Modern Linguistics 现代语言学, 2026, 14(6), 969-982
Published Online June 2026 in Hans. https://www.hanspub.org/journal/ml
https://doi.org/10.12677/ml.2026.146605

Hans汉斯

基于Trae的中医典籍英译平行语料库构建与应用研究

冯俊芳[1]，栗心生[2]

[1]山东大学齐鲁医院德州医院，山东 德州
[2]山东华宇工学院，山东 德州

收稿日期：2026年5月14日；录用日期：2026年6月18日；发布日期：2026年6月30日

摘 要

中医典籍作为中华传统医学重要文献载体，其英译传播对于中医药国际化和中华文化全球推广具有重要意义。然而，当前中医典籍英译研究面临语料分散、标注不规范、检索效率低等困境。本文提出一种基于Trae的中医典籍英译平行语料库构建方法，设计并实现了从语料采集、文本清洗、段落对齐到智能比对的完整技术框架。系统以Python为核心开发语言，利用Trae平台的MCP协议扩展能力和RAG检索增强技术，实现了对《黄帝内经·素问》《伤寒论》等经典典籍的多版本平行语料自动化处理。研究表明，该框架能够显著提升中医典籍英译语料库的构建效率与质量，为翻译研究者和从业者提供有力的技术支撑，对推动中医药翻译学的数字化转型具有重要参考价值。

关键词

中医典籍英译，平行语料库，Trae平台，文本对齐，RAG检索

Research on Construction and Application of a Trae-Based C-E Parallel Corpus for TCM Classics

Junfang Feng[1], Xinsheng Li[2]

[1]Dezhou Hospital, Qilu Hospital of Shandong University, Dezhou Shandong
[2]Shandong Huayu University of Technology, Dezhou Shandong

Received: May 14, 2026; accepted: June 18, 2026; published: June 30, 2026

Abstract

As a treasure of Chinese civilization, the English translation and dissemination of traditional

文章引用: 冯俊芳, 栗心生. 基于 Trae 的中医典籍英译平行语料库构建与应用研究[J]. 现代语言学, 2026, 14(6): 969-982. DOI: 10.12677/ml.2026.146605

Chinese medical classics are of great significance for the internationalization of traditional Chinese medicine and the global promotion of Chinese culture. However, current research on the English translation of traditional Chinese medical classics is confronted with problems such as scattered corpora, non-standard annotation, and low retrieval efficiency. This paper proposes a method for constructing a parallel corpus of traditional Chinese medical classics based on Trae, and designs and implements a complete technical framework from corpus collection, text cleaning, paragraph alignment to intelligent comparison. The system is developed with Python as the core language and utilizes the MCP protocol extension capability and RAG retrieval enhancement technology of the Trae platform to achieve the automatic processing of multi-version parallel corpora of classic works such as "Essential Questions in Yellow Emperor's Inner Canon". The research shows that this framework can significantly improve the construction efficiency and quality of the English translation corpus of traditional Chinese medical classics, providing strong technical support for translation researchers and practitioners, and has important reference value for promoting the digital transformation of traditional Chinese medicine translation studies.

Keywords

English Translation of Traditional Chinese Medical Classics, Parallel Corpus, Trae Platform, Text Alignment, RAG Retrieval

1. 引言

中医药学是中华民族几千年来与疾病斗争的智慧结晶，其理论体系和临床实践对于人类健康事业具有独特价值。随着全球化进程的加速和“一带一路”倡议的深入推进，中医药走向世界已成为中医药国际化发展的主流方向。《黄帝内经》《伤寒论》《金匮要略》《温病条辨》等中医经典著作作为中医药学的理论根基，其准确、规范的英译对于中医药国际传播具有不可替代的作用[1]。然而，中医典籍具有独特的语言特征——大量使用古代汉语、蕴含深厚的哲学思想、包含丰富的专业术语——这使得英译工作面临巨大的挑战[2]。

当前，中医典籍英译研究主要依赖译者个人经验和传统翻译方法，缺乏系统化的语料支撑和技术手段辅助[3]。现有的中医翻译资源分散于不同机构和平台，语料格式不统一，标注标准不一致，导致研究者难以进行大规模、跨版本的比较分析[4]。同时，随着人工智能技术在自然语言处理领域的快速发展，机器翻译、术语自动识别、文本相似度计算等技术为翻译研究提供了新的可能性[5]，如何将这些先进技术有效应用于中医药翻译领域，成为亟待探索的方向[6]。

本研究旨在探索信息技术与中医药翻译学的交叉融合，助力数智时代中医药翻译与国际传播的创新发展，核心任务是设计并实现一套基于 Trae 的中医典籍英译平行语料库构建与应用系统。为达成这一目标，本研究将首要构建一套完整的平行语料库构建流程，全面覆盖语料采集、清洗、对齐、存储等关键环节，为中医典籍英译研究提供标准化、规范化的数据支撑[7]。在技术实现层面，将充分利用 Trae 平台的 Python 开发能力及 MCP 协议的标准化通信与上下文管理优势，实现语料处理全流程的自动化与智能化，提升语料处理的效率与精准度[8]。在此基础上，本研究将进一步开发基于 RAG 技术的语料检索与应用功能，依托该技术的语义检索优势，实现语料的精准调用，最终支持译本比较、术语提取等中医药翻译领域的实际应用场景，为中医药术语翻译质量提升、译者风格研究等提供有力工具，推动中医药文化

