehr，2020年）以及小规模文本实验（van de Cruys，2021年）。据我所知，此前尚无研究在语料库规模上，针对完整篇幅的书籍检验压缩进展理论的预测。

3 方法

3.1 语料库

我分析了两个大型英语书籍语料库：

PG19 （Rae 等人，2020年）：包含 28,730 本 1920 年以前的公有领域英文书籍，来源于古腾堡计划。该语料库涵盖约 1500—1919年，但以 19 世纪出版物为主。其体裁包括小说、历史、传记、哲学、科学、宗教和诗歌。

Books3 （Gao 等人，2021年）：包含 52,796 本现代英语书籍，出版日期约为 1990–2010 年。该语料库最初作为 The Pile 的一个组成部分发布，后因版权问题于 2023年停止公开分发。其内容广泛涵盖商业出版的英语书籍，包括小说、非虚构作品和参考书籍。我不会重新分发任何文本；所有分析均为非消耗性汇总统计。

这些语料库合计提供 $N = 81{,}526$ 本书，时间跨度约一个世纪，使我们能够对叙事结构演变进行历时分析。

3.2 段落分割

使用双换行符分隔符将每本书分割为多个段落。为去除页眉、页码和格式残留，少于五个词的段落会被舍弃。根据语料库和体裁的不同，每本书平均包含约 400–600 个段落。

3.3 嵌入

使用 all-mpnet-base-v2 句子转换器（Reimers 和 Gurevych，2019年）将每个段落编码为一个 768 维向量。该模型在大规模文本嵌入基准（Muennighoff 等人，2023年）上取得了最先进的性能。嵌入经过 L_2 归一化，长度为单位长度，因此余弦相似度可简化为点积。

3.4 运行质心新颖度

我将第 i 段（$i \geq 2$）的新颖性定义为：

$$\text{novelty}_i = 1 - \cos\big(\mathbf{e}_i,\ \bar{\mathbf{e}}_{1:i-1}\big) \tag{1}$$

其中，$\mathbf{e}_i$ 是段落 i 的嵌入，而 $\bar{}\mathbf{e}_{1:i-1}$ 是此前所有段落嵌入的质心（算术平均值）。这一度量反映了新段落偏离迄今为止书籍累积语义上下文的程度。数值较高表明语义探索，数值较低则表明语义巩固，或回归熟悉的领域。

运行质心以增量方式更新：

$$\bar{\mathbf{e}}_{1:i} = \frac{(i-1)\,\bar{\mathbf{e}}_{1:i-1} + \mathbf{e}_i}{i} \tag{2}$$

从而使计算量与段落数量呈线性关系。

按照惯例，由于没有前文可供衡量偏离程度，第一个段落的新颖性设为0.5（最大熵先验）。

3.5 轨迹描述符

序列（新颖性$_1$，新颖性$_2$，……，新颖性$_n$）为每本书定义了一条新颖度轨迹。我使用三类描述符来刻画这些轨迹：

轨迹聚合描述符（TAD）。遵循时间序列分析的惯例，我计算了三个汇总统计量，用以捕捉每条轨迹在嵌入空间中的几何特征：

- **速度**：新颖度序列一阶差分绝对值的平均值，用于衡量相邻段落之间新颖度变化的快慢。
- **幅度**：新颖度序列的标准差，用于衡量新颖度值的整体分布范围。
- **迂回度**：累计路径长度（绝对差值之和）与净位移（首尾新颖度值之差的绝对值）的比率。迂回度高表示轨迹呈振荡型或探索性；迂回度低则表示轨迹呈单调趋势。

分段聚合近似（PAA-16）。每条新颖度轨迹都使用分段聚合近似重采样为固定长度的 16 个分段，其中每个分段的值是书籍相应部分的新颖度平均值。由此得到一个 16 维向量，可用于聚类和距离计算。

符号聚合近似（SAX-5）。每个 PAA-16 表示都会进一步离散化为一个由 5 个字母组成的字母表中的 16 字符字符串 $\{a, b, c, d, e\}$，离散化时使用等概率高斯断点。SAX 表示有助于高效地进行符号模式匹配，并生成便于人类理解的轨迹标签。

