

《中国学术期刊网络出版总库》、CNKI系列数据库、AMI及维普数据库入选集刊

# 语料库语言学

CORPUS LINGUISTICS

2026

第25辑

北京外国语大学中国外语与教育研究中心  
中国英汉语比较研究会语料库语言学专业委员会  
中国英汉语比较研究会语料库翻译学专业委员会  
何 伟 主 编  
刘鼎甲 副主编

idiom principle  
context keywords pattern grammar Sinclair  
COBUILD CLEC collocation local grammar word embeddings  
AntConc DEAP multifactorial analysis  
big data corpus WordSmith  
Brown Crown TECCL  
BNC corpus-as-method MDA semantic prosody  
COCA co-selection frequency ToRCH  
concordance iWriteBaby  
corpus-as-theory ParaConc phraseology

语料库语言学  
2026 第25辑

外研社

外语教学与研究出版社  
FOREIGN LANGUAGE TEACHING AND RESEARCH PRESS



corpus.bfsu.edu.cn

语料库语言学  
(半年刊)

Corpus Linguistics  
(Biannual)

主 管：中华人民共和国教育部  
主 办：北京外国语大学  
承 办：中国外语与教育研究中心  
中国英汉语比较研究会  
语料库语言学专业委员会  
语料库翻译学专业委员会  
出 版：外语教学与研究出版社

Administered by the Ministry of Education of China  
Directed by Beijing Foreign Studies University  
Edited at the National Research Centre for Foreign  
Language Education, Corpus Linguistics Society  
of China and Corpus Translation Studies Society  
of China  
Published by Foreign Language Teaching and Research Press

刊名题字：崔希亮

Journal Name Calligraphy: Cui Xiliang

主 编：何 伟

Editor-in-Chief: He Wei

副 主 编：刘鼎甲

Executive Associate Editor: Liu Dingjia

责任校对：李添豪 杨瑞雪  
刘若冰 江姝婷

Proofreaders: Li Tianhao, Yang Ruixue,  
Liu Ruobing, Jiang Shuting

编审委员会（按姓氏音序）  
主 任：  
许家金（北京外国语大学）  
秦洪武（曲阜师范大学）

Editorial Board (in alphabetical order)  
Chair:  
Xu Jiajin (Beijing Foreign Studies University)  
Qin Hongwu (Qufu Normal University)

委 员：  
冯志伟（教育部语言文字应用研究所）  
顾曰国（中国社会科学院）  
何安平（华南师范大学）  
胡开宝（上海外国语大学）  
雷 蕾（上海外国语大学）  
李文中（浙江工商大学）  
刘泽权（河南大学）  
陆小飞（美国宾州州立大学）  
濮建忠（浙江工商大学）  
陶红印（美国加州大学洛杉矶分校）  
王克非（北京外国语大学）  
卫乃兴（北京航空航天大学）  
文秋芳（北京外国语大学）  
杨惠中（上海交通大学）

Members:  
Feng Zhiwei (Institute of Applied Linguistics, MOE)  
Gu Yueguo (Chinese Academy of Social Sciences)  
He Anping (South China Normal University)  
Hu Kaibao (Shanghai International Studies University)  
Lei Lei (Shanghai International Studies University)  
Li Wenzhong (Zhejiang Gongshang University)  
Liu Zequan (Henan University)  
Lu Xiaofei (The Pennsylvania State University)  
Pu Jianzhong (Zhejiang Gongshang University)  
Tao Hongyin (University of California, Los Angeles)  
Wang Kefei (Beijing Foreign Studies University)  
Wei Naixing (Beihang University)  
Wen Qiufang (Beijing Foreign Studies University)  
Yang Huizhong (Shanghai Jiao Tong University)

电 话：（010）88817771  
电子邮箱：bfsucrg@sina.com  
投稿网址：http://lyly.chinajournal.net.cn

本刊地址：北京市西三环北路19号北京外国语大学  
中国外语与教育研究中心  
《语料库语言学》编辑部（100089）

\*本刊获北京外国语大学“双一流”建设经费资助

版权声明

本刊已被《中国学术期刊网络出版总库》、CNKI系列数据库及维普数据库收录。如作者不同意被收录，请在来稿时向本刊声明，本刊将作适当处理。

# 语料库语言学

CORPUS LINGUISTICS

2026 第 25 辑

何 伟 主 编

刘鼎甲 副主编

外语教学与研究出版社

FOREIGN LANGUAGE TEACHING AND RESEARCH PRESS

北京 BEIJING

## 图书在版编目 (CIP) 数据

语料库语言学. 2026. 第 25 辑 / 何伟主编 ; 刘鼎甲副主编. — 北京 : 外语教学与研究出版社, 2026. 6. — ISBN 978-7-5213-7408-7

I. H03-55

中国国家版本馆 CIP 数据核字第 2026Z0E348 号

## 语料库语言学 2026 第 25 辑

YULIAOKU YUYANXUE 2026 DI-25 JI

出 版 人 王 芳  
责任编辑 赵 雪  
责任校对 王雨舟 白小羽  
封面设计 锋尚设计  
出版发行 外语教学与研究出版社  
社 址 北京市西三环北路 19 号 (100089)  
网 址 <https://www.fltrp.com>  
印 刷 北京盛通印刷股份有限公司  
开 本 787×1092 1/16  
印 张 9.75  
字 数 206 千字  
版 次 2026 年 6 月第 1 版  
印 次 2026 年 6 月第 1 次印刷  
书 号 ISBN 978-7-5213-7408-7  
定 价 60.00 元

如有图书采购需求, 图书内容或印刷装订等问题, 侵权、盗版书籍等线索, 请拨打以下电话或关注官方服务号:

客服电话: 400 898 7008

官方服务号: 微信搜索并关注公众号“外研社官方服务号”

外研社购书网址: <https://fltrp.tmall.com>

物料号: 374080001



记载人类文明  
沟通世界文化  
[www.fltrp.com](http://www.fltrp.com)

# 目 录

## 智能语言研究

|                               |     |     |    |
|-------------------------------|-----|-----|----|
| AI赋能语言学研究：以扩展意义单位分析为例.....    | 陈 功 | 沙吉亮 | 1  |
| AI赋能语料库建设：以《论语》中俄平行语料库为例..... | 葛晓帅 | 韩 悦 | 14 |
| 智能评估嵌入翻译过程：设计与应用初探 .....      | 夏 云 | 秦洪武 | 25 |

