Modern Linguistics 现代语言学, 2026, 14(5), 359-365
Published Online May 2026 in Hans. https://www.hanspub.org/journal/ml
https://doi.org/10.12677/ml.2026.145411

Hans汉斯

基于语料库的中英文学文本语言特征对比分析

——以《骆驼祥子》与《野性的呼唤》为例

吴康雄，刘书亮*

江西理工大学外国语学院，江西 赣州

收稿日期：2026年3月29日；录用日期：2026年5月6日；发布日期：2026年5月18日

摘　要

本研究运用语料库文体学方法，对老舍《骆驼祥子》英译本与杰克·伦敦《野性的呼唤》原文进行对比分析。研究依托自建可比语料库，使用Wordsmith、AntConc等工具，从词汇、句法、语篇层面对平均词长、词汇密度、标准化类符/形符比、高频词、句长及连词使用等参数展开量化统计。在本研究选取的《骆驼祥子》英译本与《野性的呼唤》原文的对比中，观察到前者词汇密度更高，偏好从属连词构建主从复合句，信息承载更密集，符合翻译文本显化特征与现实主义叙事需求；后者的英文原文词汇多样性更强，多用并列连词，句式紧凑、节奏明快，契合自然主义文本的叙事手法。本研究也为文学文本跨语境对比与语言特征研究提供了参考。

关键词

语言特征，文学文本，语料库研究

Corpus-Based Comparative Analysis of Linguistic Features in Chinese and English Literary Texts

—A Case Study of *Rickshaw Boy* and *The Call of the Wild*

Kangxiong Wu, Shuliang Liu*

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

Received: March 29, 2026; accepted: May 6, 2026; published: May 18, 2026

*通讯作者。

文章引用: 吴康雄, 刘书亮. 基于语料库的中英文学文本语言特征对比分析[J]. 现代语言学, 2026, 14(5): 359-365.
DOI: 10.12677/ml.2026.145411

Abstract

This study, grounded in corpus-based research methods, conducts a comparative analysis of the linguistic features in Lao She's *Rickshaw Boy* and Jack London's *The Call of the Wild* across the lexical, syntactic, and discourse levels. By constructing a corpus and employing text analysis tools such as Wordsmith and AntConc, it quantitatively examines parameters including average word length, lexical density, type-token ratio, high-frequency words, sentence length, conjunction usage. In the comparison between the selected English translation of *Rickshaw Boy* and the original text of *The Call of the Wild*, it is observed that the former exhibits a higher lexical density, a preference for subordinating conjunctions to construct complex sentences, and denser information load, which aligns with the explicitation features of translated texts and the narrative demands of realism. The latter, as an original English text, shows greater lexical diversity, more frequent use of coordinating conjunctions, and a compact, fast-paced sentence structure, which fits the narrative way of naturalist texts. This study also serves as a reference for the cross-contextual comparison of literary texts and the investigation of linguistic features.

Keywords

Linguistic Features, Literature Texts, Corpus-Based Study

1. 引言

文学作品的文体特征是作者独特创作风格与深层文化背景的外在投射，而不同语言体系因本质差异与叙事传统的历史分野，形成了各具辨识度的语言表达范式[1]。老舍创作的《骆驼祥子》作为中国现代文学的经典文本，以底层劳动者的命运轨迹为叙事主线，凭借质朴凝练的语言承载深沉厚重的情感内涵；杰克·伦敦的《野性的呼唤》则秉持自然主义创作笔法，聚焦动物在极端环境中的生存博弈，其语言风格兼具野性感与生命力量。《骆驼祥子》的英译本 *The Rickshaw Boy* 与《野性的呼唤》原文 *The Call of the Wild* 为开展特定英语文学文本与汉译英文学文本语言特征的对比研究提供了文本素材。

