Modern Linguistics 现代语言学, 2026, 14(5), 824-830
Published Online May 2026 in Hans. https://www.hanspub.org/journal/ml
https://doi.org/10.12677/ml.2026.145464

Hans汉斯

汉译英翻译体的量化特征与大语言模型识别研究

张旻璐，李华东

上海海事大学外国语学院，上海

收稿日期：2026年3月19日；录用日期：2026年5月13日；发布日期：2026年5月27日

摘　要

为构建汉译英翻译体的核心风格特征体系并评估国内主流大语言模型的识别效果，本研究以106条英语原生文本与汉译英翻译文本为语料，选取词汇丰富度(TTR)、平均句长、虚词占比及标点密度四类量化指标，系统探究翻译体的多维特征，并对比豆包与DeepSeek的文本识别能力。研究发现：汉译英翻译体呈现“词汇丰富度偏低、平均句长略短、低虚词占比、高标点密度”的显著特征，其中词汇丰富度是区分两类文本最稳定的指标，印证了翻译普遍性假说与汉英语言转换的独特规律；模型评估结果显示，DeepSeek整体识别正确率达97.17%，仅在80词以内超短文本中存在特征密度不足导致的局部偏差，而豆包正确率为83.02%，存在特征识别维度单一、领域化适配不足的系统性偏差。本研究明确了汉译英翻译体的核心量化特征体系，厘清了国内大语言模型的识别能力差异与问题根源，为翻译体量化研究、机器翻译优化及语料库建设提供了实证依据与技术参考。

关键词

汉译英，翻译体，量化特征，大语言模型

A Study on Quantitative Features of Chinese-to-English Translationese and LLM-Based Identification

Minlu Zhang, Huadong Li

School of Foreign Languages, Shanghai Maritime University, Shanghai

Received: March 19, 2026; accepted: May 13, 2026; published: May 27, 2026

Abstract

To construct a core stylistic feature system for Chinese-to-English translationese and evaluate the

文章引用: 张旻璐, 李华东. 汉译英翻译体的量化特征与大语言模型识别研究[J]. 现代语言学, 2026, 14(5): 824-830.
DOI: 10.12677/ml.2026.145464

identification performance of mainstream domestic large language models, this study employs a corpus of 106 texts, including both native English texts and Chinese-to-English translated texts. Four quantitative indicators are selected: Type-Token Ratio (TTR), average sentence length, function word ratio, and punctuation density, to systematically explore the multi-dimensional characteristics of translationese and compare the text identification capabilities of Doubao and DeepSeek. The results reveal that Chinese-to-English translationese exhibits significant features of "lower lexical richness, slightly shorter average sentence length, lower function word ratio, and higher punctuation density". Among these indicators, lexical richness is the most stable index for distinguishing the two types of texts, which supports the translation universals hypothesis and the unique laws of Chinese-English language transformation. The model evaluation results show that DeepSeek achieves an overall identification accuracy of 97.17%, with only local deviations in ultra-short texts within 80 words due to insufficient feature density. In contrast, Doubao obtains an accuracy of 83.02%, suffering from systematic biases such as single-dimensional feature recognition and insufficient domain adaptation. This study establishes a core quantitative feature system for Chinese-to-English translationese, clarifies the differences in identification capabilities and root causes of domestic large language models, and provides empirical evidence and technical references for quantitative research on translationese, machine translation optimization, and corpus construction.

Keywords

Chinese-English Translation, Translationese, Quantitative Features, Large Language Models

1. 引言

1.1. 研究背景

随着全球化进程与自然语言处理技术的发展，机器翻译与人工翻译的应用日益广泛，但翻译体现象严重影响了译文的可读性与自然度。翻译体这一概念最早由 Nida 与 Taber 于 1969 年提出，指偏离目标语言规范的非自然语言变体，其典型特征包括词汇使用僵化、句法冗余以及语义表达生硬[1]。

Shuttleworth 与 Cowie 在《翻译研究词典》(1997)中进一步将其定义为一类带有评价倾向的术语，指因明显依附源语言特征而被认为不符合目标语言表达习惯、理解难度较高的目标语言使用形式[2]。这种倾向导致译文偏离目标语言常规，在文学、学术、商务等对语言自然度要求极高的场景中，负面影响尤为突出。因此，精准识别翻译体并提取其核心特征，已成为翻译研究与自然语言处理领域的关键课题。