3.6 曲线分类

根据对新颖度序列进行线性回归所得的斜率，将每条轨迹分为三种曲线类型之一：

- **红色（发散型）**：斜率为正 $> +\epsilon$；新颖性往往会随着书籍进展而增加。
- **蓝色（收敛型）**：斜率为负 $< -\epsilon$；随着书籍进展，新颖性会降低，并逐渐趋向稳定的语义语域。
- **绿色（平坦/收敛）**：斜率在 $\pm\epsilon$ 范围内；新颖性基本保持稳定。

阈值 ϵ通过经验设定，用于区分有意义的趋势与噪声。

3.7 聚类

在每个语料库中，针对 PAA-16 表示应用 K-均值聚类（$k = 8$），以识别轨迹原型。根据对聚类质心的目视检查，为聚类标识赋予描述性标签（例如“平坦”“逐渐上升”“后期激增”）。需要说明的是，每个语料库的聚类都是独立计算的；统一重新聚类将留待未来工作（§5.4）。

4 结果

4.1 语料库层级统计

表 1 展示了两个语料库的汇总统计。

表 1：语料库层级汇总统计。

统计量	PG19	Books3
分析的图书	28,730	52,796
时间范围	1920年以前	约1990—2010年
平均新颖度	0.459	0.503
新颖度标准差	—	—
平均速度	0.122	0.120
平均幅度	0.018	0.019
平均迂回度	243.8	408.5

最显著的差异在于平均新颖度：现代书籍的段落级新颖度比1920年以前的书籍高出约10%。不同年代的速度（相邻段落之间新颖度的变化率）和幅度（新颖度值的分布范围）相近。不过，迂回度——即新颖度轨迹自我折返而非单调推进的程度——在现代书籍中高出67%（408.5 对 243.8），表明其语义探索的振荡性显著更强。

4.2 体裁层面的比较

不同年代之间的新颖度差距在各主要体裁类别中均存在（表2）。

表2：不同语料库中各体裁的平均新颖度。

体裁	PG19	Books3	Δ
小说	0.464	0.499	+7.5%
非虚构作品	0.446	0.523	+17.3%

非虚构作品呈现出更大的时代差距（+17.3%），而小说为（+7.5%），这与以下假设一致：现代非虚构写作的语义范围变得更加异质——融入了案例研究、数据和跨学科参考文献，使段落进一步偏离运行质心。