## 研究论文

|                                            |         |     |
|--------------------------------------------|---------|-----|
| 基于Praat内置神经网络的自适应LPC阶数共振峰提取 .....          | 蒋博文     | 41  |
| 中外学术期刊英文摘要认知难度历时变化研究：基于信息熵与依存距离的量化分析 ..... | 刘国兵 侯佳欣 | 58  |
| 中国英语学习者与英语本族语者指物关系代词选择的多因素研究 .....         | 郝美佳 唐瑞梁 | 73  |
| 俄罗斯主流媒体对李白形象的建构研究：基于道琼斯语料库的话语历史分析 .....    | 杨玉琦     | 85  |
| 东南亚高级汉语学习者介词框式结构习得实证研究 .....               | 刘博洋 徐俊丽 | 101 |
| 基于文本挖掘的国家生态发展观探究：以“十四五”规划英译文本为例 .....      | 黎曜玮 张 朦 | 114 |

## 研究综述

|                 |     |     |
|-----------------|-----|-----|
| 西夏文数字化的进程 ..... | 孙 森 | 130 |
|-----------------|-----|-----|

## 书刊评介

|                              |     |     |
|------------------------------|-----|-----|
| 《功能性译者风格研究：语料库辅助研究法》述评 ..... | 郭嘉雯 | 140 |
| 《文体学：文本、认知与语料库》述评 .....      | 王艳伟 | 147 |

# Contents

## AI for language studies

|                                                                                                            |                           |    |
|------------------------------------------------------------------------------------------------------------|---------------------------|----|
| AI-Empowered Linguistic Research: A Case Study of Extended Units of Meaning<br>.....                       | Gong Chen and Jiliang Sha | 1  |
| AI-Empowered Corpus Construction: A Case Study of the <i>Lunyu</i> Chinese-Russian<br>Parallel Corpus..... | Xiaoshuai Ge and Yue Han  | 14 |
| Integrating AI-Powered Assessment Into Translation Process: Design and Application<br>.....                | Yun Xia and Hongwu Qin    | 25 |

## Research articles

|                                                                                                                                                                         |                              |     |
|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------|------------------------------|-----|
| Formant Extraction With Adaptive LPC Orders Based on the Praat-Embedded Neural<br>Network Module.....                                                                   | Bowen Jiang                  | 41  |
| A Diachronic Study Based on Information Entropy and Dependency Distance for<br>Cognitive Difficulty of English Abstracts in Chinese and International Journals<br>..... | Guobing Liu and Jiaxin Hou   | 58  |
| A Multifactorial Analysis of the Choice of Relativizers “which” and “that” Between<br>Chinese EFL Learners and Native English Speakers<br>.....                         | Meijia Hao and Ruiliang Tang | 73  |
| Constructing the Image of Li Bai in Russian Mainstream Media: A Historical Discourse<br>Analysis Based on the Dow Jones Corpus .....                                    | Yuqi Yang                    | 85  |
| An Empirical Study on the Acquisition of Prepositional Frame Structures by Advanced<br>Chinese Learners From Southeast Asia .....                                       | Boyang Liu and Junli Xu      | 101 |
| Research on National Ecological Development Outlook Through Text Mining: A Case Study<br>of the English Version of the 14th Five-Year Plan.....                         | Yaowei Li and Meng Zhang     | 114 |

## Review articles

|                                                      |         |     |
|------------------------------------------------------|---------|-----|
| The Digitalization Process of the Tangut Script..... | Sen Sun | 130 |
|------------------------------------------------------|---------|-----|

## Book reviews

|                                                                                                   |             |     |
|---------------------------------------------------------------------------------------------------|-------------|-----|
| A Review of <i>Researching Translator’s Functional Style: A Corpus-Assisted Approach</i><br>..... | Jiawen Guo  | 140 |
| A Review of <i>Stylistics: Text, Cognition and Corpora</i> .....                                  | Yanwei Wang | 147 |

# AI赋能语言学研究：以扩展意义单位分析为例<sup>\*</sup>

对外经济贸易大学 陈 功 沙吉亮

**提 要：**当前，生成式人工智能深入应用于多个领域，并为语言学研究带来广泛机遇，本研究旨在考察具有代表性的生成式AI工具ChatGPT-5在辅助扩展意义单位分析方面的实际表现。研究结果表明：在核心词的搭配与类联接识别及统计方面，GPT-5展现出较强的量化分析能力；在语义倾向与语义韵的判断上，该模型不仅能进行较为准确的分析，还可为研究者的深度质性解读提供思路启发。此外，指令工程在扩展意义分析全过程中具有关键作用，受大语言模型注意力机制的影响，针对性的指令词设计能够显著提升AI的语料分析能力。基于以上发现，本研究认为GPT-5可有效辅助研究者开展扩展意义单位分析，并在提升语言分析效率与质量方面具有重要价值。

**关键词：**扩展意义单位；ChatGPT；语义倾向；语义韵

## AI-Empowered Linguistic Research: A Case Study of Extended Units of Meaning

Gong Chen and Jiliang Sha

**Abstract:** Against the backdrop of the deep integration of generative artificial intelligence into various fields and the broad opportunities it brings to linguistic research, this study aims to examine the performance of a representative generative AI tool, ChatGPT-5, in assisting extended units of meaning analysis. The findings indicate that GPT-5 demonstrates strong quantitative analysis capabilities in identifying and

<sup>\*</sup> 本文系对外经济贸易大学中国特色哲学社会科学研究专项“中国特色概念国际表达的修辞策略研究”（项目编号：23YB01）的阶段性研究成果。

陈功为本文通信作者。

作者贡献：

陈功：选题构思、研究方法、数据收集、讨论结论、初稿撰写、修改润色、字数占比（60%）；

沙吉亮：选题构思、研究方法、数据分析、讨论结论、初稿撰写、字数占比（40%）。

statistically analyzing collocations and colligations of core words. In identifying semantic preferences and semantic prosody, the model not only produces relatively accurate analyses, but also provides insights that support more in-depth qualitative interpretation. In addition, prompt engineering plays a crucial role throughout the entire process of extended units of meaning analysis. Influenced by the attention mechanism of large language models, targeted design of prompt words can substantially enhance the model's corpus analysis capability. Based on these findings, this study argues that GPT-5 can effectively assist researchers in conducting extended units of meaning analysis and holds considerable value in improving both the efficiency and quality of linguistics analyses.

