主流机器翻译平台中国特色词汇对比研究
—以中国关键词为例

姚艺恒
昆明学院，云南 昆明 650214
DOI: 10.61369/RTED.2025280036

摘　　要： 自人工智能生成内容 (Artificial Intelligence Generated Content，AIGC) 出现以来，机器翻译形成专注翻译服务平台和人工智能平台共存的局面，且二者都能完成翻译任务。本文在横向对比了由 13 个机器翻译平台生成的中国特色词汇的英文译文后发现，具备 AIGC 功能的平台在翻译质量上占优势，但某些词条译文仍然不准确或不规范，并提出机器翻译平台除了要与大语言模型 (Large Language Model，LLM) 深度融合以外，还必须规范统一译本并参考和完善语料库，以提高翻译质量。
关 键 词： AIGC；机器翻译；特色词汇；翻译质量

Comparative Study of Chinese Characteristic Vocabulary on Mainstream Machine Translation Platforms - Taking Chinese Keywords as an Example

Yao Yiheng
Kunming University, Kunming, Yunnan 650214

Abstract: Since the emergence of Artificial Intelligence Generated Content (AIGC), machine translation has evolved into a landscape where dedicated translation service platforms coexist with artificial intelligence platforms, both capable of performing translation tasks. This study conducts a horizontal comparison of English translations of Chinese characteristic terms generated by 13 machine translation platforms. The findings reveal that platforms with AIGC capabilities have an advantage in translation quality. However, some translated entries remain inaccurate or non-standard. The paper proposes that machine translation platforms, in addition to deep integration with Large Language Models (LLM), must also standardize and unify translations and refine their corpora to improve translation quality.
Keywords: AIGC; machine translation; characteristic terms; translation quality

一、研究背景及意义

本文所指的机器翻译（下称 MT）包括人工智能平台和专注于翻译服务的平台，是由上世纪 50 年代统计机器翻译 (Statistical Machine Translation，SMT)、计算机辅助翻译 (CAT)、神经机器翻译 (Neural Machine Translation，NMT)，在技术与网络资源的不断融合过程中逐步演化而来。目前的 MT 开始突破以往难以企及的自然语言处理 (Natural Language Processing，NLP) 中语境、逻辑、上下文的藩篱，利用深度学习技术 (Deep Learning)、神经网络 (Neural Network) 和大语言模型 (Large Language Models，LLMs) 来处理文本。自 2022 年 11 月 ChatGPT(Chat Generative Pre-trained Transformer) 发布，神经机器翻译又向前迈进了一步。现如今已经有很多主流人工智能平台能提供 MT 服务，　如：ChatGPT，Deep Seek，Google Gemini 等。也不乏各类融入了人工智能技术、专注于翻译服务的平台，如 Lionbridge、DeepL、金山词霸等，它们并不具备人工智能生成内容 (Artificial Intelligence Generated Content，AIGC) 功能。

MT 涉及“讲好中国故事”、“建构当代中国话语体系”之类的研究方兴未艾。这些领域集中体现一个国家软实力和巧实力，决定其主流意识形态的地位和国际话语权的强弱 [1]。有了话语权便有了对自己文化和意识形态的解释权。“龙”祥瑞、正面的形象的树立离不开正确的文化和意识形态的传播 [2],[3]。“中国关键词”网站（http://keywords.china.org.cn/index.htm）是向国际社会解读、阐释当代中国发展理念、内外政策、思想文化核心话语的窗口和平台，其中的中国特色词汇蕴含有丰富的中国哲学文化思想 [4]。对特色词汇进行多平台翻译横向对比研究，可大致筛选出高效、安全、准确的平台，并通过分析造成译文优劣的原因促进人智协同和人工智能译后编辑 AIPE(Artificial Intelligence Post-editing)，为自然语言处理和大语言模型的发展提供参考。

