基于类比语料库的西安旅游英译文本语言特征研究

谢晨阳 *

西安翻译学院　陕西西安　710100

摘　要：本文以自建美国原生旅游文本语料库为参考，探索自建西安旅游英译文本语料库的文本特征。研究发现：（1）西安旅游英译文本词汇丰富度低于美国原生旅游文本；（2）西安旅游英译文本中过去时态和被动语态使用频率偏高；（3）西安旅游英译文本能精准呈现西安古都特色，但行文较为生硬，缺乏互动性。本研究结果为优化西安旅游文本的英译提供了参考依据。

关键词：旅游文本；翻译研究；语料库；语言特征

1. 研究背景

2024 年中国持续深化对外开放，过境免签政策对国际旅游市场产生了积极影响。作为历史文化名城，西安也成为国外旅客的热门打卡地。然而，当前国内旅游文本翻译仍存在目标读者意识薄弱、文本功能失衡等问题，这在一定程度上影响了外国游客对西安旅游资源的体验和理解。

2. 研究方法和思路

近年来，语料库方法广泛应用于旅游文本翻译研究。Gandin（2013）探讨了翻译与非翻译旅游文本在话语模式和文体特征上的差异；Hogg 等（2014）探讨了旅游文本翻译应遵循的文化规范。国内亦有研究从语料库数据维度关注旅游文本英译的质量提升（康宁，2012；李德超 & 唐芳，2015）。

然而目前仍鲜有关于西安旅游英译文本的量化研究，故笔者自制中美旅游文本语料库，结合量化与质性分析，揭示两者在词汇选择、句式结构等方面的异同，以期为西安旅游文本英译质量的提升提供实证支持。

2.1 语料来源

美国源语旅游文本语料库（American Native Tourism Corpus, ANTC, 187330 词）由从 VisitTheUSA.com 网站随机爬取的 413 篇文章构成。VisitTheUSA.com 是美国官方旅游指南网站，内容权威且可信。中文旅游文本英译语料库（Chinese Tourism Translation Corpus, CTTC, 48320 词）由《陕西旅游文化丛书》英译本的《陕西人文旅游》和《陕西生态旅游》西安篇构成，该丛书为陕西省社科联资助的科普读物，亦有较高的权威性和系统性。ANTC 作为参考语料库，其词数约为 CTTC 的四倍，这一比例符合类比语料研究的通行做法。

2.2 分析工具

在 Python 3.12.7 环境下，本文采用 Natural Language Toolkit 进行文本清洗、分句、分词、词性标注和句法分析；使用 Pandas 和 NumPy 进行量化数据计算。此外，借助 WordSmith Tool 8.0 基于 ANTC 词表提取 CTTC 的高频关键词。

2.3 研究方法

本文通过 Python 编程工具，对源语文本与翻译文本进行量化对比分析。重点考察标准化类符形符比（STTR）、句 / 词长均值（SLM, WLM）、名词 / 动词 / 形容词长均值（NLM, VLM, ALM）、句 / 词长标准差（SLSD, WLSD）、语态与时态分布等参数，并结合 WordSmith 8.0 生成的关键词表，系统研究西安旅游文本英译的词汇复杂度和句法结构，为数据驱动的翻译质量评估提供实证支持。

3. 文本数据对比与分析

3.1 文本基本特征

词汇丰富性和多样性是衡量文本质量的重要指标，本研究首先对 ANTC 与 CTTC 的词汇特征进行对比。

表 1　ANTC 与 CTTC 的词汇特征

语料库	Token	Type	STTR	SLM	WLM	NLM	VLM	ALM	SLSD	WLSD
ANTC	187330	17425	51.02	22.28	4.90	6.18	5.21	6.41	14.79	2.53
CTTC	48320	6005	42.77	22.98	4.69	6.00	5.17	6.41	13.40	2.46

STTR 是评估词汇丰富性和多样性的关键参数之一。其数值越高，表明词汇丰富度越高，文本表达多样性越强。对比结果表明，CTTC 的 STTR 小于 ANTC，表明源语文本词