**Keywords:** extended units of meaning; ChatGPT; semantic preference; semantic prosody

## 1 引言

深度学习（deep learning）技术的发展，特别是Transformer架构的提出，促成了作为生成式人工智能（generative artificial intelligence，简称生成式AI）的大语言模型在自然语言处理（natural language processing，简称NLP）领域的迅速崛起。目前，已有包括DeepSeek、ChatGPT等在内的一系列基于Transformer架构的大语言模型。此类模型通过数以亿万计的参数训练，采用自注意力机制（self-attention），通过并行捕捉处理长距离依赖关系，使其在自然语言处理任务中表现优秀。

基于这些技术的发展，大语言模型在辅助研究者进行研究任务处理方面表现出了强大的潜能。然而，对于大语言模型在外语研究中的实用探究却相对缺乏，尤其是针对人工智能辅助研究者进行语言学量化与质性分析表现的相关研究较少。仅有许家金等（2024）在其著作《大语言模型的外语教学与研究应用》中对生成式AI的大语言模型在语言学研究中的相关应用做了进行较为详细的探究。鉴于此，本研究将以Sinclair（1996，2004）提出的“扩展意义单位模型”为对象进行深入研究，以OpenAI最新一代大语言模型GPT-5为例，探讨人工智能在该研究过程中对研究者的效用。

## 2 研究背景与文献综述

### 2.1 生成式AI在语言研究中的应用

生成式AI的出现为语言学研究带来了巨大变革，语言学的每个细分领域都不同程度地受到生成式AI的影响，进而产生了新的研究议题。总的来说，这些变化

主要集中于两个方面。一是语言学领域研究主题的拓展。生成式AI不仅推动了传统语言学的再思考（袁毓林，2024；Piantadosi，2024），还催生了新的学科议题。譬如，研究者开始广泛关注外语教学中学习者与AI的互动过程，研究AI对外语教学的影响（蔡薇，2023；郭茜等，2023；胡壮麟，2023；秦颖，2023；陈茉等，2024；杨连瑞，2024）。二是语言学研究方法的革新。生成式AI的工具属性为语言研究的各个层面提供了支持。如在社会语言学领域，Stell等（2025）将生成式AI工具使用与Labov的访谈原则相结合，系统性地探究了其在社会语言学调查中的可行范式。在语言测试与评估领域，Luc Ha等（2024）通过对多名资深教师的访谈发现，生成式AI工具在提升命题效率与题目多样性方面表现优异。诸如此类的研究证明了生成式AI在研究主题选定、研究思路拓展、研究方法支持等多个环节具备可行之处。

当前，国内研究多聚焦于生成式AI与语言研究方向前景、政策支持等宏观领域（文秋芳等，2024），针对AI在语言研究中的具体使用未做充分探讨，少有细颗粒度地对生成式AI在某一具体语言研究任务中的作用和表现做探究。本研究将以此为侧重点，测试并评估生成式AI工具在扩展意义单位分析中的表现，详细分析其在语言研究中的可行性。

## 2.2 扩展意义单位

扩展意义单位（extended unit of meaning）这一概念最早由Sinclair于1996年在其著作《寻找意义单位》（*The Search for Units of Meaning*）中提出，并在2004年出版的《信任文本：语言、语料库与话语》（*Trust the Text: Language, Corpus and Discourse*）中进一步发展。Sinclair（1996，2004）通过COBUILD语料库实例分析，建立了扩展意义单位的分析模型，该模型由核（core）、搭配（collocation）、类联接（colligation）、语义倾向（semantic preference）和语义韵（semantic prosody）五部分组成。图1以true feelings为例，直观展现了Sinclair对扩展意义各个单位的描写。核是扩展意义单位向外扩展的起点和出发点，构成了词项作为整体出现的证据，由一个或多个词语组成。搭配是指两个或多个词语在文本中短距内的共现。类联接指某一语法类别（grammatical class）或结构型式（structural pattern）与其他类别或型式，或与某个词语或短语的共现关系。语义倾向是对共享某一语义特征的常规共现词语的限制。语义韵是整个意义单位（复合词项）所传达的态度意义（Sinclair，1996；卫乃兴，2012；濮建忠，2021）。从意核到语义韵这五个成分共同组成了扩展意义单位这一形式功能统一体，其中意核、搭配和类联接属于下面的形式（结构）层，反映统一体的“形式”面，而语义倾向和语义韵则属于上面的意义（功能）层，反映统一体的“意义”面（濮建忠，2020）。

|      |      |                         |                     |           |               |
|------|------|-------------------------|---------------------|-----------|---------------|
| 5    | 语义韵  | 犹豫或无能                   |                     |           |               |
| 4    | 语义倾向 |                         | “表达”                |           |               |
| 3    | 类联接  | 情态动词等                   | 动词                  | 所有格       | 名词词组          |
| 2    | 搭配   | 搭配词                     | 搭配词                 | 搭配词       | 节点词           |
| 1    | 意核   |                         |                     |           | 意核词（组）        |
| 典型例证 |      | reluctant to/could/will | express/communicate | their/his | true feelings |

图1 Sinclair（1996）对true feelings所在扩展意义单位的描写  
（转引自濮建忠，2020）

Sinclair提出的扩展意义单位是对Firth的“语境意义观”，即语言的意义是一种基于典型的、反复出现的，以及可重复观察到的语言事实构建起来的抽象模式，这一观点在具体的可操作层面的落实，实现了对意义的统一描写，具备重要的理论与实践意义（何安平，2011）。然而，扩展意义单位模型对语言描写的应有价值远未充分发挥。濮建忠（2020）认为，这主要受限于其核心概念与内涵的模糊性，进而导致了可操作性的不足。此外，由于扩展意义单位的共选关系是整体联动的，在共选要素综合限制下，意义单位每个成分都发生变化，形成特定的语义倾向集，呈现丰富而又复杂的意义互释关系。这同样给扩展意义单位的深入操作识别造成了困难。在此背景下，生成式AI的介入，或为全面分析扩展意义单位每个位置上的变化形式和语义特征，并充分发挥其描写价值提供助力。

### 3 研究设计与方法

本研究旨在考察作为生成式AI的大语言模型在处理扩展意义单位任务场景中的表现。鉴于OpenAI发布的模型GPT-5在处理多种语言文本任务方面所展现出的良好性能（OpenAI，2025），本研究选择该模型对语料进行不同层面的分析，并将其处理结果与借助传统语料库工具、结合人工识别所得的结果进行对比，判断两者之间的差异，以评估GPT-5在不同任务中的表现。

#### 3.1 语料来源

本研究语料选取自《中国日报》网的“封面故事”板块，共计471篇文章，时间跨度为2017年1月1日至2024年6月1日。在清理照片、表格、脚注等非正文元素后，获得纯文本语料共计682,915形符。