二、研究现状

中国外文局 CATTI 项目管理中心学术研究部发布的《2022 国内主流 AI 翻译机实测报告》发现：科大讯飞翻译机翻译质量和翻译速度占优，其 4.0 版本也已经与大语言模型结合 [5]。另外也有一些关于 MT 英汉互译、错误类型的研究，如蔡欣洁等（2021）

发现所测试四个 MT 平台在汉译英转换上，存在词和句层面翻译错误、漏译、大小写格式错误等共性化问题[6]。主要问题有用词不当、术语错误和搭配不当[7]。还有学者对 AIGC 平台语言处理功能有研究成果，如王华树等（2024）在译者主体性研究中发现：类似 ChatGPT 这样的系统的序列任务执行能力在翻译中尤为重要。通过反馈循环和持续学习，LLMs 能不断优化性能，提升翻译的准确性和流畅性，并迅速适应新的语言变化和领域术语[8]。也有一些与本研究近似的成果，如凌颖（2025）发现人工智能翻译工具在中国特色话语翻译中表现出一定的潜力，但在面对专业术语、政策表达及文化背景的准确传达方面，仍存在局限性[9]。段田园（2025）则发现，DeepL 译文准确性整体略优，但在翻译中国特色词汇方面存在不足；豆包表现较为稳定；ChatGPT 在处理语义和词汇方面表现出色，准确性相对突出[10]。

除上述研究以外，全球还有 ACL、EMNLP、NAACL（北美分会）、COLING 之类的 NLP 领域的论坛，以及世界机器翻译大会（WMT）和全国机器翻译大会 (CCMT)，它们重点讨论长文本处理、模型对齐、公平性、多模态推理、多种语言对翻译对比、MT 评价、数据训练等前沿领域。2022年后世界机器翻译大会的主要研究成果里也并不只局限在汉译英。直至2023年才首次使用大语言模型 Lan-BridgeMT 和 ChatGPT4作为翻译系统参与共享任务，但汉译英也只局限在几个领域，几乎不涉及中国特色词汇，可供参考内容有限。

鉴于上述问题，本文拟采用人工与自动评价相结合的方式，比较各个 MT 平台的译文，探讨以下问题：1). 哪种类型的 MT 更为准确？ 2). 哪种类型的原文更容易有相对准确或不准确的译文？3). 提高准确性可能的办法有哪些？

三、研究方法

（一）研究语料及平台

我们选取13个主流且免费的 MT 平台（截至2025年3月30日），包括 ChatGPT，Deep Seek R1，Google Gemini，豆包，文心一言 (2个版本)，通义千问2.5，DeepL(2种模式)，有道翻译 (2种模式)、腾讯元宝、语言桥 Dtranx。这里面既包含了专门提供翻译服务的平台也包括有多种功能的人工智能平台。我们在1700余条词条中（截至2025年3月），选取了约四分之一即442个中国特色词汇在上述平台中翻译。选词考虑到了词条的多样性、使用频率和时代性，包含成语、古文、缩略语、扶贫、经济社会发展等多个类别，且翻译过程中除了基本的翻译指令，不输入其他提示词。研究的语料没有选择句子或段落，主要是考虑到中国特色词汇在通用语境中较少出现歧义，译文生成受语境的影响较小。

（二）质量评估工具及方法

人工评测无法高效地处理大量信息，评测人员主观上理解不同，水平不同，对质量评判也很难统一，却是最为稳妥的办法。依靠类似 BLEU 模型或度量语义相似性的 BLEURT、BERTScore、COMET 等，对翻译质量进行分析，会面临很多掣肘，比如含义对等可能文化、内涵不对等或缺失。现在学术界对 MT 的质量评估主要以人工评估和机器评估互为补充[11]。作为文化载体，语言的随意性、创造性、多样性、以及精神、情感、文化内涵的细微之处也是极难量化的东西。故本文采用机器自动评价与人工评价相结合的办法评价译文质量。