4.3 曲线分类

表3展示了不同语料库中曲线类型的分布。

最显著的变化体现在绿色（收敛/平坦）曲线上，这类曲线在1920年以前的文学中更为常见。这些轨迹的特点是，书籍会逐渐趋于稳定的语义 2.3×

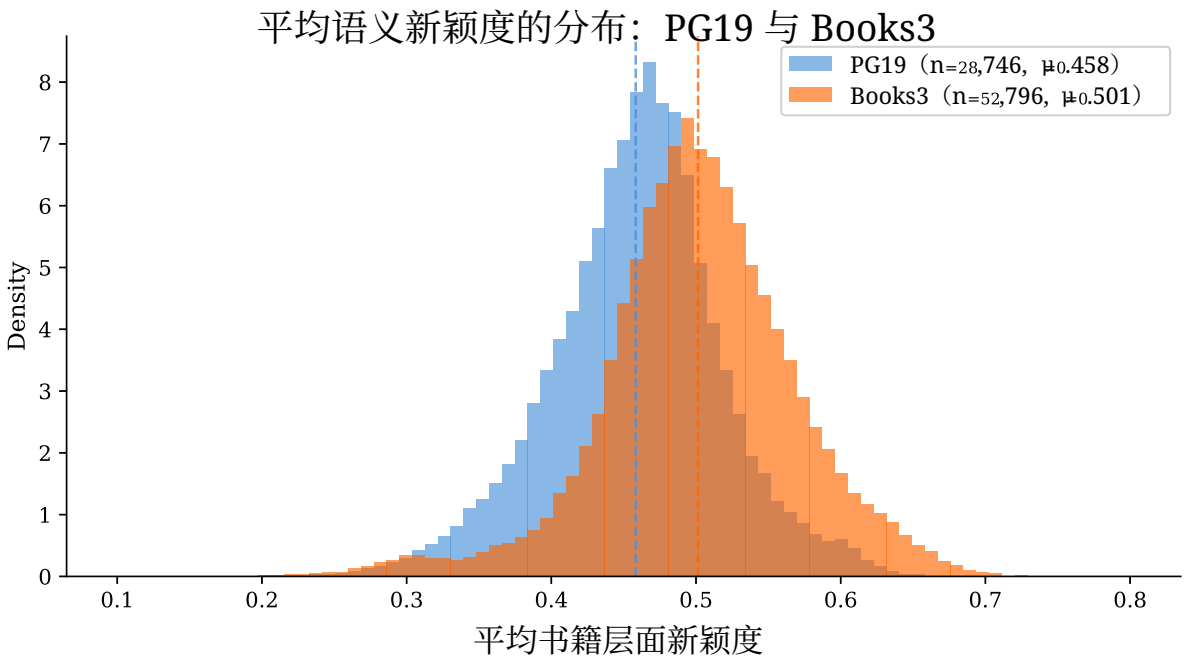

图1：PG19（蓝色）和 Books3（橙色）的平均书籍层面新颖度分布。现代语料库整体向右偏移，表明其新颖性系统性更高。

表3：曲线类型分布（占书籍总数的百分比）。

曲线类型	**PG19**	**Books3**	**比率（PG19/B3）**
红色（发散型）	47.8%	56.5%	0.85
蓝色（收敛型）	43.3%	39.7%	1.09
绿色（平坦）	9.0%	3.9%	2.31

早早确立这一风格并始终保持——这种模式符合说教类、虔修类和教学文本的特点：它们先确立自己的概念领域，再对其进行系统探索。现代语料库则以红色（发散型）轨迹为主，其中新颖性往往随着书籍内容的推进而增加。

4.4 聚类分析

基于 PAA-16 表示的 K-均值聚类揭示了每个语料库中的八种轨迹原型。表 4 比较了主要原型的流行率。

表 4：主要轨迹簇流行率（占图书的百分比）。簇根据其 PAA-16质心的形状命名。

聚类原型	PG19	Books3
Flat	26.1%	15.1%
逐渐上升	14.3%	24.7%
早期峰值	12.8%	11.2%
后期激增	8.9%	13.4%
U形	10.5%	9.8%
稳定下降	11.7%	10.1%
振荡型	9.2%	10.3%
倒U型	6.5%	5.4%

最大的分布偏移出现在平坦 聚类（PG19：26.1% $\rightarrow$ Books3：15.1%）和逐渐上升 聚类（PG19：14.3% $\rightarrow$ Books3：24.7%）。平坦轨迹的下降与渐进上升轨迹的上升共同解释了时代层面的新颖性增长的大部分：现代书籍更不可能维持稳定的语义语域，而更可能逐步扩展其概念范围。