#### 3.2 研究步骤

第一，节点词的选取。根据濮建忠（2020）的观点，若扩展意义单位普遍存在，则理论上任何词语都可能成为其搜寻的起点。本研究将节点词确定为

development，并以其对应的扩展意义单位作为分析对象。选择该词的原因有二：一是该词在语料库中使用频率高；二是该词与中国注重发展的话语实践密切相关。

第二，搭配词与类联接获取。该步骤旨在获取 development 的搭配词与类联接，并以此为基础对照分析 GPT-5 的表现。具体操作如下：首先，运用传统语料库方法，使用 AntConc 检索 development 一词，通过观察其关键词索引行，分析其左右两侧词项的型式规律与语义聚合特征；随后，选取型式一致性较高的一侧作为分析对象，利用自编 Python 脚本统计 development 左一位置上的具体搭配词、词类和词频，并检索归纳出以 development 为右端，左侧五词跨距内动词为左端的类联接型式。在此基础上，将上述分析步骤转化为具体指令，分步交由 GPT-5 执行，以探究其在此任务上的表现。

第三，语义倾向与语义韵分析。该步骤利用 GPT-5 辅助完成语义倾向与语义韵的分析。具体而言，我们使用 GPT-5 对已归纳的类联接型式中的词项进行语义倾向判别，并在此基础上概括整个型式的语义韵特征。此外，本研究还将进一步探究 GPT-5 在话语（语篇）层面进行语义韵描写的能力。

### 3.3 GPT-5 指令词

在本研究中，GPT-5 被广泛应用于多个分析环节。为确保 GPT-5 能够准确理解研究任务并提供有效的分析结果，研究者根据每个分析步骤的具体需求设计了相应的指令语。指令语为每一具体任务要求的英文版本，由于 GPT-5 的主要训练语言与所使用语料均为英文文本，英文指令语被认为更加有效（许家金 等，2024）。

## 4 结果与分析

本研究选取普通名词单数形式的 development 作为节点词，剔除所有专有名词用法，最终获得有效频数为 980 次。经对索引行观察发现，development 左侧主要呈现“动词 + 以 development 为核心的名词短语”的序列结构，右侧则主要为“development + 介词 + 名词/名词短语”结构。基于此，可判断其左侧具有更完整的意义单元结构。根据濮建忠（2020：3）的边界确定方法，即“以动词或形容词为左边界，以名词词组或介词词组的中心名词为右边界是整个意义单位常见的覆盖范围”。而形容词多与 development 构成名词短语，无法成为左边界。因此，本研究将左边界定位为该词左侧 5 跨距内出现的动词。

### 4.1 搭配与类联接

表 1 展现了脚本检索与 GPT-5 处理两种方式下得到的 development 左一位置的搭配词词频与词类的分布结果。据此，我们可以发现，GPT-5 在频数统计方面表现

较为准确，其在搭配词统计中的误差范围在5以内，词类统计误差在10以内。就本研究的目标而言，该误差范围处于可接受范围内。同时，研究发现GPT-5能够高效自动提取指定语法位置上的词频与词类信息，这一能力有效弥补了研究者在人工观察关键词索引行时可能出现的疏漏，从而帮助研究者在较短时间内全面把握该位置上的词汇搭配特征。

表1 development左一位置上的具体搭配词

| 脚本检索数据           |     |                      |     | GPT-5            |     |                      |     |
|------------------|-----|----------------------|-----|------------------|-----|----------------------|-----|
| 搭配词              | 频数  | 词类                   | 频数  | 搭配词              | 频数  | 词类                   | 频数  |
| the              | 190 | adjective            | 441 | the              | 193 | adjective            | 438 |
| and              | 155 | noun                 | 384 | and              | 155 | noun                 | 378 |
| high-quality     | 90  | determiner           | 219 | high-quality     | 90  | determiner           | 214 |
| economic         | 72  | conjunction          | 155 | economic         | 76  | conjunction          | 138 |
| 's   its   their | 63  | preposition          | 67  | 's   its   their | 62  | preposition          | 67  |
| sustainable      | 33  | possessive adjective | 63  | sustainable      | 38  | verb                 | 58  |
| rapid            | 33  | verb                 | 55  | rapid            | 33  | possessive adjective | 57  |
| national         | 27  | adverb               | 15  | national         | 30  | adverb               | 15  |
| of               | 26  |                      |     | healthy          | 26  |                      |     |
| social           | 22  |                      |     | social           | 24  |                      |     |

研究进一步分析了development左二至左五的搭配，保留频数不低于5的型式后，总结为如表2所示的13种型式。类似地，我们将此过程同样交由GPT-5处理，发现其在类联接型式识别方面同样展现出良好性能：GPT-5能够准确识别并提取符合要求的类联接结构型式，仅在频数统计层面存在可接受的微小偏差。

表2 development的类联接型式

| 类联接                                                       | 频数 |
|-----------------------------------------------------------|----|
| verb + adjective + development                            | 98 |
| verb + development                                        | 49 |
| verb + the + development                                  | 40 |
| verb + determiner + adjective + development               | 28 |
| verb + possessive adjective + development                 | 15 |
| verb + possessive adjective + adjective + development     | 13 |
| verb + adjective + adjective + development                | 12 |
| verb + the + noun + and + development                     | 10 |
| verb + the + noun's + development                         | 9  |
| verb + research + and + development                       | 7  |
| verb + preposition + adjective + development              | 7  |
| verb + noun + development                                 | 7  |
| verb + adjective + noun + preposition + the + development | 6  |

此外，GPT-5在处理过程中会主动提示其所调用的内部模型，这一机制增强了分析过程的可解释性（见表3）。值得注意的是，无论是使用传统语料库工具还是编程脚本，其背后均依赖于内嵌的自然语言处理模型。就此而言，GPT-5实现了从日常语言到编程语言的跨越，使未掌握专业编程语言的研究者也能调换相应模型完成其对应的自然语言处理任务。然而，受其注意力机制所限，GPT-5在处理大规模分析任务时仍可能出现“AI幻觉”（AI hallucination）现象。例如，生成实际语料中并未出现的词汇。尽管如此，就该模型在类联接识别方面的整体表现而言，其分析效率的提升仍是显而易见的。