语料库的兴起促使翻译研究从传统的语言学与比较语言学视角，拓展至政治、意识形态、经济及文化等宏观层面。基于这一方法，研究者得以揭示翻译实践中的概然性规律，系统归纳出翻译文本及翻译过程中的“翻译总特征”，从而实现对翻译现象更为全面、广角的认知[2]。本文以自建的“《骆驼祥子》英译本《野性的呼唤》英文原文可比语料库”为研究基础，运用语料库文体学的核心研究方法，从词汇复杂度、句法结构及语篇衔接三个核心维度，对《骆驼祥子》英译本与《野性的呼唤》英文原文展开系统性对比分析。基于上述研究框架，本文设定的核心研究问题如下：(1) 两文本在词汇多样性、信息密度及语言正式程度三个维度存在何种具体差异？(2) 在句法层面，句式复杂度与文本可读性的差异如何映射并强化两者的叙事风格分野？(3) 语篇衔接手段的运用特征的分布规律，如何具象化呈现本研究所涉文本语篇衔接的本质差异？

2. 语料库文体学研究方法

本研究采用语料库文体学的研究方法，通过量化与定性分析相结合的方式，对《骆驼祥子》英译本

The Rickshaw Boy 与《野性的呼唤》英文原版 *The Call of the Wild* 语言特征进行对比。具体步骤如下。

2.1. 语料收集与处理

搜集《骆驼祥子》英译本 *The Rickshaw Boy* 与《野性的呼唤》英文原版 *The Call of the Wild* 的完整文本，确保语料的真实性与代表性。

对语料进行清洗与标准化处理，去除无关符号与格式噪音，并进行分词与词性标注(使用 CLAWS 等工具)。

2.2. 语料库构建

分别建立《野性的呼唤》原文以及《骆驼祥子》英译本的两个子语料库，确保语料容量均衡。

使用 Wordsmith 和 AntConc 等软件对语料进行统计分析，提取词汇、句法和语篇层面的特征数据。

2.3. 分析维度

词汇层面：通过平均词长、词汇密度、类符/形符比等指标，分析词汇复杂度与多样性。

句法层面：计算平均句长、句式类型(简单句、复合句等)及连接词使用频率，评估句式结构与逻辑性。

语篇层面：考察衔接手段(如连词)的分布情况，分析语篇的连贯性与紧凑性。

通过以上方法，本研究旨在客观揭示两种文学文本的文体特征差异，为文学语言研究提供实证依据。

3. 文体特征对比结果

本文基于自建语料库，对《野性的呼唤》英文原版以及《骆驼祥子》英译本的文本从词汇、句法和语篇三个层面进行语言特征对比分析。

3.1. 词汇方面

3.1.1. 平均词长和长词数据

平均词长指文本中词汇的平均字母数，用于衡量词语长度；词长标准差则反映各单词长度与平均词长的偏离程度。一般而言，平均词长越大，文本中的复杂词汇越多。将《骆驼祥子》英译本与《野性的呼唤》英文原文分别导入 Wordsmith7.0，执行 WordList 功能后，即可获得这两个子库的平均词长与词长标准差，如表 1 和表 2：

Table 1. The two sub-corpora's average word length and standard deviation of word length
表 1. 两个子库的平均词长及词长标准差

参数类型	*The Rickshaw Boy*	*The Call of the Wild*
平均词长	4.26	4.26
词长标准差	2.14	2.11

Table 2. The two sub-corpora's total long words and their proportion
表 2. 两个语料子库的长词总数与占比

参数类型	*The Rickshaw Boy*	*The Call of the Wild*
长词	14,563	4605
总词数	95,046	32,382
长词占比	15.32%	14.22%

根据表 1 数据，《野性的呼唤》与《骆驼祥子》在平均词长上持平，前者的词长标准差略小于后者，说明两部作品在词汇复杂度上处于同一水平，但前者的正式程度略高于前者，几乎相同。

另外，英语中的长词通常将 6 个字母以上的词。表 2 数据表明，《骆驼祥子》子库的长词占比略高于《野性的呼唤》，尽管两部作品的题材、时代和作者迥异，但它们在词汇层面的复杂程度或正式程度上可能具有相似性。长词占比的微小差异可能指向更细微的文体风格或叙事视角的区别，但整体而言，两部作品并未在用词的显性复杂度上表现出显著差异。