1. 机器自动评价

我们用 BLEURT、BERTScore、COMET 三种机器自动评价模式为平台的译文赋分，参考译文选用中国关键词网站所公布的官方译文。具体计算逻辑如下：BLEURT 近似人工评分，分值接近1，意味着译文与参考译文在语义、用词和流畅性上高度一致；BERTScore 关注语义对齐，F1值（精确度和召回率的调和平均值）高于0.9意味着译文与参考译文在词或语义层面高度相似；COMET 分值接近1通常意味着译文忠实保留了原文含义，与人类对翻译质量的判断一致。总体上来说，三种机器自动评价模式的分值越高意味着译文越准确。

2. 人工评价

人工评价我们以国家质量监督检验检疫总局、中国国家标准化管理委员于2005年发布的翻译服务译文质量要求和中国质量协会、华为技术有限公司于2022年联合提出的《机器翻译服务质量评价规范—中英双向》为参考。为避免个人对平台的偏见，并没有把平台名字列出，只是将平台编号。再横向对每个平台生成的442条译文逐条和参考译文进行比较。除团队成员（包括教授和博士各一名）外，还邀请获得省级翻译比赛一等奖的学生来参与。参考译文除中国关键词网站以外，还包括《中国时政话语翻译基本规范·英文》以及其他类似资源。考虑到本文的目的是找出 MT 发展至今尚存的最大的问题并探讨解决之道，所以我们并没有按照翻译多维质量标准 Multidimensional Quality Metrics (MQM) 来标注译文打分，只要求参与人员将核心语义差错、专用名词术语错译或完全不准确的、无法达到交际效果的译文进行错误标记。反之，达到或基本达到参考译文标准的则标记为准确。标记为准确的词条总数与442个词条的比值就是相对准确率。

四、结果及讨论

（一）机器自动评分和人工评测结果

粗略来看，自动评分结果（表1）在一定程度上反映了各个平台的翻译质量。在 BERTScore 中，所有平台表现都较为一致，分值差距小，唯有 DeepL 单独翻译的平均分略低于其他平台，这可能与 MT 不断进化更新和 BERTScore 的得分机制有关。COMET 被誉为“最佳性能指标”(WMT2020)，各平台分值差距较小，但较低分数 (0.76以下) 出现在 Deep Seek R1、DeepL 的两种模式和有道翻译的神经网络模式中，且 DeepL 和有道翻译的神经网络模式的标准差也高于其他平台，显示出其翻译质量的不稳定性。就 BLEURT Score 而言，0.1以下的分数出现在 DeepL 和有道翻译的神经网络模式，且 DeepL 单独翻译的标准差 (0.787) 明显高于其他平台。就总分而言，三种评价模式下，DeepL 的两种模式、Deep Seek R1和有道翻译的神经网络模式都未能有较高评分。虽 DeepL 自诩“最准确的翻译”，但与自动评测的结果相悖。有道

翻译(两种模式)和 Deep Seek R1在三种评价模式下的表现并没有超越其他平台。

就人工评测结果(表2)来看，四名评测人员标记出的较低相对准确率与机器自动评价分值有较高的一致性，即 DeepL 的两种模式和有道翻译的神经网络模式相对准确率最低，且略低于 Dtranx 文档翻译。这三个平台的标准差也略高于其他平台(大于0.02)。

从上述两个表不难看出，DeepL 的两种模式和有道翻译的神经网络模式是得分最低的平台，说明在某些词汇上有不小的短板，Dtranx 文档翻译次之。Deep Seek R1的人工评测分值相对较高但自动评价分值相对较低，其原因有待进一步探讨。