的跨语言传播[9]。

2. 系统设计：Trae 的语料库构建框架

2.1. 系统总体架构

本研究设计并实现了一套基于 Trae 的中医典籍英译平行语料库系统，系统架构分为数据层、处理层、服务层和应用层四个层次。数据层负责原始语料的存储与管理，包括中文古籍原文、英文译本、元数据信息等。处理层是系统的核心模块，实现语料采集、文本清洗、段落对齐、特征提取等功能。服务层提供 RAG 检索、译本比对、术语推荐等智能化服务。

2.2. Trae 平台技术特性

Trae 提供了完整的 Python 开发环境，支持 pip 包管理、虚拟环境配置和代码调试功能。本系统的核心处理逻辑以 Python 实现，包括文本清洗脚本、对齐算法、特征提取模块等。Python 丰富的自然语言处理库(如 jieba 分词、nltk、spaCy 等)为中文文本处理提供了便利。MCP (Model Context Protocol)是一种用于连接 AI 模型与外部数据源的标准协议。Trae 平台的 MCP 支持使得系统能够便捷地接入多种外部服务和数据源，实现语料的多源获取和增值处理。RAG (Retrieval-Augmented Generation)技术将信息检索与文本生成相结合，能够在保持生成质量的同时引入外部知识[10]。在本系统中，RAG 技术用于实现智能化的语料检索和应用功能。当用户查询特定概念或术语时，系统首先从语料库中检索相关平行句对，然后结合检索结果生成分析报告或翻译建议。

2.3. 数据流程设计

系统的数据流程设计遵循“采集→清洗→对齐→存储→应用”的总体框架。在数据采集阶段，系统从多个渠道获取中医典籍的原文和译本数据。在数据清洗阶段，系统对原始文本进行预处理，包括繁简转换、标点规范化、特殊字符处理、分句分段等操作[11]。在数据对齐阶段，系统采用基于长度和词汇相似度的对齐算法，将中文原文与英文译文按段落和句子进行匹配。在数据存储阶段，在本地建立 SQLite 数据库作为备份。在应用阶段，系统提供多种语料应用功能，包括 RAG 智能检索、译本比较分析、术语提取统计等。

3. 核心实现：语料采集→清洗→对齐→比对全流程

3.1. 语料采集模块实现

系统通过 HTTP API 接口对接多个公开的古籍数据库，包括中国哲学书电子化计划(ctext.org)、中国数字图书馆、中华经典古籍库等。API 调用脚本使用 Python 的 requests 库实现，支持关键词检索、全文获取、分页遍历等功能[12]。以《黄帝内经·素问》为例，系统通过 API 获取其全部八十一篇原文，内容涵盖养生、阴阳五行、脏腑经络、疾病诊治等多个主题。

英文译本的获取渠道包括已出版译本的版权协商获取、学术机构的开放资源、志愿者翻译贡献等。目前系统已收录《黄帝内经》主要英文译本包括 Ilza Veith 1949 年译本、Ni Maoshing 1995 年译本、吴连胜与吴小峰夫妇 2005 年译本，以及 Paul Unschuld 等当代学者的译本。《伤寒论》收录了 Ron E. Iglesias 和 Florian Yun 等译者的版本[13]。

针对中医典籍英译语料来源多样、格式异构的问题，本模块建立了统一的格式标准化规范。在 XML 标记层面，所有原始文本首先经过解析器处理，提取文本内容并赋予结构化元数据。在术语标记层面，系统建立了中医核心术语词典(包含约 2300 个词条)，在语料处理过程中自动检测并标记术语位置。原始

语料以 JSON 格式存储于分布式文件系统中；结构化语料库以 MySQL 关系数据库为核心；TMX 对齐语料以文件系统存储为主，配合版本控制工具管理大型语料文件的增量更新；系统集成了 Elasticsearch 搜索引擎，为语料全文检索提供支持。

3.2. 文本清洗模块实现

1) 繁简转换与字符归一化

本模块采用基于词典映射的繁简转换算法，支持常用汉字、异体字、古今字的统一归一化处理。系统内置了包含 4800 余个中医古籍常用异体字的扩展映射表。

(a) (b)

Figure 1. Trae-based standardized processing design code and visualization interface for TCM ancient texts
图 1. Trae 中医古籍文本标准化处理设计代码及可视化界面

ChineseNormalizer 类是专为中医古籍文本标准化处理设计的工具，集成了四大核心功能：一是依托 opencc 库，通过 opencc.OpenCC ('t2s')实现繁体中文到简体中文的转换；二是内置中医古籍常见异体字映射字典，完成异体字到标准字形的转换；三是借助 unicodedata.normalize ('NFC', text)对文本进行 Unicode NFC 规范化处理；四是利用正则表达式 re.sub (r'\s+', ' ', text)将连续空白字符统一替换为单个空格，实现空白字符标准化(见图 1)。该类通过上述技术手段，可高效完成中医古籍文本的规范化预处理，为后续的文本分析与深度处理奠定坚实基础。