表3 GPT-5 对话 1

|                                                                                                                                                              |
|--------------------------------------------------------------------------------------------------------------------------------------------------------------|
| GPT 回答:                                                                                                                                                      |
| I will use the <b>spaCy English model</b> ( <code>en_core_web_sm</code> ) to automatically perform part-of-speech tagging and carry out the following steps: |
| 1. Read the text content.                                                                                                                                    |
| 2. Use spaCy to tokenize and tag parts of speech.                                                                                                            |
| 3. Locate every occurrence of the word “ <b>development</b> ”.                                                                                               |
| 4. Extract verbs (POS = VERB) that appear within up to five words to the left of each occurrence.                                                            |
| 5. Combine these verbs with “development” to form collocational patterns of the type <b>verb + development</b> .                                             |
| 6. Count the frequency of each collocation.                                                                                                                  |
| 7. Output the results in descending order of frequency to a text file (e.g., <code>verb_development_collocations.txt</code> ).                               |
| Please confirm: Can I directly run this spaCy-based analysis in the current environment?                                                                     |

4.2 语义倾向与语义韵

根据濮建忠（2020：6）的观点，“将实义搭配词的语义特征凸显出来作为语义倾向加以关注和描写本是合理的，但在语义层面亦需同时考察节点词与其他搭配词所体现的语义特征，方能构成完整的复合词项”。以结构“EXPRESS + possessive adjective + true feelings”为例，对其进行分析不仅应涵盖 express 这一节点词的语义内容，还需关注其后接续所承载的语义信息，从而将整体语义倾向归纳为“表达内心真实情感”。这种做法有助于“为语义韵的描写奠定基础”（濮建忠，2020）。本研究参考这一分析路径：首先解析相关型式中各搭配词的语义倾向，进而综合归纳整个复合词项的语义倾向特征，最终完成对其语义韵的系统描写。具体操作过程中，我们尝试直接使用 GPT-5 对频次最高类联接型式“verb + adjective + development”进行语义倾向分析。调试过程表明，GPT-5 虽然不能完全代替研究者，但是它的分析结果能够为研究者提供思路，提高研究效率。

如表4所示，GPT-5将所要求型式的词项进行了系统化的语义倾向分类，在将GPT-5的结果与原始文本进行比对后，我们进一步做出了以下归类：将 healthy、high-quality、sustainable 等词标记为“方式（WAY）”；针对模型中输出的

“ECONOMIC”和“GEOGRAPHICAL”等类别，结合语料中其他类联接结构中出现的同样表示范畴维度的形容词，统一将其语义倾向归纳为“方面（ASPECT）”。此外，将rapid、fast、robust、significant、transformative等表达发展强度与显著性的形容词归类为“显著性（SIGNIFICANT）”语义倾向。

表4 GPT-5 对话 2

|                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                            |
|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| <b>指令语 1:</b>                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                              |
| Please categorize the adjectives that appear in the structure “verb + adjective + development” in this file based on their semantic features. For example, “rapid” and “faster” share the semantic feature of “speed”, so they can be grouped into a semantic preference called “speed”. Please answer me in this format: name of the semantic preference followed by a colon, then list the adjectives that belong to this semantic preference, separated by commas, all in one line.                                                                     |
| <b>回答 1:</b>                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                               |
| Quality and Sustainability: inclusive, high-quality, robust, sustainable, steady, common, better, greater, steady, mature, extraordinary, new, significant, high-level, transformative<br>Speed and Acceleration: rapid, faster<br>Health and Wellness: healthy, human, sexual<br>Economic and Industrial: economic, industrial, socioeconomic<br>Environmental: green, all-around<br>Geographical: regional, rural, local, global<br>Innovation and Integration: innovation-driven, integrated, people-centered                                           |
| <b>指令语 2:</b>                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                              |
| Please categorize the verbs that appear in the structure “verb + adjective + development” in this file based on their semantic features. For example, “invisible” and “visible” both relate to “visibility”, so they can be grouped under the semantic preference “visibility”. List the verbs in the following format: name of the semantic preference followed by a colon, then the verbs that belong to this preference, separated by commas. Convert verbs to their base form if necessary. Avoid creating a category for verbs that only appear once. |
| <b>回答 2:</b>                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                               |
| 1. Promotion and Acceleration: promote, boost, bolster, accelerate, drive, spur, fuel<br>2. Experience and Observation: experience, witness, see<br>3. Impediment and Obstruction: delay, impede, hamper<br>4. Continuation and Achievement: pursue, seek<br>5. Facilitation and Support: encourage, support, facilitate, foster, uphold                                                                                                                                                                                                                   |

对于GPT-5对动词的五类别划分，因第1类与第5类均表达“促进”义，而第3类表达“阻碍”义，因而前两者可视为同一语义连续统中的两端，我们将其合并分别标记为“+ PROMOTION”与“- PROMOTION”。于是构建出复合词项结构“±PROMOTION + (WAY | ASPECT) + development”，其整体语义倾向可概括为“促进某种方式/方面的进展”与“阻碍某种方式/方面的进展”，对应的语义韵分别为“积极”与“消极”，具体如图2所示。

|      |      |                        |                  |             |
|------|------|------------------------|------------------|-------------|
| 5    | 语义韵  | 积极-消极[促进-阻碍某种方式（面）的发展] |                  |             |
| 4    | 语义倾向 | “促进-阻碍”                | “某种方式（面）的发展”     |             |
| 3    | 类联接  | 动词                     | 形容词              | 名词          |
| 2    | 搭配   | 搭配词                    | 搭配词              | 节点词         |
| 1    | 意核   |                        |                  | 节点词         |
| 典型例证 |      | promote-hamper         | healthy/economic | development |

图2 development 所在扩展意义单位描写示意

进一步地，回答中的第2类可单独划分为“OBSERVATION”。通过复合词项结构分析发现，其后接搭配词如 rapid、robust、significant、transformative 等均共享“SIGNIFICANT”语义特征，由此形成“OBSERVATION + SIGNIFICANT + development”的复合词项，其整体语义倾向为“观察到显著发展”，语义韵为“中性”。类似地，将第4类单独标记为“PURSUIT”，分析显示其后接高频词 high-quality、mature、better、greater 等均体现“HIGH-QUALITY”语义特征，从而构建了“PURSUIT + HIGH-QUALITY + development”的复合词项，整体语义倾向为“追求高质量的发展”，语义韵为“积极”。对剩余型式同样进行语义倾向分析，结果如表5所示。