3.1.2. 词汇密度

词汇密度指实词数量在总词数中所占的百分比，用于衡量文本的信息量与词汇丰富程度；英语中的实词主要包括名词、实义动词、形容词和副词四类，如表 3：

Table 3. The total number of content words and the lexical density of the two sub-corpora
表 3. 两个子库的实词总数及词汇密度

参数类型	*The Rickshaw Boy*	*The Call of the Wild*	X2
名词	17,352	4605	13.4946
实义动词	15,082	32,382	344.3557
形容词	5384	2105	0.2388
副词	5183	1592	67.7302
实词总数	43,001	14,424	385.8861
总词数	93,627	36,162	
词汇密度	45.92%	39.88%	

由表 3 数据可知，《骆驼祥子》英译本词汇密度高达 45.92%，表明其实词比例高，文本信息密集、内容具体，语言承载的实质性信息量大；它通过高频、高比例地使用实词(尤其是名词和动词)，构建出一个信息稠密、叙述具体、细节丰富的写实世界，以匹配其描绘广阔社会图卷的需求。而《野性的呼唤》英文原文的词汇密度为 39.88%，相对较低，意味着功能词(如介词、连词)比例更高，句式可能更复杂或更偏重逻辑衔接，信息呈现相对稀疏。采用了相对凝练、聚焦的实词使用模式，将语言重心更多地投向有限核心要素的深度刻画与内在变化，以适应其探索自然与野性本质的主题。

3.1.3. 类符/形符比

在语料库语言学领域，类符指的是文本中不重复且忽略大小写后的不同词语，形符指的是文本中所有词语出现的总频次，即语料的总词数[3]。通过计算类符与形符的比值，可以初步判断译者在用词上的差异：该比值越高，说明文本中使用的不同词汇越丰富；反之，则表明词汇变化相对有限。然而，这一比值容易受到文本长度的影响，为消除这一干扰，Scott 提出了标准化类符/形符比(standardized type/token ratio, STTR)，将其作为衡量词汇多样性的更稳健指标。STTR 的具体算法是：先将文本按每 1000 词划分为若干区间，分别计算每个区间的 TTR 值，然后对所有区间的 TTR 结果取算术平均数，如表 4。

Table 4. The two sub-corpora's type-token ratio
表 4. 两个子库的类符与形符比

参数类型	*The Rickshaw Boy*	*The Call of the Wild*
类符	8118	4750
形符	95,046	32,382
标准化类/形符比率(%)	8.54%	14.73%

表 4 数据表明，从类符与形符数量来看，《骆驼祥子》的类符(8118)与形符(95,046)均远超《野性的呼唤》的类符(4750)与形符(32,382)，表明前者的文本篇幅更长、整体语料规模更大；从标准化类/形符比率(反映词汇丰富度的核心指标)来看，《野性的呼唤》以 14.73%的比率显著高于《骆驼祥子》的 8.54%，说明前者的词汇多样性更强，文本中出现的不同词汇类型占比更高。综合来看，《骆驼祥子》虽凭借更大的文本体量拥有更多的词汇总数，但词汇重复使用的频率相对更高；而《野性的呼唤》则在有限的文本规模内展现出更优的词汇丰富度，词汇运用的多样性更突出。

3.1.4. 高频词

高频词指在文本中反复出现、出现频率显著高于一般词语的词汇，这类词不仅构成了文本的主体内容，奠定了文本的基本词汇基调，同时也是考察译者语言习惯与文本特征的重要分析对象。根据 Laviosa 的观点，当一个词在语料库中出现的频率达到总词数的千分之一及以上(即占比 ≥0.10%)时，便可归为高频词[4]。本研究沿用这一标准，借助 Wordsmith 7.0 生成词频表，并从中提取出两个语料子库中满足该条件的高频词，相关数据汇总于表 5。

Table 5. High-frequency words in the two sub-corpora
表 5. 两个子库的高频词数据

参数类型	*The Rickshaw Boy*	*The Call of the Wild*
高频词数目	142	125
累计比例	7.56%	14.55%
高频词重复率	66.68	71.98
高频词与低频词之比	0.1632	0.1703