汇密度与多样性更优。此现象符合 Laviosa（1998）提出的词汇简化假说：译者倾向于选择高频词项以降低认知负荷，导致译文词汇丰富度衰减。旅游文本往往涉及较多的专有名词和文化负载词，这些词汇的翻译难度较高，为了方便目标语读者理解，译者往往会采用省译，替换等策略简化表达提升译文的可读性。例如中文“亭台楼阁”统一译为“pavilions”，忽略建筑类型差异。

CTTC 的平均句长（22.98）略长于 ANTC（22.28 ），但句长标准差（13.40）低于 ANTC（14.79）。合理的解释是：汉语的意合特征使源文本多用流水句，而英译时译者常借助连接词（如 which 和 that）实现形合转换，导致句长增加但结构趋同。例如，依山傍水，风景秀丽被译为 The resort, which is surrounded by mountains and rivers, boasts magnificent scenery，定语从句显著增加了句长。此外，根据翻译规范理论（Toury, 2012），译者倾向于“规范化”，体现为词汇多样性降低和句子长度增加，例如在汉英翻译过程中添加解释性成分以增强目标文本的可理解性。

在词汇层面，CTTC 的平均词长（4.69 ）低于 ANTC（4.90），名词（6.00 vs. 6.18）和动词（5.17 vs. 5.21）长度均有所缩减。汉语依赖语境和句法关系，英语强调形态变化（Halliday & Matthiessen, 2014），因此，译者可能倾向于采用更通俗、固定的表达方式提高可读性，如用 show 替代 demonstrate。形容词无显著差异（6.41），可能因旅游文本需维持描述性词汇的感染力，如 picturesque 和 magnificent 等长词具有不可替代性。

3.2 词汇特征

本文借助 WordSmith 的 KeyWords 功能，以 ANTC 为参考语料库，采用 BIC(Bayesian Information Criterion) 评分排序，剔除虚词后，获得了共计 30 个 CTTC 关键词，详见表 2 和表 3。

表 2 CTTC 正向关键词（BIC 正值）

Keywords	BIC	Keywords	BIC	Keywords	BIC
dynasty	1,273.69	were	272.94	pagoda	146.6
was	628.59	mausoleum	238.74	tableland	133.89
temple	542.49	imperial	156.05	ancient	122.94
emperor	365.84	built	153.25	wall	112.06
palace	358.15	gate	151.52	relics	102.74

表 3 CTTC 负向关键词表（BIC 负值）

Keywords	BIC	Keywords	BIC	Keywords	BIC
rocks	–9.4	gallery	–9.28	love	–8.57
climbing	–9.4	games	–9.13	nation	–8.57
well	–9.38	takes	–9.12	travel	–8.57
peak	–9.38	species	–9.05	look	–8.45
boating	–9.36	markets	–8.59	works	–8.33

BIC 值是衡量文本词汇显著性差异的参数。BIC 值综合考虑了词频、文本离散度及语料规模影响，有着较高的可靠性。BIC 值为正表示该词在目标语料库中的使用频率远高于参考语料库，反之则表示使用频率低于参考。关键词的显著性不仅反映出文本的语言使用习惯，还能揭示其文体风格。

从名词的角度来看，正向词表中，历史文化相关词汇占据主导，如 dynasty、temple、emperor、palace、mausoleum、relics 等。这种词汇分布特征与西安深厚的历史文化背景相契合，体现了翻译文本在跨文化传播中的信息选择倾向。相比之下，负向词表包括了 species、rocks、climbing、boating、gallery 等，说明 ANTC 更倾向于描述自然景观、户外活动和艺术展览等主题，强调现代旅游体验和互动性。这一差异可能源于西安旅游文本在国际推广中的策略选择，即通过强化文化遗产叙述以塑造城市形象并吸引具有文化兴趣的游客。

形容词方面，正向词表中的 imperial, ancient, ecological，表明译文倾向于使用具有历史或自然特征的描述性词汇，以增强文本的文化厚重感。然而，过多的描述性词汇可能会使文本显得过于正式，而缺少轻松、互动的语言风格。

动词方面，正向词表中 be 动词（was, were）占据多位，这可能表明，英译文本在句法结构上更依赖被动语态和过去式。为了进一步验证，笔者借助分词工具分析了两个语料库中被动语态和过去式的使用情况。