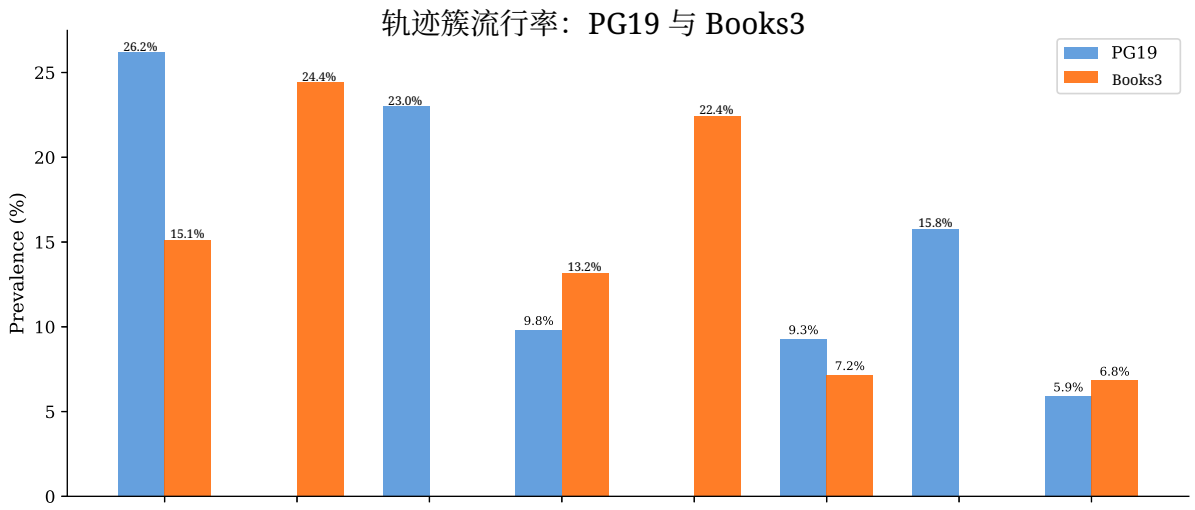

图2：PG19 与 Books3 中的轨迹簇流行率。平坦原型在1920年以前的文学中占主导地位；逐渐上升在现代语料库中占主导地位。

4.5 新颖度与读者质量

朴素趣味性理论的一个核心预测是：“越新颖 = 越受好评。”我通过将平均书籍层面新颖度与Books3语料库中可用的读者质量评分进行相关分析来检验这一点。

平均新颖度与读者评分之间的皮尔逊相关为 $r = -0.002$ ($p > 0.05$)，表明两者完全没有关系。通过运行质心方法测得的语义新颖度，与感知到的文学质量彼此正交。

对于 PG19 语料库，下载次数可作为粗略的流行度代理指标；我发现迂回度与下载次数之间存在适度的正相关（$r = 0.406$）。不过，这种相关很可能受到书籍长度的混杂影响：较长的书籍自然会产生更高的迂回度，也可能因其被认为更有价值或具有经典地位而吸引更多下载。在控制段落数后，偏相关大幅下降。

4.6 诗歌：结构变迁的案例研究

诗歌在不同时代之间经历了最显著的结构变迁。在 PG19 中，诗歌的平均迂回度在所有体裁中最低（69.4），这反映了格律诗严格的形式约束：押韵方案、诗节结构和传统意象限制了语义游移。在 Books3 中，诗歌的迂回度飙升至 481.4，在所有体裁中名列前茅，增幅达 594%。

这一转变精准追踪了从形式严谨的格律诗到实验自由诗、散文诗和语言中心诗学的历史演变。现代诗歌摆脱了格律和押韵的束缚，以振荡般的强烈态势探索语义空间，其程度甚至超过了实验小说。