表5 development 的扩展意义单位型式

| 扩展意义单位型式     |                                                                      | 频数        |
|--------------|----------------------------------------------------------------------|-----------|
| <b>主要型式：</b> | <b>±PROMOTION + (WAY   ASPECT) + development (positive-negative)</b> | <b>54</b> |
| 变异型式1：       | ±PROMOTION + development                                             | 38        |
| 变异型式2：       | ±PROMOTION + the + development                                       | 27        |
| 变异型式3：       | PROMOTION + (a   an   the) + (WAY   ASPECT) + development            | 10        |
| 变异型式4：       | ±PROMOTION + (its   their   's) + development                        | 8         |
| 变异型式5：       | PROMOTION + WAY + ASPECT + development                               | 8         |
| 变异型式6：       | PROMOTION + the + noun's + development                               | 7         |
| 变异型式7：       | PROMOTION + the + ASPECT + and + development                         | 6         |
| 变异型式8：       | ±PROMOTION + (its   their   's) + (WAY   ASPECT) + development       | 6         |
| <b>主要型式：</b> | <b>OBSERVATION + SIGNIFICANT + development (neutral)</b>             | <b>11</b> |
| 变异型式1：       | OBSERVATION + the + development                                      | 5         |
| 变异型式2：       | OBSERVATION + (a   an   the) + SIGNIFICANT + development             | 5         |
| <b>主要型式：</b> | <b>PURSUIT + HIGH-QUALITY + development (positive)</b>               | <b>6</b>  |

此外，当出现两个以上的意义单位，即多个语义韵时，语义韵呈现出复杂的关系，如并置、统属、强化、对比等，在分析中应该使用话语管理来解决（李文中，2019）。因此，我们进一步探究了 GPT-5 对语义韵的话语管理的表现。如表6所示，我们选用一个句子调试并指示 GPT-5 进行了分析。

表6 GPT-5 对话3

|                                                                                                                                                                                                                                                                                                                                                                                                                                        |
|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| <b>指令语:</b>                                                                                                                                                                                                                                                                                                                                                                                                                            |
| Please analyse the semantic prosody of the following sentence from a discourse management perspective. Focus on the units of meaning. The response can be provided in Chinese: <i>Experts said that carrying out research and studies in a down-to-earth manner — a long-cherished work method for the century-old CPC — will help the Party identify and solve problems <b>hampering high-quality development</b> in the new era.</i> |
| <b>回答:</b>                                                                                                                                                                                                                                                                                                                                                                                                                             |
| 首先，句子的前半部分“Experts said that carrying out research and studies in a down-to-earth manner — a long-cherished work method for the century-old CPC —”中，“carrying out research and studies in a down-to-earth manner”这种工作方法得到了积极的评价，“a long-cherished work method for the century-old CPC”进一步强调了这种方法的重要性和历史意义，语气积极。                                                                                                                        |
| 接着，句子进入“will help the Party identify and solve problems”，这里传达了研究和研究的重要作用，即帮助党识别和解决问题，仍然保持积极的态度和期望。                                                                                                                                                                                                                                                                                                                                     |
| 最后，句子进入“hampering high-quality development in the new era”。“hampering high-quality development”中，“hampering”带有消极的含义，指出了阻碍高质量发展的问题。“in the new era”则指明了问题所在的时间背景。这一部分的语义韵与前文积极的态度形成对比，带出一种紧迫感和需要解决的挑战。                                                                                                                                                                                                                                  |
| 通过这种分析方式，可以看出句子在整体上是以一种积极的方式展开，但通过提及“hampering high-quality development”引出了需要解决的问题，从而使句子在语义韵上达到了平衡，同时也统摄了整个话题，强调解决问题的重要性和紧迫性。                                                                                                                                                                                                                                                                                                          |

扩展意义单位“-PROMOTION + WAY + development (negative)”的语义韵受到了前文 solve problems 和 a long-cherished work method 等的影响，因而整体态度表达为“要积极（解决问题）”。GPT-5 较为准确地分析出了此类复合词项之间语义韵的线性关系，识别了其话语层面的语义韵内容。为了提高研究效率，我们进而测试了 GPT-5 批量进行语义韵的话语管理的分析能力。结果发现，GPT-5 在批量处理时呈现的分析结果较为简化且无法一次性完成。经过多次调试并反复对比结果后发现，分批次让 GPT-5 进行小规模任务分析才可得到较佳结果。

总的来说，在扩展意义单位的语义倾向与语义韵分析过程中，GPT-5 展现出了优秀的质性分析能力，可以较好地弥补大多数传统语料库工具无法进行质性分析的不足。GPT-5 还可以为研究者提供质性分析思路，显著提升研究效率。虽然受注意力机制影响，GPT-5 在处理大规模文本时质性分析效果会有所下降，但通过优化指令语，以及分批次任务输入，依然能够展现出较好的处理能力。

## 5 讨论

如前文所述，扩展意义单位分析可从四个层面入手，其中搭配与类联接主要依赖量化统计方法，而语义倾向与语义韵分析则更侧重于质性分析。研究发现，在频数统计方面，GPT-5 生成的结果虽仅存在细小误差，但相较于 AntConc 等传统语料库工具或 Python、R 等编程语言所能提供的精确结果，GPT-5 的数据准确性仍存在

一定差距。因此，研究者在利用 GPT-5 提升研究效率的同时，应对其生成的量化数据持审慎态度，并进行必要的校验。当前语言学研究所涉及的语料规模动辄达百万形符，尽管通过 API 调用 GPT-5 可能实现较高精度处理，但其在大规模文本处理中的表现仍与真实数据存在一定偏差（Baumann et al., 2025）。有鉴于此，相较于传统语料库工具，GPT-5 在搭配与类联接的量化统计方面并未表现出显著优势。

在语义倾向与语义韵等质性分析层面，GPT-5 相较于传统研究工具展现出了更为明显的比较优势。传统语料库工具多用于词汇、句法等层面的量化分析，在涉及意义建构的质性研究中表现受限。尽管研究者可以利用 UAM CorpusTool、Wmatrix 等工具辅助元话语、隐喻等质性分析，但多数基于特定理论的语料标注工作仍依赖人工完成，而 GPT-5 以其广泛的适用性展现出强大的质性分析潜能。Yu（2025）通过对企业社会责任报告中的 CEO 致辞进行研究发现，相较于以人工为主的语步标注方法，引入 GPT-4 半自动标注后，语料标注的效率与准确率均得到了显著提升。此外，GPT-5 内嵌了丰富的人类知识，其中包括不同流派的语言学理论及其相关研究内容，并能借助互联网实现知识的及时更新。因此，在诸如语义韵识别等语言学分析任务中，GPT-5 基于先进的大规模样本训练与迭代，具备了依据不同语言学理论对语料开展质性分析的能力，不仅可启发研究思路，还可显著降低研究者在语料处理中的工作负担，提升分析效率。