1.2. 文献综述

译文的核心特征、识别方法及文体量化分析，是翻译研究与自然语言处理交叉领域的核心议题。Baker 提出“翻译普遍性”假说，指出译文存在显化、简化、规范化等固有特征，为后续研究奠定了理论框架[3]。近年来，翻译体量化研究与大语言模型识别评估成为领域内研究热点，相关成果为汉译英翻译体的特征挖掘与模型测试提供了多元思路与方法参考，本研究也将在现有前沿研究基础上，进一步聚焦汉译英场景的本土化特征与国内模型的识别表现。

特征研究聚焦词汇、句法、语义三维度：词汇层面，Laviosa 证实译文词汇丰富度低于源语文本，王子瑞等补充了译者个体文体印记的影响[4] [5]；句法层面呈现明显语言对差异，Liu & Afzaal 发现汉英译

文句法复杂度低于英语原生文本，而秦洪武、王克非基于英译汉语料指出英语复杂句法易被拆分，二者印证翻译体句法特征的语言对依赖性，汉译英场景下翻译体的系统性特征仍有待结合本土化语料进一步提炼与验证[6] [7]；语义层面，Wang & Jiang 揭示译文语义显化与“折中”特性，Church 等强调翻译方向的影响，但未深入探讨语义与词汇、句法特征的内在关联[8] [9]。

识别技术方面，Lembersky 等通过混合模型增强特征捕捉，Zhang 利用注意力机制神经网络提升识别精度，但现有研究多聚焦国外模型与通用场景，缺乏对国内主流模型(如豆包、DeepSeek)的针对性评估与能力归因，难以满足国内实践需求[10] [11]。

综合现有研究，当前领域核心缺口有三：其一，特征研究多聚焦单一维度，缺乏汉译英场景下“词汇–句法–衔接”三维量化体系的系统构建，未能充分体现汉英语言转换独特规律及各维度特征的协同作用；其二，大语言模型识别研究“重算法、轻评估”，缺乏对国内主流模型的针对性测试与偏差归因；其三，短文本(80 词以内)、专业领域文本的翻译体特征捕捉无统一标准，相关场景识别策略仍需完善。

本研究针对上述缺口，结合汉译英本土化语料特征，构建三维量化特征体系，聚焦国内主流大语言模型的识别效果与差异，兼顾特殊场景特征分析，旨在填补汉译英翻译体系统性研究与国内模型评估空白，为识别技术优化提供实证支撑。

1.3. 研究问题

针对翻译体识别的现实需求与现有研究的局限性，本研究聚焦两大核心研究问题：其一，适配汉译英场景的翻译体核心风格特征体系应如何构建？拟从词汇、句法、衔接三个维度切入，系统挖掘并提炼基于本研究语料的汉译英翻译体具有普适性的核心风格特征，明确各维度特征的内在关联与表征规律。其二，如何实现汉译英翻译体与英语原生文本的精准有效区分？国内主流大语言模型(如豆包、DeepSeek)在汉译英翻译体与英语原生文本的识别任务中表现如何，其识别能力的差异与核心影响因素是什么。

1.4. 研究意义

本研究围绕汉译英场景下翻译体核心风格特征体系构建、国内主流大语言模型翻译体识别效果探究展开，兼具重要理论与实践意义。理论上，从词汇、句法、衔接三维度系统提炼本研究语料中汉译英翻译体普适性核心特征，完善翻译体特征体系化研究，为翻译普遍性假说提供汉译英场景新实证[3]；同时探究豆包、DeepSeek 等模型的识别表现与能力影响因素，填补国内大语言模型在翻译体识别应用研究的空白，推动翻译学与自然语言处理的跨学科融合。实践中，提炼的汉译英翻译体特征体系可为机器翻译模型优化、翻译学语料库纯净文本筛选提供依据，也能为翻译实践与质量评估提供量化标准；而对模型识别能力的分析，能为国内大语言模型翻译相关任务的优化升级明确方向，研究成果亦可应用于跨文化传播、学术写作、商务文案等场景，有效检测翻译体、提升文本表达的规范性与传播效果，规避相关应用问题。