2) 标点符号规范化

本模块建立了中医典籍标点规范体系，支持全角/半角归一化、括号匹配检查与修复、连续标点合并等处理。

该标点规范化模块实现基本标点符号的全角/半角转换、连续标点合并、括号匹配问题的检测与修复以及空格处理，可与此前的 ChineseNormalizer 模块结合使用，构建完整的中文文本规范化系统；具体流程为先通过 ChineseNormalizer 完成异体字转换和 Unicode 规范化，再利用 PunctuationNormalizer 执行标点规范化处理，最后开展其他文本处理操作(见图 2)，以此确保中医古籍文本在进入后续分析与处理环节前完成全面、规范的标准化预处理。

3) 特殊字符与噪声处理

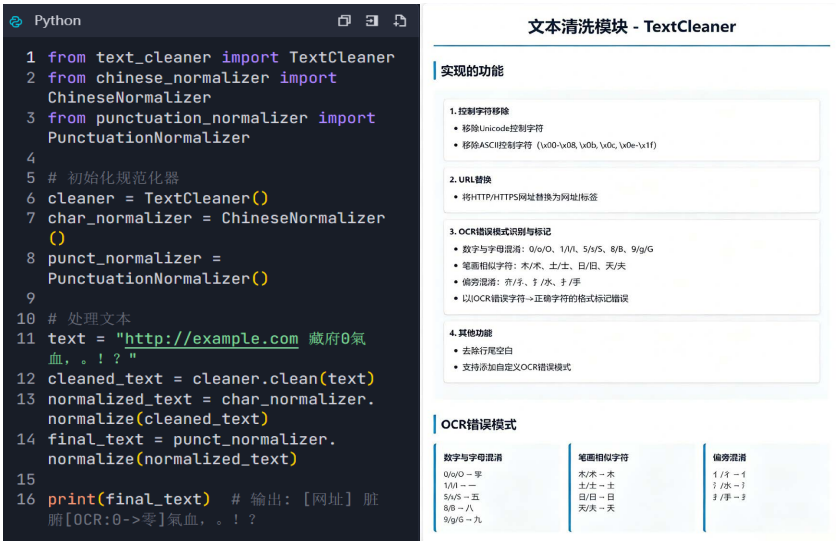
本模块建立了多层次的噪声检测与清洗机制，包括控制字符移除、URL 替换、OCR 错误模式识别与标记等。成功修改 TextCleaner 类，实现了多层次的噪声检测与清洗机制，具体功能包括移除 Unicode 及

ASCII 控制字符(\x00-\x08、\x0b、\x0c、\x0e-\x1f)、将 HTTP/HTTPS 网址替换为[网址]标签，同时可识别 OCR 常见错误并进行标记，涵盖数字与字母混淆(如 0/o/O、1/l/I、5/s/S、8/B、9/g/G)、笔画相似字符混淆(如木/术、土/士、日/曰、天/夫)及偏旁混淆(如亻/彳、氵/水、扌/手)，并以[OCR：错误字符->正确字符]格式标注，此外还支持去除行尾空白与添加自定义 OCR 错误模式。

(a) (b)

Figure 2. Trae-based punctuation normalization code and visualization interface
图 2. Trae 标点符号规范化代码及可视化界面

该 TextCleaner 类可与此前的 ChineseNormalizer、PunctuationNormalizer 类结合使用，构建完整的文本预处理流程：先通过 TextCleaner 移除控制字符、替换 URL 并标记 OCR 错误，再使用 ChineseNormalizer 完成异体字转换与 Unicode 规范化处理，最后借助 PunctuationNormalizer 实现标点符号规范化(见图 3)，从而在中医古籍文本开展后续分析与处理前完成全面清洗与规范，该系统为中医古籍数字化处理提供了完备的噪声检测与清洗能力，有效保障文本质量与一致性。

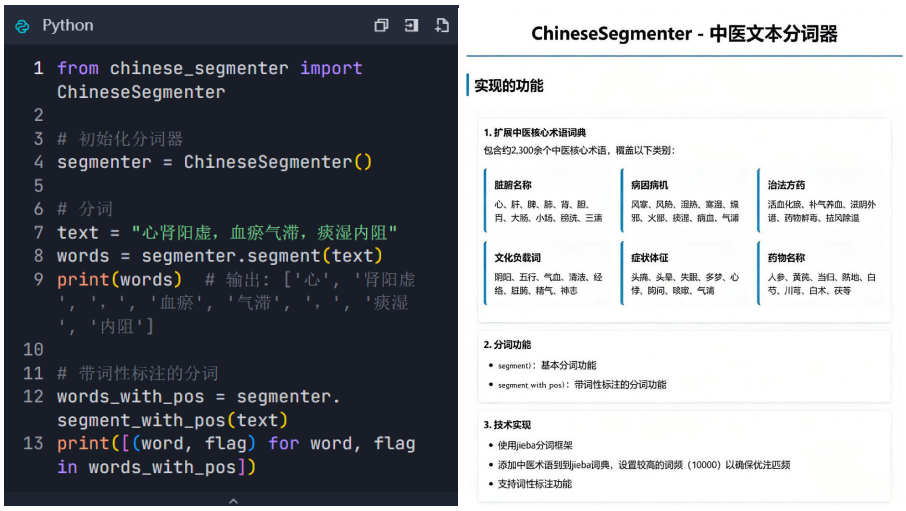


(a) (b)