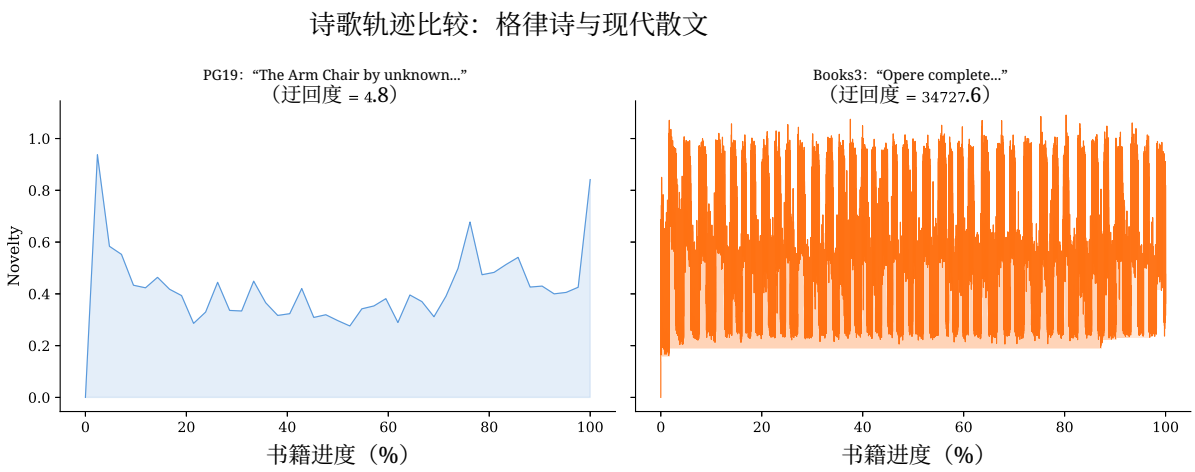

图 3：诗歌轨迹比较。左：PG19 中具有代表性的诗作，迂回度较低（语义振荡紧密）。右：Books3 中具有代表性的诗作，迂回度较高（语义探索范围广）。

4.7 体裁层面的结构平行性

尽管不同时代整体呈现出变化，某些结构上的相似性依然存在。在两个语料库中，小说都是最“蓝色”（趋同）的体裁，这与讲故事对叙事收束的预期相符。在两个时代中，非虚构作品和科学写作都更偏向“红色”（发散），反映了新概念、证据和论证的渐进式引入。

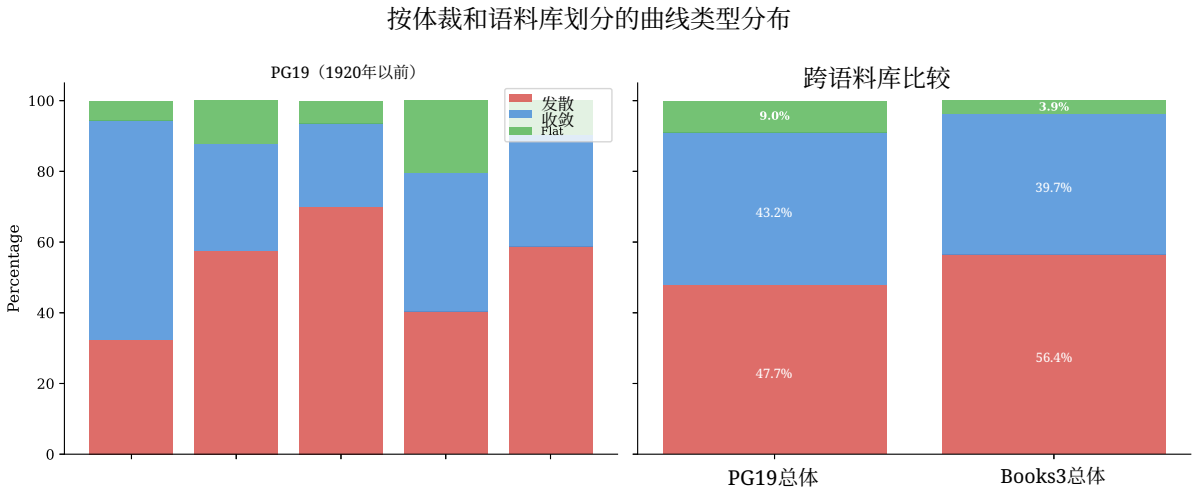

图4：不同语料库中各体裁的曲线类型分布。尽管新颖性水平整体有所变化，体裁结构特征仍跨越不同时代持续存在。

5 讨论

5.1 对现代新颖性提升的解读

现代书籍的新颖度始终高出 10%，对此可以有几种相辅相成的解释。首先，二十世纪见证了文学现代主义和后现代主义的兴起，这些运动有意识地推崇陌生化、碎片化和体裁混合——所有这些都会体现为更高的嵌入空间新颖度。其次，现代商业出版面临追求独特性的市场压力：在竞争激烈的市场中，无法在语义上脱颖而出的书籍可能会被淹没。第三，现代非虚构语料库反映出一种日益跨学科的知识文化，单本书籍可能涉及神经科学、经济学、历史和个人轶事，从而产生高度的语义异质性。

值得强调的是，更高的新颖性并不等同于更高的质量。新颖性与读者评分之间的零相关（§4.5）凸显了这一点。Schmidhuber 的理论预测，压缩进展——即观察者模型改进的速率——才是驱动趣味性的因素，而非新颖性本身。一本新颖性达到极致（即语义随机）的书，在这一指标上会得分很高，但读者体验到的会是不连贯，而非有趣。趣味性的最佳点位于最优压缩进展区域：新材料既足够出人意料，能够推动模型更新，又具有足够的结构性，便于学习。