表1. 平台所有译文得分平均值

	ChatGPT 4omini	Deep seek R1	Gemini 2.0 flash	文心一言 3.5	文心一言 4.0	豆包	DeepL单独翻译	DeepL文档翻译	通义千问	有道翻译（神经网络翻译YNMT）	有道翻译（AI模式）	腾讯元宝	Dtranx 文档翻译
BERT Score 平均值（bert-base-multilingual-cased）	0.917	0.904	0.919	0.919	0.915	0.924	0.893	0.911	0.918	0.909	0.921	0.916	0.914
BERT Score标准差	0.148	0.137	0.049	0.155	0.147	0.156	0.148	0.161	0.149	0.156	0.153	0.157	0.155
COMET 平均值（wmt22-comet-da）	0.780	0.741	0.785	0.782	0.781	0.781	0.689	0.756	0.785	0.754	0.785	0.786	0.774
COMET 标准差	0.147	0.139	0.142	0.147	0.140	0.148	0.169	0.157	0.148	0.169	0.150	0.152	0.155
BLEURT 平均值（bleurt-tiny-128）	0.171	0.165	0.179	0.168	0.208	0.207	-0.216	0.082	0.194	0.089	0.221	0.191	0.125
BLEURT标准差	0.600	0.648	0.629	0.607	0.604	0.612	0.787	0.656	0.614	0.664	0.615	0.613	0.668

表2. 人工评测相对准确率

	ChatGPT 4omini	Deep seek R1	Gemini 2.0 flash	文心一言3.5	文心一言4.0	豆包	DeepL单独翻译	DeepL文档翻译	通义千问	有道翻译（神经网络翻译YNMT）	有道翻译（AI模式）	腾讯元宝	Dtranx 文档翻译
人工相对准确率1	0.943	0.957	0.975	0.952	0.968	0.964	0.792	0.810	0.952	0.796	0.948	0.975	0.873
人工相对准确率2	0.941	0.964	0.975	0.950	0.968	0.959	0.771	0.796	0.955	0.787	0.943	0.982	0.862
人工相对准确率3	0.921	0.930	0.962	0.928	0.957	0.952	0.731	0.753	0.930	0.744	0.923	0.937	0.894
人工相对准确率4	0.932	0.950	0.968	0.941	0.964	0.962	0.771	0.799	0.952	0.776	0.943	0.973	0.846
标准差	0.010	0.015	0.006	0.011	0.005	0.005	0.025	0.024	0.011	0.023	0.011	0.020	0.013

（二）平台译文比较

1. 几个平台纵向比较

为了进一步验证得分较低的上述几个平台是否确实产生较多的误译，我们将分析这几个平台的部分典型译文。

有道翻译的神经网络模式：《编户齐民》The editors and the people are in harmony、和实生物 eubiotic、格物究理 Look at things。

DeepL 单独翻译：《齐民要术》The Essentials of Qi Min、《天工开物》The Work of Heaven、老子 "I, your father" (in anger, or out of contempt)、天下大同 the whole country under heaven、龙凤呈祥 brilliant calligraphy or painting of the dragon and phoenix (idiom); fig. brilliant calligraphy。DeepL 作为国外研发的平台，归化倾向比较明显，译文靠近读者，如“协和万邦的国际观”译成：Concordia Viva's international perspective，Concordia Viva 是拉丁语，意为“活力和谐”或“生命的和谐”，与原文相去甚远；《诗经》译成：The Book of Psalms，Psalms 是圣经的“诗篇”，用在《诗经》里实为不妥；授时历译成：The Gregorian calendar，如按此译法，读者只会认为授时历就是由教宗格列高利十三世 (Pope Gregory XIII) 改革历法而来，以上种种均违背了“异化为主、归化为辅”的原则。

DeepL 文档翻译：盘古开天 Pangu opens the sky、靡不有初，鲜克有终 If there is no beginning, there is no end 等。

Dtranx 文档翻译：浩然之气 Noble Qi、扶真贫、真扶贫 Targeted poverty alleviation 等。

产生“离谱”的译文是因为这些应用或平台深度学习和神经网络发展滞后，尚处于迭代阶段，如无法识别文化专有项 (Culture-Specific Items，CSI)，(平行)语料库不充实[12]，如缺乏中国传统文化数据库。