2. 研究方法

2.1. 语料来源

本研究以权威双语新闻语料为研究对象，中译英语料来源于 https://www.china.org.cn/ 2022 年 1 月—2023 年 12 月发布的汉语–英语新闻翻译文本(均为人工翻译)，英语原生语料来源于同期 BBC News 的国际新闻板块。语料筛选遵循以下标准：一是主题聚焦时政、文旅、科技三大领域，确保主题分布均衡；二是剔除包含特殊格式(如表格、列表)。经筛选后最终形成 106 条有效语料，其中英语原生文本(标注为 0)与汉译英翻译文本(标注为 1)各 53 条，样本比例 1:1 以避免类别偏倚。

2.2. 风格指标定义

本研究选取 4 类易计算、强区分度的风格指标，从词汇–句法–衔接三维度系统分析翻译体核心特征，指标选取参考秦洪武、王克非、Liu & Afzaal 等经典研究，兼顾特征辨识度与可操作性[6] [7]。

2.2.1. 词汇丰富度(TTR)

指文本中不同词汇数量与总词数的比值，为避免文本长度对结果的干扰，采用标准化类符形符比(STTR)计算，即每 1000 词为一个区间计算类符/形符比后取均值[4]。该指标用于衡量词汇多样性，是翻译体“词汇僵化”特征的核心量化指标，与已有研究中单一 TTR 计算相比，STTR 更适用于多长度文本的跨样本对比。

2.2.2. 平均句长

指文本总词数与句子数量的比值(词/句)，反映句法复杂程度[6]。相较于仅关注单句长度的传统指标，该指标能更全面体现文本整体句法结构特征，契合汉译英译者拆分源语长句的常见策略。

2.2.3. 虚词占比

参考秦洪武、王克非的分类标准，虚词界定为无实质语义、主要起语法衔接作用的词汇，核心类别包括：① 介词(如 in、with)；② 连词(如 and、but)；③ 副词(如 very、often)；④ 助词(如 be 动词的辅助形式) [7]。计算公式为“虚词占比 = 单句虚词数量/单句总词数”，该指标可有效反映文本形合程度，契合汉英语言“意合–形合”转换的核心差异。

2.2.4. 标点密度

指文本中标点符号数量与总词数的比值，计算公式为“标点密度 = 文本总标点数量/文本总词数” [12]。该指标体现文本断句习惯与逻辑衔接方式，汉译英文本常因虚词使用不足而通过标点补偿逻辑关系，是区别于原生文本的独特衔接特征指标，弥补了传统研究侧重词汇、句法而忽视衔接细节的不足。

3. 研究与讨论

3.1. 汉译英翻译体核心特征及机制分析

从词汇丰富度(STTR)、平均句长、虚词占比与标点密度四个维度的统计结果来看，英语原生文本(label = 0)与汉译英翻译文本(label = 1)呈现出显著差异，既印证了 Baker 翻译普遍性假说中的“简化”“显化”原则，也深度反映了汉英语言类型差异对翻译体特征的塑造作用[3]。

在词汇层面，见图 1，英语原生文本的平均标准化词汇丰富度(STTR)为 0.6285 (SD = 0.042)，显著高于汉译英翻译文本的 0.4812 (SD = 0.051) ($t = 11.36$, $p < 0.001$) [4]。这一结果验证了翻译体“词汇僵化”的核心特征，其深层机制与翻译普遍性中的“简化”原则高度契合——译者为降低跨语言转换难度，倾向于选择高频基础词汇，同时受中文源语词汇体系的制约，缺乏原生文本中常见的同义替换与灵活搭配，导致词汇多样性显著不足[13] [14]。

在句法层面，见图 2，英语原生文本的平均句长为 28.32 词/句(SD = 3.15)，略长于中译英翻译文本的 24.92 词/句(SD = 2.87) ($t = 4.72$, $p < 0.001$) [6]，这与传统“翻译体句长冗余”的认知形成反差。核心成因在于汉英语言句法差异：中文多流水句、意合衔接，而英文重主谓结构、形合逻辑，译者为适配英文句法规范，会主动拆分中文长句，形成“短而精”的句法特征。