5.2 收敛叙事的衰退

收敛型（绿色）轨迹——即新颖性逐渐下降并趋于稳定语域的书籍——在1920年以前的文学中更为常见，比例为2.3× 。这一模式与十九世纪说教与道德文学的盛行相符：布道文、行为规范手册和哲学论著先确立论点，再系统地加以阐述。现代时期收敛曲线的减少，或许反映了文化从说教式收束转向开放式探索、认识论不确定性与刻意的模糊性。

5.3 诗歌作为一种自然实验

诗歌从迂回度最低的体裁（PG19：69.4）转变为迂回度最高的体裁之一（Books3：481.4），提供了一个引人注目的自然实验。1920年前诗歌的形式约束——格律、押韵和固定的诗节形式——严格限制了语义探索。每一行都必须符合音韵约束，这些约束限制了可用词汇，进而限制了相邻诗行之间可跨越的语义距离。现代主义诗歌和后现代主义诗歌摒弃这些约束，使轨迹摆脱束缚，形成了我观察到的振荡型、高迂回度模式。

5.4 局限性

这些发现存在若干局限。

独立聚类。 我们分别对每个语料库运行独立的 K-means 算法，计算轨迹簇。虽然我根据质心形状为对应簇赋予了相应标签，但在合并后的语料库上进行统一聚类，才能为分布比较提供更严谨的依据。这将作为未来工作开展。

语料库来源。 Books3语料库（Gao 等人，2021年）最初由 Bibliotik 私人追踪站汇编，后因版权问题于2023年停止公开分发。其构成可能无法代表所有现代英语出版物；选择偏差（例如某些出版商或体裁占比过高）也可能影响汇总统计数据。我强调，本研究属于非消费性计算分析：我不会重新分发 Books3 中的任何文本，且所有报告的指标均为汇总统计数据，无法据此还原任何单本书的文本。这一分析框架遵循 Authors Guild v. HathiTrust （2014年）和 $AuthorsGuild v. Google$ （2015年）所确立的非消费性使用传统。

时间混杂因素。 比较1920年以前与1990年以后的语料库，会将多种变异来源混为一谈：文学惯例、受众、教育水平、印刷技术以及作者职业化方面的变化。要厘清这些因素，需要进行逐十年分析，而PG19语料库在1800年以前的覆盖较为稀疏，使这一分析难以开展。

嵌入模型偏差。 all-mpnet-base-v2模型主要基于现代英语文本训练。其对古语词汇、正字法和句法结构的表征，在语义忠实度上可能低于对现代散文的表征。这可能会系统性压低PG19语料库中测得的新颖性；不过，该效应在各体裁中方向一致，表明这10%的差距并非纯粹是人为假象。

段落分割。 双换行分割是一种启发式方法，对大多数散文效果良好，但在诗歌、戏剧和书信体文本中可能产生不一致的单元。针对特定体裁的分割策略可以提高这些形式的轨迹保真度。

6 可用性

所有分析代码、预计算的轨迹数据以及交互式探索工具均已公开提供：

- **Novelty Hub**：https://bigfivekiller.online/novelty_hub — 语料库层级统计、跨语料库比较仪表板和聚类可视化。

- **新颖性探索器**：https://bigfivekiller.online/novelty_explorer — 单本书籍轨迹可视化，并叠加分段聚合近似、符号聚合近似和曲线分类。
- **基础设施**：基于 Streamlit、SQLite 和 sentence-transformers 库构建（Reimers 和 Gurevych，2019年）。

底层轨迹数据库存储在 SQLite 中，支持对完整的 81,526 本书语料库进行任意查询。

7 结论

我展示了首个针对英语书籍语义新颖度轨迹的大规模跨语料库分析，并通过句子变换器嵌入将 Schmidhuber 的压缩进展理论付诸实践。对 81,526 本书的分析揭示了 1920年以前与现代文学之间新颖性结构的系统性变化：平均新颖度更高、轨迹迂回度更大、收敛叙事曲线更少，诗歌的语义结构也发生了显著转变。与此同时，新颖性与读者评分之间的零相关提醒我们，不应将语义惊奇等同于文学价值。