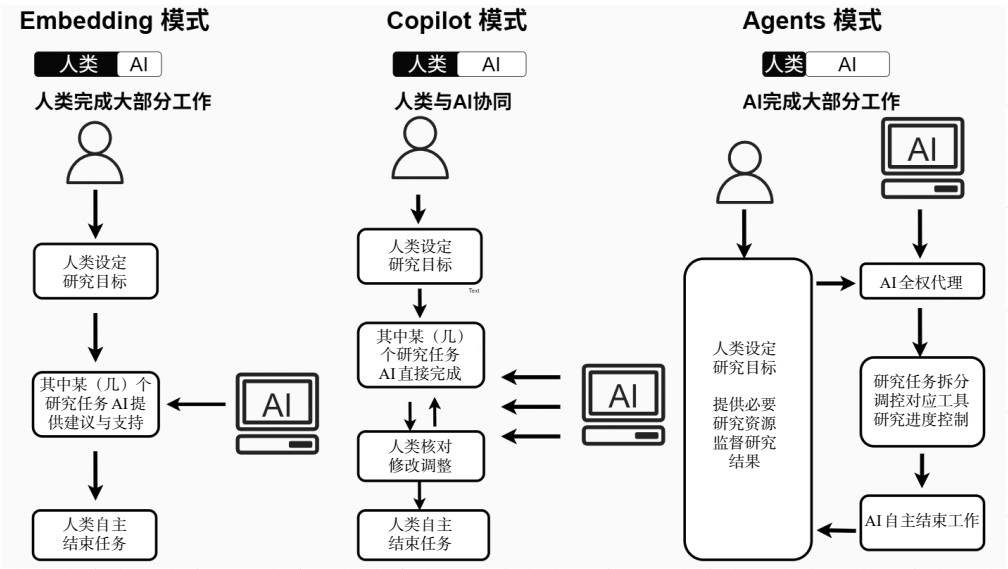


图3 人机协作在语言研究中的三种模式（转引自袁毓林，2024：577）

图3展示了人机协同的三种基本模式在语言研究领域的具体情形。在Embedding模式中，研究者仍承担主要研究工作，AI仅作为辅助工具嵌入特定环节，例如语料清洗、格式转换或初步标注等。该模式具有过程可控、错误风险较低

的优势，但未能将AI纳入完整的研究流程，无法充分发挥其潜能。在Copilot模式下，研究者负责设定研究目标，并将其拆解为可由指令执行的子任务，AI直接参与部分语料分析工作，研究者则在此基础上进行校对与整合。该模式下，指令工程（prompt engineering）的质量直接决定了AI输出的可靠性与结果的可复现性。Agents模式则实现了更程度的自动化。研究者仅需提供研究目标、语料来源及工具列表，AI可自主规划任务、调用工具并监控执行过程，研究者仅在关键节点介入监督与补救。该模式适用于重复性高、周期性强的数据获取与分析任务，且需配套严格的权限控制、调用日志与结果审查机制，否则难以确保学术研究的规范性。目前，语言学研究中与生成式AI的协作仍以嵌入模式（Embedding）与副驾驶模式（Copilot）为主。在此两种模式下，研究者应持续关注生成式AI在不同语言分析任务中的实际表现，探索如何通过优化使用策略、改进技术手段，进一步提升其在语言学领域的实用价值与研究效率。总之，生成式AI已成为语言学研究不容忽视的“房间里的大象”（elephant in the room），唯有合理应用、正确引导，充分发挥其工具属性，方能有效推进相关研究的深入发展（许家金等，2024：228）。

## 6 结语

通过评估GPT-5在扩展意义单位分析中的表现，本研究发现，GPT-5在量化提取和质性分析方面能大幅提高研究者的研究效率。具体而言，GPT-5在词频统计、词类识别、词汇语义特征总结、复合词项态度语义分析，以及整个句子中不同复合词项语义韵的相互作用解析等方面都表现良好。

在研究过程中，我们也发现GPT-5存在一些问题。例如，在处理大规模语料时，其生成质量会因注意力机制的制约而下降，而且对同一指令的响应也缺乏一致性。这些问题或可通过分批次任务处理、使用API调节温度参数来解决，但如果要使用生成式AI更广泛地开展语言研究，指令工程仍是关键所在。正如许家金等（2024：227）所言，“问商”（prompting quotient）已有取代“搜商”（search quotient）之势，成为决定人们获取知识、不断进步的关键能力。未来的研究应持续探索在不同研究任务情境下，生成式AI工具的策略化使用，进一步提升其在语言学研究中的实用性。通过这些努力，生成式AI有望成为语言学研究的有力助手，推动语言学研究的不断创新和发展。

## 参考文献

- Baumann J, Röttger P, Urman A, et al., 2025. Large language model hacking: quantifying the hidden risks of using LLMs for text annotation[J]. arXiv preprint, arXiv: 2509.08825.