在衔接维度，二者呈现“低虚词、高标点”的汉译英专属特征：见图 3，英语原生文本的平均虚词占比为 0.4163 (SD = 0.038)，显著高于翻译文本的 0.3338 (SD = 0.043) ($t = 10.25$, $p < 0.001$)；见图 4，翻译文本的平均标点密度为 0.1725 (SD = 0.021)，显著高于原生文本的 0.1313 (SD = 0.018) ($t = 9.68$, $p < 0.001$)

[12]。这一特征的形成机制可从两方面解释：其一，中文意合特征使译者在转换过程中易忽视英文形合所需的虚词(如介词、连词)，导致虚词使用不足[15]；其二，为补偿逻辑衔接的缺失，译者通过增加逗号、分号等标点划分意群，形成“标点代偿虚词”的衔接策略，这既是翻译普遍性中“显化”原则的体现，也是汉英语言类型差异的必然结果[3]。

综合来看，经随机森林算法特征重要性排序验证，词汇丰富度权重占比 38.7%，为区分两类文本的核心指标，而平均句长、虚词占比与标点密度的差异则共同揭示了汉译英翻译过程中“适配性简化”“逻辑显化”的核心规律，印证了 translationese 作为“第三语码”的独立性特征[14] [16] [17]。

平均词汇丰富度（TTR）对比：英语原生文本vs汉译英翻译文本

文本类型	平均标准化词汇丰富度（STTR）
英语原生文本	0.6285（SD=0.042）
汉译英翻译文本	0.4812（SD=0.051）
t=11.36，p<0.001	

Figure 1. Comparison of Average Lexical Richness (TTR): Native English texts vs Chinese-to-English translated texts
图 1. 平均词汇丰富度(TTR)对比：英语原生文本 vs 汉译英翻译文本

平均句长对比：英语原生文本 vs 汉译英翻译文本

英语原生文本	汉译英翻译文本
平均句长：28.32词/句（SD=3.1.5）	平均句长：24.92词/句（SD=2.87）
t=4.72，p<0.001	

Figure 2. Comparison of average sentence length: Native English texts vs Chinese-to-English translated texts
图 2. 平均句长对比：英语原生文本 vs 汉译英翻译文本

虚词占比对比：英语原生文本vs汉译英翻译文本

文本类型	平均虚词占比（标准差）
英语原生文本	0.4163（SD=0.038）
汉译英翻译文本	0.3338（SD=0.043）
t=10.25，p<0.001	

Figure 3. Comparison of function word ratio: Native English texts vs Chinese-to-English translated texts
图 3. 虚词占比对比：英语原生文本 vs 汉译英翻译文本

文本类型	平均标点密度	标准差（SD）
英语原生文本	0.1313	0.018
汉译英翻译文本	0.1725	0.021
t=9.68, p<0.001		

Figure 4. Comparison of punctuation density: Native English texts vs Chinese-to-English translated texts
图 4. 标点密度对比：英语原生文本 vs 汉译英翻译文本

3.2. 大语言模型识别效果及差异归因

以 106 条语料为测试集，人工标注结果为金标准，对比豆包与 DeepSeek 的二分类识别性能，结果显示：DeepSeek 整体正确率达 97.17% (精确率 0.96、召回率 0.98、F1 值 0.97)，豆包整体正确率为 83.02% (精确率 0.81、召回率 0.85、F1 值 0.83) [10] [11]。与同类研究相比，本研究中 DeepSeek 的识别性能优于 Lembersky 等提出的混合模型(正确率 89.7%)，与 Zhang 基于注意力机制的模型(正确率 96.3%)接近，而豆包表现略低于上述模型，核心差异源于对翻译体特征捕捉的完整性与场景适配性。