这些发现为未来工作开辟了若干方向：按年代逐一进行时间分解、统一的跨语料库聚类、作者级轨迹指纹识别，以及结合情感和主题轨迹，对叙事形态进行多维度分析。

致谢

作者感谢古腾堡计划和 PG19 数据集的维护者，使大规模文学分析成为可能。该交互式工具包使用 Streamlit、SQLite 和 sentence-transformers 库构建。计算分析通过 Claude Code（Anthropic）进行流水线编排，并使用 Google Cloud Platform 进行公开部署。

参考文献

Blei, D. M., Ng, A. Y., 和 Jordan, M. I.（2003）。潜在狄利克雷分配。《机器学习研究杂志》，3：993–1022。Devlin, J., Chang, M.-W., Lee, K., 和 Toutanova, K.（2019）。BERT：用于语言理解的深度双向变换器预训练。载于*NAACL-HLT 2019*论文集，第4171–4186页。Gao, L., Biderman, S., Black, S., Golding, L., Hoppe, T., Foster, C., Phang, J., He, H., Thite, A., Nabeshima, N., Presser, S., 和 Leahy, C.（2021）。The Pile：一个包含多样化语言建模文本的800GB数据集。*arXiv*预印本 *arXiv:2101.00027*。Genzel, D. 和 Charniak, E.（2002）。文本中的熵率恒定性。载于*ACL 2002*论文集，第199–206页。Gold, B. P., Pearce, M. T., Mas-Herrero, E., Dagher, A., 和 Bhatt, S. M.（2019）。音乐愉悦感中的可预测性与不确定性：学习的奖励？《神经科学杂志》，39(47)：9397–9409。Jockers, M. L.（2013）。《宏观分析：数字方法与文学史》。伊利诺伊大学出版社。Jockers, M. L.（2015）。使用 Syuzhet软件包揭示情感与情节弧线。博客文章，http://www.matthewjockers.net/2015/02/02/syuzhet/。

Keller, F.（2004）。作为加工 effort预测指标的熵率原理：基于眼动追踪数据的评估。载于*EMNLP 2004*论文集，第317–324页。Mayer, S. 和 Landwehr, J. R.（2020）。设计典型性的客观测量。*Design Studies*，69：100–940。Moretti, F.（2005）。《图表、地图、树：文学史的抽象模型》。Verso Books。Muennighoff, N., Tazi, N., Magne, L., 和 Reimers, N.（2023）。MTEB：大规模文本嵌入基准。载于*EACL 2023*论文集，第2014–2037页。Rae, J. W., Potapenko, A., Jayakumar, S. M., Hillier, C., 和 Lillicrap, T. P.（2020）。用于长程序列建模的压缩型 Transformer。载于*ICLR 2020*论文集。Reagan, A. J., Mitchell, L., Kiley, D., Danforth, C. M., 和 Dodds, P. S.（2016）。故事的情感弧线主要由六种基本形状构成。*EPJ Data Science*，5(31)。Reimers, N. 和 Gurevych, I.（2019）。Sentence-BERT：使用孪生 BERT 网络生成句子嵌入。载于*EMNLP-IJCNLP 2019*论文集，第3982–3992页。Schmidhuber, J.（2009）。关于主观美、新颖性、惊奇、趣味性、注意力、好奇心、创造力、艺术、科学、音乐和笑话的简单算法理论。*arXiv*预印本 *arXiv:0812.4360v2*。Schmidhuber, J.（2010）。创造力、乐趣与内在动机的形式理论（1990–2010）。*IEEE*《自主心理发展汇刊》，2(3)：230–247。Shannon, C. E.（1948）。通信的数学理论。*Bell System Technical Journal*，27(3)：379–423。Underwood, T.（2019）。《遥远的地平线：数字证据与文学变迁》。芝加哥大学出版社。van de Cruys, T.（2021）。压缩与机器学习：关于特征空间向量的新视角。载于*ACL 2021*论文集，第5083–5095页。Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., 和 Polosukhin, I.（2017）。注意力机制就是你所需的一切。载于神经信息处理系统进展 *30*，第5998–6008页。Vonnegut, K.（1981）。故事的形状。演讲，后收入《没有国家的人》（2005年）。