再进一步横向比较上述词条的译文后发现，其他平台虽然时而出现不准确的现象如：“龙凤呈祥”译为：“Dragon and Phoenix Bringing Prosperity”，严重误译则相对较少，如：“和实生物”都译出了重点词“Harmony”，“伊儒会通”译出了重点词“Islam-Confucianism”，“上善若水”译出了“virtue/goodness”。不难发现，这几个得分较低、无 AIGC 功能的平台虽然以提供翻译服务为主，声称有 AI 技术，或融入了神经网络机制，但在翻译特色词汇方面并不具优势。

2. 平台高分译文横向比较

对于平台在哪些词汇上有准确译文，我们先对三个自动评测模式下所有平台译文得分进行排序，发现得分前10%(约40条)的词条译文都在以下方面体现出了优势：1. 中国传统历史文化，如：道教、针灸、天干地支；经典著作，如：《道德经》、《论语》；人名，如：孔子、王阳明；传统节日，如：清明节、中秋节；标志性地理名称，如：黄河、敦煌莫高窟；某些历史事件，如：郑和下西洋。2. 不易造成误解的、国际通用的词汇，如：石狮、实现共同富裕、包容性增长等。以上的类别，包括几个得分较低平台的译文都与官方译文一致(偶有冠词和语序的差别)。这也说明，有足够语料库、数据库和技术的支持，是可以将某些 CSI 或特色词汇准确译出的。

3. 低分译文横向比较

我们再将三个自动评测模式下13个平台得分后10% 出现10次及以上的词条列出，共55条(表3)，以便分析什么样的词条会得到低分。

表3得分后10% 出现10次及以上的词条与低分出现次数

词条	低分出现次数	词条	低分出现次数	词条	低分出现次数
《本草纲目》	37	美人之美，美美与共	20	增强中华文明传播力影响力	13
《伤寒论》	37	盘古开天	19	行省制	13
浑天仪	36	太极图	19	读万卷书，行万里路	13
四合院	35	“回头看”	19	《史记》	13
“六个精准”	34	愚公移山	19	泰山	12
三大攻坚战	32	厚积薄发	19	和实生物	12
“五个一批”	32	《营造法式》	18	摘帽不摘政策	12
“七个强化”	30	旗袍	18	孝老爱亲	12
建档立卡	30	合	18	补短板	11
《资治通鉴》	27	投桃报李	16	君子一言，驷马难追	11
和谐共生	26	从善如登，从恶如崩	16	驻村帮扶	11
中央统筹、省负总责、市县抓落实	26	郡县制	15	社会保障兜底	11

《黄帝内经》	25	和合天下	15	“绣花”功夫	11
“两手抓、两手要硬”	24	青铜器	14	和合五教	11
和	24	言必信，行必果	14	城乡建设用地增减挂钩	10
内生动力	22	“两线合一”	14	见微知著	10
《四库全书》	21	苟日新，日日新，又日新	13		
和合文化	20	修身处世	13		

例1：中国文化特有事物。《伤寒论》、浑天仪、四合院、太极图、泰山、旗袍等，出现了译本不统一的问题。以上词条被人工标注为错误的较少，但自动评分较低，主要还是归咎于三种评价指标的算法逻辑是以参考译文为准绳来赋分。”

例2：特有缩写简略词汇。这些词条都是从我国政策文本中提取出来的缩略语，也正是目前MT的硬伤之一：各平台译文流于原文表层符号，无法译出简化却有较多文化历史背景的表达或四字短语，未能触及其语用意义或内涵。如平台中几乎无一能将“三大攻坚战”、“五个一批”、“两线合一”、“两手抓、两手都要硬”、“‘绣花’功夫”正确译出。“和合五教”(古代也称“五典”)，部分平台曲解了其含义，以致将“教”译为：“religion”。若译为“teachings教义，学说”，虽与“Ethical Codes伦理准则”有重叠之处，仍与原文含义有差别。“两手抓、两手都要硬”，指的是“一手抓物质文明，一手抓精神文明”的方针，如官方译文为：A two-pronged approach；但各平台译文却五花八门，如ChatGPT 4omini译文为：“Two Hands Grasp, Both Must Be Strong”；有道翻译(神经网络翻译)译文为：Both hands should be hard；DeepL单独翻译译文为：Grasp both hands and do both；DeepL文档翻译译文为：do both；文心一言3.5译文为："Grasp with both hands, and both hands must be strong"；豆包译文为："Attach equal importance to two aspects of work"；有道翻译(AI模式)译文为"Grasp Both and Let Neither Go"。