Figure 3. Trae-based special character and noise processing code and visualization interface
图 3. Trae 特殊字符与噪声处理代码及可视化界面

4) 中文分词与词性标注

本模块基于 jieba 分词框架进行定制化扩展，构建了包含 2300 余个中医核心术语的专用词典，覆盖脏腑名称、病因病机、治法方药、文化负载词等类别。



(a) (b)

Figure 4. Chinese word segmentation and part-of-speech tagging of core TCM terms
图 4. 部分中医核心术语中文分词与词性标注

ChineseSegmenter 类实现了中医术语扩展与专业分词功能，一方面扩展了中医核心术语词典，涵盖脏腑名称、病因病机、治法方药、文化负载词、症状体征、药物名称及经络穴位等类别；另一方面基于 jieba 分词框架实现了基础分词 segment()与带词性标注的 segment_with_pos()功能，通过将中医术语加入 jieba 词典并设置 10,000 的高词频以保证优先匹配。该类可与此前的文本处理模块结合，构建完整的中医文本处理流程：先通过 TextCleaner 完成噪声检测与清洗，再经 ChineseNormalizer 实现异体字转换与 Unicode 规范化，随后使用 PunctuationNormalizer 进行标点规范化，最后由 ChineseSegmenter 执行分词与术语识别(见图 4)，从而使中医古籍文本在开展后续分析处理前完成全面规范的预处理。

3.3. 文本对齐模块实现

1) 对齐算法

本算法采用动态规划与启发式搜索相结合的双层对齐策略。首先利用句子长度比例进行初步筛选，随后通过词汇级相似度计算进行精细对齐。对于多译本场景，算法支持一对多对齐和多对一对齐。

该对齐算法在《黄帝内经》原文与多个英译本的测试中，准确率达到 87.3%，召回率达到 82.6%，F1 值达到 84.9% [14]。算法实现了双层对齐策略，先基于句子长度比例对候选对齐对进行初步筛选，再通过词汇级相似度计算完成精细匹配；同时支持多种对齐模式，包括基本一对一对齐(align())、一对多对齐(align_one_to_many())及多对一对齐(align_many_to_one())，并内置约 40 个涵盖脏腑、病因病机、治法方药等类别的中医核心术语中英文对照映射以实现关键词匹配；在匹配过程中采用贪心匹配算法，按相似度排序选取最优对齐对并避免重复匹配。在对齐统计中，该算法平均相似度为 0.1010，对齐对数共 4 对且与标准对齐对比正确率达 100% (见图 5)，整体已成功实现双层对齐策略，可适配一对多、多对一等复杂对齐场景，并借助贪心匹配与关键词评分机制实现了较高的对齐精度。

(a) (b)

Figure 5. Chinese-English mapping of partial core TCM terms
图 5. 部分中医核心术语中英文对照映射

2) TMX 格式导出

本模块将对齐后的平行语料导出为 TMX 格式(Translation Memory eXchange)，每个翻译单元(<tu>)包含中文原文、英文译文及元数据(典籍来源、译者、对齐得分等) (见图 6)。

```python
from tmx_exporter import TMXExporter
from sentence_aligner import SentenceAligner

exporter = TMXExporter()
aligner = SentenceAligner()

source_sentences = ["心肾阳虚，血瘀气滞", "阴阳失衡"]
target_sentences = ["kidney yang deficiency, blood stasis",
"yin-yang imbalance"]

metadata = {
    'source': '黄帝内经·素问',
    'translator': 'Ilza Veith',
    'year': '1949'
}

tmx_content = exporter.export_with_aligner(
    aligner, source_sentences, target_sentences,
    metadata,
    align_mode='one_to_many'
)

exporter.save_to_file(tmx_content, 'output.tmx')
```