3.2.1. 模型错误案例的具体分析

DeepSeek 的 3 条错误均集中于 80 词以内超短文本(该阈值参考 Liu & Afzaal 的短文本界定标准，且本研究语料中 80 词以下文本占比仅 7.5%，特征密度显著低于中长文本) [6]，以“Chinese intangible cultural heritage techniques are widely praised overseas”(68 词，翻译文本)为例，其 STTR 为 0.572 (接近原生文本下限 0.512)，虚词占比 0.38 (高于翻译文本均值 0.3338)，标点密度 0.14 (低于翻译文本均值 0.1725)，因文本长度限制，核心量化特征密度不足(每 10 词仅含 0.7 个区分特征，中等长度文本为 1.5 个)，导致模型无法通过“词汇–句法–衔接”特征组合完成判断，属于“特征被动缺失型错误”。

豆包的 18 条错误中，7 条为原生文本误判，典型案例为专业领域文本“Quantum entanglement enables secure communication across long distances”(72 词，原生科技文本)：其专业术语密度达 15% (专业术语界定参考《牛津高阶英汉双解词典》(第 10 版)中的科技类术语列表，计算方式为“专业术语数量/文本总词数”) [13]，STTR 为 0.586 (处于原生文本区间 0.5123~0.7319)，虚词占比 0.42 (符合原生文本均值 0.4163)，但豆包将“术语密集”简单等同于翻译体词汇僵化，忽略核心量化特征的匹配度，属于“特征认知偏差型错误”；另有 11 条为翻译文本漏判，如深度润色文本“Environmental protection is the common responsibility of all mankind”(65 词)，虽残留“共同责任 common responsibility”的搭配偏差，且 STTR = 0.532、虚词占比 = 0.34，均符合翻译体特征，但豆包未整合表层搭配与深层量化特征，导致漏判。

3.2.2. 模型差异的核心根源

DeepSeek 的高正确率源于其量化特征和隐性特征的双重捕捉能力：既能精准识别词汇丰富度、虚词占比等显性量化指标，又能捕捉语义搭配偏差、逻辑衔接模式等隐性特征，仅在短文本特征密度不足时出现局部偏差，可通过补充短文本特征增强规则快速优化[10]。

豆包的表现偏差本质是系统性问题：一是特征识别维度单一，过度依赖表层特征(如术语密度、文本长度)，忽视 STTR、虚词占比等核心量化指标的协同作用；二是领域化特征库缺失，未针对科技、法律等领域构建专属量化特征区间(如法律翻译文本的虚词占比更低、标点密度更高)，无法区分“领域原生特征”与“领域翻译体特征” [11]。综上，DeepSeek 以 97.17%的高正确率成为更可靠的工具，核心是其能精准捕捉“显性 + 隐性”翻译体特征，且能整合词汇丰富度、平均句长、虚词占比与标点密度等量化特征，仅在翻译体特征(含量化特征)密度不足的短文本中出现局部偏差，可通过补充短文本量化特征增强规则快速优化；而豆包 83.02%的正确率，本质是翻译体特征识别维度单一、领域适配不足的系统性问题，需从底层重构判断逻辑，明确词汇丰富度(TTR)、平均句长、虚词占比与标点密度的两类文本区分标准，补充科技、法律、医疗等领域的专属量化特征库，才能提升对翻译体特征的综合捕捉能力，进而改善文本类型分辨的准确性。

4. 结论

本研究以 106 条汉译英翻译文本与英语原生文本为研究语料，从词汇、句法、衔接维度选取词汇丰富度(TTR)、平均句长、虚词占比与标点密度四类量化指标，系统探究汉译英翻译体的核心风格特征，并

评估豆包、DeepSeek 两款国内主流大语言模型对两类文本的分辨质量[3] [7] [12]。研究发现，汉译英翻译体呈现出鲜明的量化特征：词汇丰富度显著偏低，经特征重要性排序验证为核心区分指标，体现出翻译体词汇僵化的特点；平均句长略短于英语原生文本，反映出译者适配英文句法的拆分策略；同时存在“低虚词、高标点”的专属衔接模式，是汉译英过程中意合向形合转换的典型体现，这些特征既印证了翻译普遍性假说，也揭示了汉英语言转换的独特规律。在模型分辨质量评估中，DeepSeek 整体正确率达 97.17%，豆包为 83.02%，二者的核心差距源于对翻译体特征捕捉的完整性与精准性不同[10] [11]。DeepSeek 仅在 80 词以内超短文本(参考 Liu & Afzaal 界定标准)中出现局部错误，此类错误系文本特征密度不足导致的被动偏差，其核心能精准整合量化特征与隐性翻译体特征完成判断；而豆包的错误呈多场景分散分布，存在原生文本误判与翻译文本漏判问题，本质是特征识别维度单一、领域化特征库缺失的系统性偏差。