ChatGPT 4omini、有道翻译神经网络翻译、DeepL单独翻译、DeepL文档翻译都出现了较大的偏差或错误。其他的平台基本上体现了原文字面含义，但难以体现其内涵。豆包、有道翻译(AI模式)与官方译文不太一致，但没落入按字面硬译的窠臼。至于到底是“grasp with both hands”、“Two Hands Grasp”、还是“Grasp both hands”，对于平台来说无疑是个挑战。“‘绣花’功夫”指的是各项工作都要像绣花一样精准施策。但如果只译成embroidery skill/kung fu/work就与原义相去甚远。“读万卷书，行万里路”指的是要获取知识，积累实践经验，但大多数平台的译文“Read ten thousand books, travel ten thousand miles.”都停留在词汇表面意义。

例3：成语及古代思想。此类表达简略，属文化负载词，存在“和”、“穷”这样一字多义的现象，易导致误译或不准确。这类文本对译文效果的影响显著，文本口语化表达较多，涉及谚语、成语等具有独特文化信息的词汇，语言表达形式较为灵活，难以被机器准确理解语言背后的文化内涵。“穷则变”中，“穷”指极限，但没有平台正确译出“limit”一词。

综上，MT译本不一致导致低分可能是平台和自动评价指标的无心之失，而缩写简略词汇、成语及古代思想比较容易导致在机器自动评价中出现低分。但如果是已经被目标语接受的、统一译法的词条，无论什么平台都能有准确译文。

（三）提高准确率

1.MT平台是否能够通过对话生成准确译文？受王华树等（2024）研究的启发，我们用统一的提示模板[8]，在有AIGC功能的平台对误译出现频率最高的几个词条单独用中英文输入提示词：什么是“六个精准”、“两手抓、两手都要硬”、“天人合一”？如何译成英文？译文表达什么含义，是否正确？，发现这些平台对词条解释正确，并给出了多个译本。“天人合一”没有在低分词条中出现，但这一经典哲学思想内在的价值观很难用一两个词汇表达。“天”既可以是“law”、“nature”，甚至是“heaven”；“合”既可以是“harmony”或者“unity”。只是“heaven”一词只体现“天国之天”的内涵，忽视了当前人与自然互生、相容、和谐的生态文明思想[13]。虽然ChatGPT 4omini给出的译文是：Harmony between heaven and humanity但也在追问下给出了准确解释，见图1：

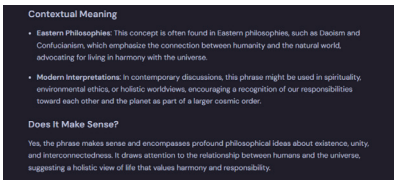

图1.ChatGPT给出“天人合一”的回答

此时，我们再向Deep Seek R1询问它自己生成的译文，见图2。“修身处世cultivate oneself and manage one's affairs”。“处世”和“manage one’s affairs(管理个人事务)”并不能划等号。虽然Deep Seek R1在自动评分中未必占优，但从人工评测来看也不逊色，若无特别提示语，则可能呈现多种译本共存的局面。我们用另外几个有AIGC功能的MT平台针对其他误译率高的词条进行中英文对话，都能给出相应的正确解释和译文。不具备AIGC功能的平台，如DeepL，只要语料库足够充实，也能有体现内涵的译文：如“读万卷书，行万里路”被译为“Gain knowledge from books, gain experience from travel”。