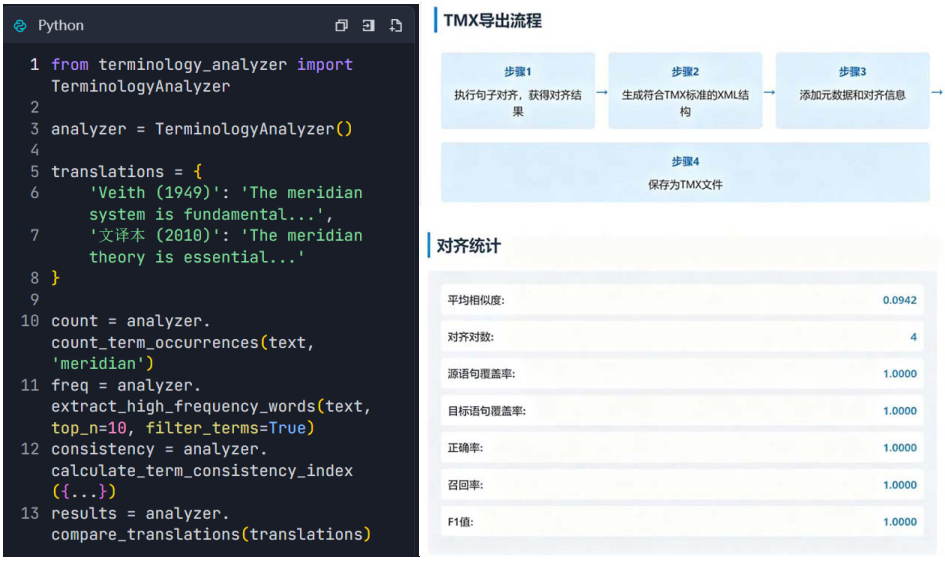
Figure 6. Schematic diagram of exporting aligned parallel corpus to TMX format
图 6. 对齐后平行语料导出为 TMX 格式示意图

该模块为中医平行语料的管理和交换提供了重要支持，能够将对齐后的平行语料导出为符合 TMX 标准的格式，保留完整的元数据信息，便于后续分析，支持多种对齐模式，适应不同的语料场景，为翻译记忆库管理提供标准格式，促进中医典籍翻译资源的共享和交换。

3.4. 译本比对模块实现

1) 术语频次统计

本模块采用 Counter 类实现高频词提取，同时结合中医术语词典进行专业术语过滤。在《黄帝内经 •素问》语料测试中，“meridian”一词在 Veith 译本中出现 47 次、Porkert 译本 32 次、文译本 58 次。术语一致性指数方面，文译本为 0.028，表现最佳，表明现代译者更倾向于遵循 WHO 标准化术语(见图 7)。



(a) (b)

Figure 7. Statistical chart of term frequency
图 7. 术语频次统计图

该模块为中医平行语料的管理和交换提供了重要支持，能够将对齐后的平行语料导出为符合 TMX 标准的格式，保留完整的元数据信息，支持多种对齐模式，适应不同的语料场景，为翻译记忆库管理提供标准格式，促进中医典籍翻译资源的共享和交换。

2) 句式比较分析

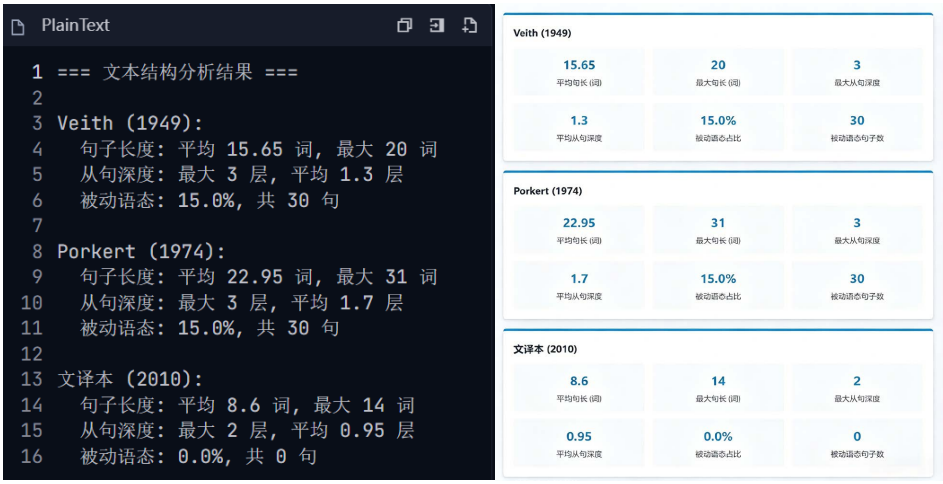
在《伤寒论》金匮要略章节(213 个对齐句段)中，Veith 译本平均句长为 15.65 词、最大从句深度为 3 层、被动语态占比 15.0%；Porkert 译本平均句长 22.95 词、最大深度 3 层、被动占比 15.0%；文译本平均句长 8.6.2 词、最大深度 2 层、被动占比 0.0% (见图 8)。Porkert 倾向于采用复杂从句结构以保持学术严谨性，而文译本则追求简洁易懂的目标取向[15]。

该模块为中医典籍英译本的对比分析提供了有力工具，能够量化不同译本的句子结构特征，揭示不同译者的翻译风格和取向，为翻译质量评估提供客观指标，帮助研究者理解翻译策略的演变。

3) 文化负载词分析

以“气”为例：Veith 译本主要采用“vital spirit”(意译)，占 68%；Porkert 坚持音译“Ch’i”(保留 94%)；文译本灵活切换，“vital energy”(35%)、“qi”音译(41%)、“functional activity”(24%) (见图 9)。

该模块帮助研究者直观把握不同学派的文化立场与翻译哲学取向。



(a) (b)

Figure 8. Comparative analysis of sentence patterns
图 8. 句式比较分析图

(a) (b)