表 5 的数据表明，两个语料子库在高频词数量方面存在一定程度的差别。对比分析显示，尽管《骆驼祥子》的高频词总数(142 个)多于《野性的呼唤》(125 个)，但后者的词汇使用呈现出更高的集中性和重复性。具体表现为后者的高频词累计比例(14.55%)显著高于前者(7.56%)，同时其高频词重复率(71.98%)也更高。这表明在《野性的呼唤》中，数量较少的核心词汇承担了更大比例的文本表达功能，且这些词汇的内部复用更为频繁。

3.2. 句法层面

3.2.1. 平均句长

平均句长及句长标准差，通常被用作评估作者文体风格的重要参数指标。平均句长是衡量文本句子难易程度的一个标准，其指代的是句子长度的平均值，即：平均句长 = 形符数/句子数。本文的理解难度一般随平均句长的增加而上升，反之则下降。句长标准差用于衡量句子长度相对于平均句长的波动情况，其数值越大，意味着句子长短差异越显著，句式越富变化，可读性也随之增强。如表 6。

Table 6. Average sentence length of the two sub-corporas
表 6. 两个子库的平均句长数据

参数类型	*The Rickshaw Boy*	*The Call of the Wild*
句子个数	5194	1686
平均句长	18.30	19.21
句长标准差	11.81	12.39

表 6 数据显示，《骆驼祥子》的平均句长数值小于《野性的呼唤》的平均句长数值，显示其句法结构可能更为复杂，但二者仅 0.91 词的微小差距表明其整体可读性水平相近。在句长分布上，两部作品均表现出较高的离散性(句子标准差分别为 12.39 和 11.81)，说明它们都采用了长短句交替的句式结构，此分布特征是形成各自文体节奏的重要基础。此外，《骆驼祥子》的句子总量(5194 句)约为《野性的呼唤》(1686 句)的三倍，这一样本量差异使得前者的句长数据具有更高的统计稳定性。

3.2.2. 句子结构类型

从结构角度划分，英语句子共有三类，即简单句、并列句与复合句。简单句的核心特征是有且仅有一个主谓结构；并列句是用并列连词把两个以上的简单句串联起来；复合句则依靠从属连词实现多个简单句的组合。将已完成词性标注的两个语料子库导入 AntConc，利用其 Word 功能，即可检索出并列连词(CC, CCB)与从属连词(CS, CSA, CSN, CST, CSW)的出现频次，如表 7。

Table 7. Connecting words in the two sub-corporas
表 7. 两个子库的相关连词数据

参数类型	*The Rickshaw Boy*	*The Call of the Wild*	X2
从属连词	2818	801	60.7917
并列连词	3771	1853	75.6568
总词数	95,046	32,382	
从属连词占比	2.96%	2.48%	
并列连词占比	3.97%	5.72%	

表 7 数据显示，《骆驼祥子》显著更多地使用从属连词(绝对数量与占比均更高)，句式结构偏向主从复合。语言特征服务于小说复杂的社会写实与心理剖析主题，通过构建多层次、逻辑严密的句子，深入刻画人物命运、社会关系与内心矛盾。

与之相对，《野性的呼唤》并列连词的使用比例显著更高，句式结构偏向并列平行。这种语言特征营造出简洁、明快、富有动感的叙事节奏，通过连接一系列短促的动作与事件，生动再现了荒野生存中的直接体验与原始冲动，契合其自然主义与冒险故事的主题。

3.3. 语篇层面

3.3.1. 语篇衔接

Halliday (韩礼德)指出，衔接是一个语义概念，所谓衔接，指的是语篇内部各语言成分在语义层面相互关联的特性，合理运用衔接手段有助于构建文本前后之间的逻辑一致性[5]。逻辑联系语作为一种衔接方式，能表达句间的多种语义关系，其形式包括词(连词、连接副词)、短语以及分句(限定或非限定)，两个字语料库数据如表 8。

Table 8. Discourse connectives in the two sub-corporas
表 8. 两个子库的逻辑联系语数据

参数类型	*The Rickshaw Boy*	*The Call of the Wild*	X2
从属连词	2818	801	60.7917
并列连词	3771	1853	75.6568
连接副词	1060	267	39.9774
总词数	95,046	32,382	
逻辑联系语占比	8.05%	9.02%	