综上，本研究明确了汉译英翻译体的核心量化风格特征体系，也厘清了两款大语言模型在翻译体识别中的能力差异与问题根源，为翻译体特征的量化研究提供了实证依据，同时为国内大语言模型在翻译体识别相关任务中的优化升级指明了具体方向，也为翻译实践、语料库建设等场景提供了可参考的量化标准与工具选型依据。后续研究可进一步拓展语料类型，纳入文学、法律等专业领域文本，同时探究模型参数调整与特征融合策略对识别性能的提升作用，进一步完善翻译体识别的技术体系。

参考文献

[1] Nida, E.A. and Taber, C.R. (1969) The Theory and Practice of Translation. Brill Publishers, 12-15.

[2] Shuttleworth, M. and Cowie, M. (1997) Dictionary of Translation Studies. St. Jerome Publishing, 89-90.

[3] Baker, M. (1993) Corpus Linguistics and Translation Studies: Implications and Applications. *Target*, **5**, 223-243.

[4] Laviosa, S. (1998) The Corpus-Based Approach: A New Paradigm in Translation Studies. *Journal des traducteurs*, **43**, 474-479. https://doi.org/10.7202/003424ar

[5] 王子瑞, 李红满. 译者文体印记与翻译体特征互动研究[J]. 外语教学, 2023, 44(5): 98-104.

[6] Liu, K.L. and Afzaal, M. (2021) Syntactic Complexity in Translated and Non-Translated Texts: A Corpus-Based Study of Simplification. *PLOS ONE*, **16**, e0253454. https://doi.org/10.1371/journal.pone.0253454

[7] 秦洪武, 王克非. 基于对应语料库的英译汉语言特征分析[J]. 外语教学与研究, 2009, 41(2): 131-136.

[8] Wang, L. and Jiang, Y. (2024) Do Translation Universals Exist at the Syntactic-Semantic Level? A Study Using Semantic Role Labeling and Textual Entailment Analysis of English-Chinese Translations. *Humanities and Social Sciences Communications*, **11**, Article No. 848. https://doi.org/10.1057/s41599-024-03317-6

[9] Church, K., Li, B., Vickers, P., Dudy, S. and Yue, R. (2025) Emerging Trends: Translationese. *Natural Language Processing*, **31**, 965-981. https://doi.org/10.1017/nlp.2025.1

[10] Lembersky, L., Goldberg, Y. and Levy, O. (2023) Identifying Translationese with Neural Models. *Computational Linguistics*, **49**, 457-492.

[11] Zhang, T. (2022) Deep Learning Classification Model for English Translation Styles Introducing Attention Mechanism. *Mathematical Problems in Engineering*, **2022**, 1-10. https://doi.org/10.1155/2022/6798505

[12] 杨晓琳, 李德超. 语料库翻译研究背景下的“translationese”与“翻译共性”刍议[J]. 山东外语教学, 2024, 45(3): 108-119.

[13] 李德超, 王克非. 汉英同传中词汇模式的语料库考察[J]. 现代外语, 2012, 35(4): 409-415.

[14] 庞双子, 王克非. 基于历时语料库的文学翻译文本和原创文本语体特征演变研究[J]. 外国语, 2023, 46(6): 78-88.

[15] 周彦君. 英汉翻译中的翻译体研究[J]. 河北理工大学学报(社会科学版), 2009, 9(3): 172-174.

[16] Frawley, W.J. (1984) Translation and Language: Linguistic Theories of Translation. Garland Publishing, 210-215.

[17] 柴秀娟. Translationese 及相关概念探析[J]. 当代外语研究, 2012(3): 104-108.