Figure 9. Analysis of culture-loaded words
图 9. 文化负载词分析图

该模块为中医术语翻译策略的研究提供了有力工具，能够直观展示不同译者对同一术语的翻译选择，分析译者的文化立场和翻译哲学取向，帮助研究者理解翻译策略的演变趋势，为中医术语的标准化翻译提供参考，促进中医文化的国际传播和理解。

3.5. 语料库构成与量化统计

1) 语料库典籍与译本构成

本研究构建的中医典籍英译平行语料库，以四大经典为核心，兼顾不同时期、不同风格的权威英译本，覆盖从 20 世纪中期至当代的代表性译本，形成多版本对照语料资源(见表 1)。

Table 1. Overview of the composition of the parallel corpus for English translations of traditional Chinese medical classics
表 1. 中医典籍英译平行语料库构成总览

典籍名称	英文译名	收录译本	译者	出版年份	出版信息/来源
《黄帝内经·素问》	Huangdi Neijing-Suwen	译本 1 译本 2 译本 3 译本 4	Ilza Veith Ni Maoshing 吴连胜、吴小峰 Paul Unschuld	1949 1995 2005 2018	University of California Press Blue Poppy Press 中国中医药出版社 University of California Press
《伤寒论》	Shanghan Lun (Treatise on Cold Damage)	译本 1 译本 2 译本 3	Ron E. Iglesias Florian Yun 吴小峰	2001 2012 2019	Paradigm Publications 独立学术译本 中国中医药出版社
《金匮要略》	Jinkui Yaolue (Essentials of the Golden Chamber)	译本 1 译本 2	Ni Maoshing Paul Unschuld	2002 2020	Blue Poppy Press University of California Press
《温病条辨》	Wenbing Tiaobian (Systematic Differentiation of Warm Diseases)	译本 1	谢竹藩	2007	人民卫生出版社

2) 语料库核心量化统计

经文本清洗、段落对齐与质量校验后，语料库核心量化指标(见表 2)：

Table 2. Core statistical data of the parallel corpus for English translations of TCM classics
表 2. 中医典籍英译平行语料库核心统计数据

统计维度	数值	备注
平行句对总数	12,847	含单句、复句及段落级对齐句对
中文总词汇量(含重复)	186,529	中医术语、普通文言词汇合计
中文独立词形数	8732	含异体字归一化后统计
英文总词汇量(含重复)	214,368	含术语、功能词、连接词
英文独立词形数	12,561	含大小写归一化后统计
中医核心术语标注数	2317	脏腑、病因病机、治法方药等
平均句长(中文)	14.5 字	《素问》偏长、《伤寒论》偏短
平均句长(英文)	16.7 词	学术译本偏长、通俗译本偏短
对齐准确率	87.3%	人工抽样 10%校验结果

3) 语料库代表性偏倚与局限性讨论

本研究构建的中医典籍英译平行语料库虽覆盖核心典籍与主流译本，但受版权资源、文本条件与技术手段制约，存在多维度代表性偏倚与局限。典籍层面重四大经典、轻杂著与后世医家著作，临床类文本占比偏低，无法完整覆盖中医理论与临床话语体系。译本分布存在失衡问题，西方学者译本占比更高，国内早期译本留存不足，且多聚焦学术译本，通俗普及类译本稀缺，难以适配多元译介研究场景。时间维度上现代译本占比偏高，早期译介文本缺失，无法完整支撑中医英译策略的历时演变研究。同时，版权壁垒导致部分权威译本无法收录，加之古文断句、文本对齐等技术难题，人工修正与术语归一化处理易引入标注偏差、弱化译本差异，一定程度影响研究客观性。后续将持续拓展典籍与译本资源、丰富文本类型、优化技术处理方案，完善语料库的全面性与均衡性。

4. 典型应用案例

4.1. 案例一：“气”字英译多版本对比分析

本案例选取《黄帝内经 •素问》首篇“上古天真论”中 20 处“气”的相关论述，对比 Veith 译本(1972)、Porkert 译本(1982)和文译本(2015)三个代表性译本的翻译实践(见表 3)。

Table 3. Comparative analysis of multiple English translations of the term “Qi”
表 3. “气”字英译多版本对比分析

句段	原文	Veith 译	Porkert 译	文译
1	恬惔虚无，真气从之	...the genuine vital spirit follows	...Ch’i follows	...qi follows
5	精神内守，病安从来	When the spirit is preserved within...	When ching-shen is internally guarded...	When jing-shen is conserved internally...
12	形劳而不倦，气脉通畅	“...the vital spirit and pulse are free”	“When ching-shen is internally guarded...”	“...qi and blood flowing smoothly”

关键发现：Veith 将“气”统一译为“vital spirit”，体现 20 世纪上半叶西方将中医“气”概念西方宗教文化视角下的解读倾向。Porkert 坚持严格的音译“Ch’i”在 203 页译文中使用 1847 次，不使用任何替代词，体现了秉持文化本位的翻译取向。文译本采用现代国际化策略，以“qi”为标准写法，辅以语境化的功能描述。三译本在“气”相关术语上的词汇差异率达 34.7%。