图2.Deep Seek给出“修身处世”的回答

可见，有AIGC的平台在人为干预下可生成准确译文；不具备该功能的平台，则要有充实的数据库或专用语料库。但得注意，AIGC功能极有可能受AI幻觉(AI Hallucination)的影响，生成可

信度存疑的内容[14]，需要验证甄别。

2. 充分参考外部资源和知识库，如平行或专用语料库，或能提高翻译的规范性和准确性。一方面，翻译质量的评价要考虑翻译语言与目的语语言规范的一致化程度，翻译语言特点越符合目的语语言规范（习惯），越可能便于目的语读者理解或接受[15]。MT 平台似乎并没有充分借鉴语料库内容，在生成译文时有规范性问题，这又是 MT 的又一短板。另一方面，（术语和准术语）语料库的语料量不足或功能缺陷，会导致译文质量的下降[6]。“解放和发展社会生产力”较为常见的译文是：Emancipation and development of social productive forces，纵然在网络有此表达，但在外国媒体和语料库罕见。

Emancipate 主要是指“to give someone the political and legal rights that they did not have before，解放，给（某人）政治、法律权利”。它和 productive forces 搭配并不妥当。对英语母语者来说，Emancipation and development of social productive forces 更像是在说：人的解放和生产力的发展。但以“liberate productive forces”为关键词查询语料库得到2个境外媒体的结果：1. Taiwan’s economic development and its relationship to the international environment.，ACAD: Asian Affairs: An American Review，Summer91，Vol. 18 Issue 2，p63，15p；2. The New China -- A Special Report.; Chinese Communism’s Secret Aim，NEWS: New York Times，1992 (19921019))，故不失为一种可取的表述。通过成对地类比不同语料来系统考察两种语言相关主题语言规范的偏离程度，可以避免或减少主观性，提高语言规范程度评价的客观性和系统性[15]。

3.MT 需不断升级优化

就目前语料库而言，训练数据短缺成为全球共性问题，是制约 AI 大模型发展的重要因素[17]。语料库的存在是让译文“有据可考”。提高翻译准确性需要各平台扩充自身数据库、语料库，引入知识图谱，如中文通用百科知识（CN-DBpedia）、Wikidata 等，使特定词汇、术语检索匹配更精确。但单靠增加计算数据量和“思考时间”并不能显著提高翻译质量[18]。如果把语料库看成食材，那平台就是厨师。平台本身的算法、自然语言处理技术、大语言模型数据训练也是生成准确、规范译文的必要条件。通过学习大量的语言规律和语言特征，针对性地改进语言模型，能够显著提升机器翻译的质量[16]。可以针对性地训练成语、谚语、古汉语和中国特色词汇语言模型。DeepL 已经开始与语言模型结合，并在其专业版中加入编辑工具，词汇表，风格切换功能。有道翻译神经网络翻译也已提供多种翻译结果选择功能。这两个翻译服务提供商都已经意识到了自身平台服务发展的瓶颈，开始逐步改进。腾讯研究院推出的新型 AI 模型——DRT-o1 系列采用长思维链 (CoT) 技术，旨在处理比喻和隐喻等修辞手法方面提高文学作品的翻译质量。总之，高质量的译文是数据和平台共同发展的结果。

五、结语及展望

本文大范围地对比了多个平台翻译能力，发现没有 AIGC 功能的平台在翻译某些词条时有很多掣肘之处，而有这项功能的平台总体上译文准确率相对较高，但无论哪种类型的平台都会出现不准确或错误译文。这主要是数据库或语料库相关词条缺失、译文没有完全异化或异化不完整、各种媒介译本“各执一词”之故。有 AIGC 功能的平台少有“知识盲区”，对于某些误译率较高的词条，通过提示词和追问的方式，同样能有高质量译文。既然很多中国古籍如《道德经》、《论语》、《孙子兵法》等都可以正确译出，说明统一规范译本、不断扩充语料库对生成准确、无歧义译文至关重要。