表 4 ANTC 和 CTTC 被动语态与过去式使用情况统计

语料库	被动语态数量	占比	过去式数量	占比 %
ANTC	762	4.00	4369	22.87
CTTC	674	12.94	2623	50.36%

表 4 显示，CTTC 过去式和被动语态使用占比远高于 ANTC。这可能源于汉英翻译过程中对源语信息的忠实传递，尤其是在描述历史遗迹和文化遗产时，译者倾向于使用被动句来强调客体（如 was built），而非突出主体的行为。此外，这种倾向或也表明，英译文本更偏向于静态描述，使其呈现

出较为正式和书面化的风格。这一特征表明西安旅游文本的翻译策略较为保守，强调信息传递的准确性，而忽略了文本的互动性。

另一方面，负向词表中 work、take、look、love 等高频动词的出现，表明 ANTC 可能更倾向于使用简单、直白的表达。例如，Take a walk along the river、Look at the stunning view、Love this place! 等句式强调游客的直接体验和互动。这种表达方式符合旅游文本的交际功能，即通过轻松、亲切的语言吸引游客。相比之下，翻译文本更受源文本结构的约束，较少采用类似的互动表达。

4. 结语

本研究基于语料库方法，以美国旅游文本语料库为参考，探讨了西安旅游英译文本的文本特征和词汇特征。数据分析表明，西安旅游英译文本的标准形符比相对较低，词汇丰富度有限，表明译文在词汇使用上较为保守。此外，英译文本的平均词长明显高于美国旅游文本，说明其用词偏正式。句长分析显示，英译文本的平均句长较长，句法结构较复杂，符合翻译文本较多使用被动语态、定语从句等结构的趋势。与此同时，关键词分析显示，美国旅游文本更倾向于使用简短、高频词，反映出其更注重可读性和吸引力，以降低阅读门槛、增强传播效果。综合来看，西安旅游文本英译在文化信息传递方面具有较强的历史文化导向，但在语言风格上表现出正式性与复杂性，与英语母语旅游文本的通俗化表达存在一定差异，这为进一步优化旅游文本翻译策略提供了参考。

本研究不仅为旅游文本英译的优化提供了数据支持，同时也对提升中国文化的国际传播效果具有实际意义。然而，本研究仍存在一定局限性，如语料库规模的限制以及对语篇层面特征的深入分析尚待进一步开展。未来研究可扩展语料库范围，结合人工智能技术进一步探讨旅游文本翻译的自动化策略，以提升翻译质量和跨文化传播效果。

参考文献：

[1]Gandin, S. (2013). Translating the language of tourism. A corpus-based study on the Translational Tourism English Corpus (T-TourEC). Procedia-Social and Behavioral Sciences, 95, 325-335.

[2]Halliday, M. A. K., & Matthiessen, C. (2014). Halliday’s Introduction to Functional Grammar (4th ed.). Routledge.

[3]Hogg, G., Liao, M. H., & O’Gorman, K. (2014). Reading between the lines: Multidimensional translation in tourism consumption. Tourism Management, 42, 157-164.

[4]Laviosa, S. (1998). Core patterns of lexical use in a comparable corpus of English narrative prose. Meta, 43(4), 557-570.

[5]Molina, L., & Albir, A. H. (2002). Translation techniques revisited: A dynamic and functionalist approach. Meta, 47(4), 498-512.

[6]Toury, G. (2012). Descriptive Translation Studies – and Beyond (2nd ed.). John Benjamins.

[7] 康宁 .(2012). 基于类比语料库的中国网站英语旅游文本语言分析 . 青岛科技大学学报 (社会科学版)(04),105-109.

[8] 李德超 & 唐芳 .(2015). 基于类比语料库的英语旅游文本文体特征考察 . 中国外语 (04),88-96.

作者简介：

谢晨阳（1993—），通讯作者，男，汉，河南驻马店，硕士研究生，助教，翻译、语料库语言学。

基金项目：

西安翻译学院 2023 校级科研项目青年专项：基于双语平行语料库的西安旅游文本英译研究，编号 :23B37。