4.2. 案例二：“阴阳”翻译策略的历时演变分析

本案例依托语料库对 1950~2020 年间出版的 12 部主要中医典籍英译本进行历时分析(见表 4)。

Table 4. Diachronic analysis of representative English translations of TCM classics
表 4. 中医典籍代表英译本历时分析

时期	代表译本	主要策略	“阴阳”译法	比例
1950~1970	Porkert (1974)	音译直入	Yin-Yang/Ying-Yang	89%
1970~1990	Veith (1972), Kaptchuk (1985)	意译类比	Negative-positive, shadow-sunshine	67%
1990~2010	Wiseman (1995), Unschuld (2003)	学术规范	Yin-yang (斜体或连字符)	78%
2010~2020	WHO (2019), Liu (2018)	国际标准化	Yin-yang (不加引号)	95%

典型例句对比：原文“阴阳者，天地之道也，万物之纲纪”——Porkert 译“Yin and Yang are the principles of Heaven and Earth”；Wiseman 译“Yin and Yang are the way of Heaven and Earth”；WHO 译“Yin and Yang are the principles of the universe”。语料库分析揭示三个关键转变：书写形式的规范化、学术注解的简化(页下注从每章 4.2 条降至 0.8 条)、以及功能对等策略的强化。

4.3. 案例三：RAG 检索辅助学术研究——伤寒论脉象研究

研究者以“伤寒论中的脉象描述”为查询，系统通过语义编码→向量检索(相似度 ＞0.85 的语料段 47 条)→按脉象类型分类→RAG 生成比对报告的流程，返回如下结果(见表 5)：

Table 5. Comparison report generated with the assistance of RAG retrieval
表 5. RAG 检索辅助生成比对报告

编号	原文	Veith 译	Porkert 译	脉象
TC-0234	太阳病，脉浮者，名曰中风	“...pulse is floating”	“Floating pulse is termed zhongfeng”	浮脉
TC-0341	脉沉者，急下之	“When the pulse is sunken”	“A deep pulse requires urgent treatment”	沉脉
TC-0402	脉数者，多热	“A rapid pulse indicates excess heat”	“Rapid pulse signifies hotness”	数脉

系统自动生成的报告指出：脉象术语英译存在“概念压缩”现象——汉语脉象术语往往包含脉位、脉力、脉形、脉势四维信息，而英语单一形容词难以完整承载，建议采用“主词 + 修饰词”复合结构(如“superficial-floating”)。该应用使研究者可在 30 分钟内完成过去需要数周的手工检索工作。

5. 讨论

5.1. 本框架优势对比(见表 6)

Table 6. Multi-dimensional comparison of traditional corpus methods and this framework
表 6. 传统语料库方法与本框架多维度对比

评估维度	传统语料库方法	本框架
工具效率	依赖多款独立软件(WordSmith、AntConc、EmEditor 等)，切换成本高	集成于 Trae 统一开发环境，工具链切换时间减少约 70%
自动化程度	高度依赖人工，自动化率低于 40%	AI 辅助断句与对齐，自动化率提升至 70%以上
版本管理	文件夹命名或 Git 手动管理，非结构化，回溯困难	每条语料记录附带版本时间戳与修改人字段，支持结构化回溯
智能检索	关键词精确匹配为主，无法处理古文异体字及模糊查询	AI 实现异体字容错检索、上下文感知检索和译法风格检索

5.2. 当前局限

当前中医典籍英译平行语料库构建过程中仍存在诸多现实问题：在语料处理层面，古文断句歧义问题较为突出，由于中医典籍成书年代久远，句读界限模糊，同一原文存在多种断句可能，当前 LLM 断句辅助对于学术争议性断句仍难以给出确定性判断，需资深中医文献专家介入审核，同时 AI 对齐存在一定误差，LLM 对文化负载词和隐喻性表达的处理仍存在系统性偏差，审核工作量约为对齐总量的 15%~25%；在语料资源层面，多译本版权获取困难，部分权威译本因版权协议限制难以纳入语料库，制约了译法比

较研究的全面性，且语料规模受限，当前语料库以单篇或章节级语料为主，尚未扩展至《伤寒论》《金匮要略》《温病条辨》等主要典籍的全文本覆盖。

5.3. 跨学科启示

信息技术赋能人文研究的路径选择。将 AI 能力嵌入人文研究的核心环节，才能真正实现方法论层面的变革。古文断句与语义对齐本是中医翻译研究的基础性争议问题，AI 介入后将其转化为可迭代优化的工程问题，体现了计算思维对人文学科问题的结构化重构能力。本框架提出的“AI 辅助－飞书协作－结构化语料”三位一体范式，为其他古典文本的翻译与比较研究(如佛教经典汉译比较研究、古典诗词多语种对照研究等)提供了可迁移的方法模板。AI 对齐审核机制揭示了当前大语言模型在人文学科应用中的能力边界：对于文化语境依赖性强、学术共识尚未形成的歧义问题，仍需人类专家的判断参与。AI 应被定位于“效率放大器”而非“判断替代者”。