表 8 表示，尽管两个语料库在逻辑联系语的总体使用比例上接近，但在所有三类具体逻辑联系语(从属连词、并列连词、连接副词)的绝对使用数量上，《骆驼祥子》均远超《野性的呼唤》。由于两个语料库的总词数不同(分别为 95,046 词和 32,382 词)，这种数量上的差距可能部分源于文本规模差异，但极高的卡方值表明，前者的作者或文本类型可能更倾向于频繁使用各类连接词来构建句间和从句间的逻辑关系。

4. 结语

在本研究选取的《骆驼祥子》英译本与《野性的呼唤》英文原文的对比中，观察到了二者在词汇、句法和语篇层面的语言特征各有不同。尽管两部作品均为文学经典，但由于文化背景、创作意图及目标读者的差异，二者呈现出显著的文体分野。在词汇层面，《野性的呼唤》凭借较高的长词比例与标准化类符/形符比，体现出较强的词汇丰富度与叙事流畅性；而《骆驼祥子》英译本则通过更高的词汇密度与实词比例，增强了文本的信息承载能力，这在一定程度上反映了译者对原作社会批判内涵的忠实传达。句法方面，《野性的呼唤》原文借助长句与灵活句式提升语篇逻辑连贯性；相较之下，《骆驼祥子》英译本对从属连词侧重(后者在汉英文学翻译中常伴随人称代词的显化现象[6])，则凸显了中国文学对人物内在经验的主观聚焦。

然而，本研究在语料选取方面存在若干局限。首先，比较对象分别为翻译文本与源语文本，其中《骆驼祥子》英译本可能受到译者策略(如归化或异化)的影响，导致跨语言对比时产生潜在偏误。翻译过程中为兼顾原文表达习惯与目标语读者接受度，其语言特征未必完全反映原作的文体风格。此外，仅选取两部作品进行对比，样本规模有限，语料库的代表性与多样性不足，可能影响结论的普遍适用性。

其次，研究未充分纳入文化背景与主题差异对语言特征的塑造作用。两部作品在文化内涵与创作意图上差异显著(例如《骆驼祥子》的社会批判取向与《野性的呼唤》的自然主义倾向)，但量化方法难以完全揭示语言形式与主题思想之间的深层关联[7]。例如，词汇密度差异不仅关乎文体选择，也可能与叙事焦点密切相关，而本研究对此未作深入探讨。

在分析方法层面，研究工具与方法亦存在一定的局限性。词性标注工具(如 CLAWS)对文学文本中隐喻、方言等特殊用法的处理可能不够准确，连词分类标准若不一致也可能影响统计结果的可信度。此外，语料库的平衡性有待进一步考量：两部文本的长度差异较大(《骆驼祥子》英译本词数远多于《野性的呼唤》)，可能对词汇重复率、类符/形符比等指标的跨文本可比性造成干扰。未来研究应在扩大语料规模、控制变量、完善方法验证等方面持续推进。

参考文献

[1] 陈伟. 翻译英语语料库与基于翻译英语语料库的描述性翻译研究[J]. 外国语, 2007(1): 67-73.

[2] 申丹. 叙事学与小说文体学研究[M]. 北京: 北京大学出版社, 1998.

[3] Baker, M. (1995) Corpora in Translation Studies: An Overview and Some Suggestions for Future Research. *Target. International Journal of Translation Studies*, **7**, 223-243. https://doi.org/10.1075/target.7.2.03bak

[4] Laviosa, S. (2002) Corpus-Based Translation Studies. Brill. https://doi.org/10.1163/9789004485907

[5] Halliday, M.A.K. and Hasan, R. (1976) Cohesion in English. Longman.

[6] 王克非, 胡显耀. 汉语文学翻译中人称代词的显化和变异[J]. 中国外语, 2010, 7(4): 16-21.

[7] Stubbs, M. (2005) Conrad in the Computer: Examples of Quantitative Stylistic Methods. *Language and Literature*: *International Journal of Stylistics*, **14**, 5-24. https://doi.org/10.1177/0963947005048873