因本研究的对象是特色词汇，所以没有涉及语境对词汇译文的影响，这也是后续研究的着力点。在本研究进行的同时，各种 AI 大模型的迭代更新并未停止，如 ChatGpt 5、Gemini 2.5、文心一言 4.5 Turbo、文心一言 X1.1、Deep Seek 3等，使得横向研究有时效性问题。另外，MT 技术上的一些缺陷如 AI 幻觉、大语言模型黑箱 (LLMs Black Box) 可能对译文有影响，这也要求我们对生成译文谨慎甄别。鉴于以上种种问题，往后研究方法要进一步改良，如限定研究文本范围或对象、单平台纵向研究、AI 与专用语料库融合等。

参考文献

[1] 韩庆祥，陈远章 . 建构当代中国话语体系的核心要义 .《光明日报》，2017年05月16日15版 .

[2] 肖雪 . 专家称应将龙译为 loong：龙与 dragon 有重大差别 [EB/OL][2015-07-16] http://culture.people.com.cn/n/2015/0716/c22219-27316631.html.

[3] 外文出版社 . 中国龙，是 "Dragon" 还是 "Long/Lonng"，这是一个问题！ [EB/OL][2024-03-07]https://www.163.com/dy/article/ISMHAGHN05565V7N.html.

[4] 中国网 . 项目介绍 [EB/OL][2014-11-13]http://keywords.china.org.cn/2014-11/13/content_34036371.html.

[5] 科大讯飞官方账号 .IDC 发布中国 AI 翻译技术评估报告：科大讯飞8项评测全部第一，6项满分领跑行业 [EB/OL][2025-10-13]https://news.qq.com/rain/a/20251013A08BN300.

[6] 蔡欣洁，文炳 . 汉译英机器翻译错误类型统计分析——以外宣文本汉译英为例 [J]. 浙江理工大学学报（社会科学版），2021,46(2):162-169.

[7] 雷鹏飞；张浮凌 . 基于机器翻译软件的外宣文本翻译质量评估研究 [J].《未来与发展》2024.(6).

[8] 王华树，刘世界 . 大语言模型对译者主体性的冲击及化解策略研究 [J]. 外语与翻译，2024,(第4期),13-19，I0001.

[9] 凌颖 . 人工智能中国特色话语汉英翻译质量研究 [J]. 现代语言学 .2025,13(4),575-581.

[10] 段田园 . 人工智能时代机器翻译汉译英质量评测 [J]. 数字技术与应用，2025,43(5):9-11.

[11] 王均松，庄淙茜，魏勇鹏 . 机器翻译质量评估：方法，应用及展望 [J]. 外国语文，2024,40(3):135-144.

[12] 韦佑武，李娜，赵良威 . 机器翻译的译文质量、高频错误类型及解决对策研究：基于机器翻译的发展史 [J]. 现代语言学，2022,10(9):1944-1949.

[13] 杨艳霞，王雨婷，向毓 . 机器翻译质量影响因素研究：来自元分析的证据 [J]. 外语学刊，2025(3):26-32.

[14] 汤一介 . 论 " 天人合一 "[J]. 中国哲学史，2005(2):5-1078.

[15]Waldo Jim，Boussard Soline. GPTs and Hallucination[J].Communications of the ACM，2025，68(1).DOI:10.1145/3703757.

[16] 张威，张蕾 . 中国特色话语对外译介的语言规范及质量评估 [J]. 外语学刊，2025(5):84-90.

[17] 李兴腾，冯锋，黄鹂强 . 突破人工智能大模型的 " 数据瓶颈 "——构建国家级语料库运营平台的思考 [J]. 中国科学院院刊，2025,40(3):522-529.

[18]Zihao Li，Shaoxiong Ji，Jörg Tiedemann，Test-Time Scaling of Reasoning Models for Machine Translation，arXiv - CS - Computation and Language Pub Date：2025-10-07，DOI：arxiv-2510.06471.