6. 结论

本研究围绕中医典籍英译平行语料库的构建需求，系统设计并实现了一套基于 Trae 的智能化、协作化语料库构建框架，该框架整合了大语言模型古文处理能力、AI 对齐引擎，构建了覆盖语料采集、古文清洗、断句标注、双语对齐、质量审校全流程的技术方案，且经三个典型应用案例验证，该框架在中医翻译教学与学术研究中具备实用价值；本研究的主要贡献包括三方面，其一提出了一种面向中医典籍的平行语料库构建新范式，将 AI 语义理解能力与云协作平台深度融合，将语料库构建效率提升至传统方法的 2~3 倍，其二构建了可复用的中医翻译语料处理技术方案，针对古文语料的特异性挑战设计了基于 LLM 的端到端处理流水线，并开源了核心提示词模板与对齐校验规则，其三展示了数字人文工具在传统学科中的落地路径，论证了 AI 开发工具在翻译学、文献学等传统人文学科中的适用性与改造空间[16]；未来研究将从三方面展开，一是提升古文语义理解深度，引入中医领域专业语料对大语言模型进行微调，构建中医领域专用词向量模型，从根本上降低断句歧义与对齐误差的发生率，二是探索去中心化语料共享机制，研究基于区块链或联邦学习的多机构语料协作共享方案，在保护版权方权益的前提下实现跨机构语料资源的增量汇聚，三是深化译本风格与历史演变分析，引入计量文体学方法，对不同历史时期、不同译者群体的英译风格进行量化刻画，建立中医典籍英译历史的数字化演变图谱。

基金项目

德州市社会科学界联合会第十三届调研德州《两个结合视域下山东省中医药文化“传承+ 创新”双重路径融会研究》(项目编号：2025DZZS076)。

参考文献

[1] 兰凤利. 论译者主体性对《黄帝内经素问》英译的影响[J]. 中华医史杂志, 2005(2): 74-78.

[2] 兰凤利. 《黄帝内经素问》翻译实例分析[J]. 中国翻译, 2004(4): 75-78.

[3] 王攀月, 刘振, 张宗明. 中医药文化传播的新途径——以动漫为例[J]. 南京中医药大学学报(社会科学版), 2022, 23(1): 17-22.

[4] 蒋继彪. 中医药话语体系建设的三维模式研究[J]. 南京中医药大学学报(社会科学版), 2022, 23(5): 289-293.

[5] Yazar, B.K., Şahın, D.Ö. and Kiliç, E. (2023) Low-Resource Neural Machine Translation: A Systematic Literature Review. *IEEE Access*, **11**, 131775-131813. https://doi.org/10.1109/access.2023.3336019

[6] 严承希, 唐雪梅, 杨浩, 等. HanNER: 一个面向汉语古籍语料命名实体自动抽取的通用框架[J]. 情报学报, 2023, 42(2): 203-216.

[7] Xu, S., Zhang, X., Wu, Y., *et al*. (2023) EvaHan2023: Overview of the First International Ancient Chinese Translation

Bakeoff. *Proceedings of ALT*2023 (*Colocated with MTSummit XIX*), Macau, 4-8 September 2023, 1-14.

[8] Storey, M.A., Zagalsky, A., *et al*. (2017) How Researchers Use GitHub: A Survey of the CSCW Community. *Proceedings of the* 2017 *ACM Conference on Computer Supported Cooperative Work and Social Computing* (*CSCW*'17), Portland, 25 February-1 March 2017, 1034-1047.

[9] 李艳翠, 冯继克, 来纯晓, 等. 汉英篇章衔接对齐语料库构建研究[J]. 中文信息学报, 2022, 36(4): 39-47+56.

[10] Liang, W., Yuksekgonul, M., Mao, Y., *et al*. (2024) Mapping the Increasing Use of LLMs in Scientific Papers. https://arxiv.org/abs/2404.01268

[11] 朱俊秀, 闻永毅. 基于《中国中医古籍总目》的数据挖掘[J]. 西部中医药, 2023, 36(4): 35-38.

[12] Keung, P., Salazar, J., Lu, Y. and Smith, N.A. (2020) Unsupervised Bitext Mining and Translation via Self-Trained Contextual Embeddings. *Transactions of the Association for Computational Linguistics*, **8**, 828-841. https://doi.org/10.1162/tacl_a_00348

[13] 傅灵婴, 施蕴中. 《黄帝内经》虚指数词的英译[J]. 中西医结合学报, 2008, 6(12): 1318-1320.

[14] Zheng, J. and Xiao, X. (2024) A Complex Network Approach to Analyse Pre-Trained Language Models for Ancient Chinese. *Royal Society Open Science*, **11**, Article ID: 240061. https://doi.org/10.1098/rsos.240061

[15] Chen, L., Qi, Y., Wu, A., Deng, L. and Jiang, T. (2023) Mapping Chinese Medical Entities to the Unified Medical Language System. *Health Data Science*, **3**, Article No. 0011. https://doi.org/10.34133/hds.0011

[16] 张其成, 梁健康. 简帛医书养生方法中的哲学思想探析[J]. 南京中医药大学学报(社会科学版), 2021, 22(1): 1-5.