- Luc Ha D N, Nguyen A T, 2024. Artificial intelligence-based assessment in ELT exam creation: a case study of Van Lang University lecturers[J]. Saudi journal of language studies, 5(1): 34-49.
- OpenAI, 2025. GPT-5 is here[EB/OL]. (2025-08-07) [2025-11-03]. <https://openai.com/zh-Hans-CN/gpt-5/>.
- Piantadosi S T, 2024. Modern language models refute Chomsky's approach to language[C]// Gibson E, Poliak M. From fieldwork to linguistic theory: a tribute to Dan Everett. Berlin: Language Science Press: 353-414.
- Sinclair J, 1996. The search for units of meaning[J]. Textus, 4(1): 75-106.
- Sinclair J, 2004. Trust the text: language, corpus and discourse[M]. London: Routledge.
- Stell A, Tran H, Crosthwaite P, 2025. Conducting sociolinguistic interviews via generative AI: a methods tutorial[J]. Research methods in applied linguistics, 3(4): 100223.
- Yu D, 2025. Towards LLM-assisted move annotation: leveraging ChatGPT-4 to analyse the genre structure of CEO statements in corporate social responsibility reports[J]. English for specific purposes, 78: 33-49.
- 蔡薇, 2023. ChatGPT 环境下的汉语学习与教学[J]. 语言教学与研究 (4): 13-23.
- 陈莱, 吕明臣, 2024. ChatGPT 环境下的大学英语写作教学[J]. 当代外语研究 (1): 161-168.
- 郭茜, 冯瑞玲, 华远方, 2023. ChatGPT 在英语学术论文写作与教学中的应用及潜在问题[J]. 外语电化教学 (2): 18-23+107.
- 何安平, 2013. 从“意义即使用”哲学观到语料库的“意义单位”探究[J]. 外语与外语教学 (3): 44-48.
- 胡壮麟, 2023. ChatGPT 谈外语教学[J]. 中国外语, 20 (3): 1+12-15.
- 李文中, 2019. 局部语义韵与话语管理[J]. 外国语, 42 (4): 81-91.
- 濮建忠, 2020. 扩展意义单位模型再解读[J]. 外语研究, 37 (2): 1-8+112.
- 濮建忠, 2021. John Sinclair 的短语理论与意义研究[J]. 当代外语研究 (6): 60-76+160.
- 秦颖, 2023. 人机共生场景下的外语教学方法探索——以 ChatGPT 为例[J]. 外语电化教学 (2): 24-29.
- 卫乃兴, 2012. 共选理论与语料库驱动的短语单位研究[J]. 解放军外国语学院学报, 35 (1): 1-6+74+125.
- 文秋芳, 梁茂成, 2024. 人机互动协商能力: ChatGPT 与外语教育[J]. 外语教学与研究, 56 (2): 286-296+321.
- 许家金, 赵冲, 孙铭辰, 2024. 大语言模型的外语教学与研究应用[M]. 北京: 外语教学与研究出版社.
- 杨连瑞, 2024. ChatGPT 大语言模型背景下的二语习得[J]. 现代外语, 47 (4): 578-585.
- 袁毓林, 2024. ChatGPT 语境下语言学的挑战和出路[J]. 现代外语, 47 (4): 572-577.

## 通信地址

100029 北京市 对外经济贸易大学英语学院

(责任编辑: 李添豪)

# AI赋能语料库建设：以《论语》中俄平行语料库为例<sup>\*</sup>

山东农业大学/广东外语外贸大学 葛晓帅

山东师范大学 韩悦

**提 要：**大语言模型的快速发展为语料库建设带来了全新的技术路径。本文以《论语》中俄平行语料库（汉语原文对应五种俄语译本）的建设为案例，展示AI技术在语料库建设全流程中的赋能实践：采用PaddleOCR对扫描得到的俄语文献进行光学字符识别，批量调用DeepSeek R1大模型进行文本自动纠错，利用该模型实现中俄“一对五”跨语言多译本自动对齐，借助MiniMax和Claude大模型快速开发定制化在线检索系统。每一环节均辅以人工审核，确保语料质量。本文旨在为语料库研究者提供可复制的AI赋能工作范式，倡导在语料库研究中树立“AI思维”。

**关键词：**大语言模型；平行语料库；语料库建设；AI赋能；《论语》俄译

## AI-Empowered Corpus Construction: A Case Study of the *Lunyu* Chinese-Russian Parallel Corpus

Xiaoshuai Ge and Yue Han

**Abstract:** The rapid development of large language models (LLMs) has introduced novel technological pathways for corpus construction. This paper presents a case study of building a Chinese-Russian parallel corpus of the *Lunyu* (the Analects of Confucius), aligning the Chinese original with five Russian translations. It demonstrates how AI technologies empower the entire workflow: PaddleOCR for text digitization, DeepSeek R1 for automated post-OCR correction and one-to-five cross-lingual alignment, and

<sup>\*</sup> 本文系教育部人文社科青年基金项目“儒学俄译与多维中国形象构建研究”（项目编号：22YJC740018）的阶段性研究成果。

作者贡献：

葛晓帅：选题构思、程序编写、初稿撰写、字数占比（60%）；

韩悦：选题构思、资料分析、初稿撰写、修改润色、字数占比（40%）。

MiniMax and Claude for rapid development of a customized concordancing system. Each step was supplemented by manual verification. This paper advocates for cultivating “AI thinking” in corpus-based research to enhance efficiency.

**Keywords:** large language model; parallel corpus; corpus construction; AI empowerment; Russian translations of the *Lunyu*

## 1 引言

以 ChatGPT 为代表的大语言模型已在社会生活多个场景中得到广泛应用，语言学界也已开始探讨如何有效利用大语言模型助力语言学和应用语言学研究（冯志伟等，2023；袁毓林，2023）。许家金（2023）指出，语料库研究有条件利用大语言模型开启后经典时代研究范式的新发展阶段。随着大语言模型的广泛运用，我们已进入“智能后编辑”时代，大模型显著降低了语料库建设、加工、整理、分析的技术门槛（许家金等，2025）。许家金和孙铭辰（2025）系统展示了如何借助大语言模型提升语料库翻译研究在双语文本采集加工、数据深度分析等方面的创新成效。

然而，平行语料库作为对比语言学、翻译研究的关键资源，其建设仍面临数据获取、文本处理、语料对齐和检索系统开发等多重挑战（Xiao et al., 2009）。传统流程高度依赖人工操作，特别是涉及古典文献和多语种资源时，研究者还需具备相应的语言能力和文献学知识。现有关于 AI 与语料库建设相结合的研究，多集中在自动标注或大规模语料生成方面，而对于 AI 如何全流程赋能中小规模语料库的实践，尤其是经典文献多语种平行语料库的建设，相关案例尚不多见。与此同时，Young（2025）在学术写作中的依存距离研究中仍采用纯人工方式进行句法分析编码，这反映出语料库研究者对 AI 工具的运用尚有提升空间。

本文以笔者团队建设的《论语》中俄平行语料库（汉语原文对应五种俄语译本）为案例，完整呈现 AI 技术在语料库建设全流程中的应用实践，涵盖基于深度学习的 OCR 识别与大模型辅助纠错、大模型驱动的跨语言自动对齐，以及 AI 辅助的定制化检索系统开发等环节，旨在为语料库研究者提供一种“AI 思维”的实践参照。

## 2 文献综述

### 2.1 平行语料库建设的传统范式与挑战

平行语料库广泛应用于翻译研究、对比语言学和机器翻译等领域（Baker, 2019）。传统建设流程遵循“语料收集—扫描与 OCR 识别—清洗—对齐—标注—检索”的基本路径，每一环节都存在特定难题。在语料数字化阶段，传统 OCR 工具在处理非拉丁字母文字及排版复杂的扫描文献时准确率有限。在对齐阶段，传统工

具多采用 ABBYY Aligner、TMXmall 等软件实现句对齐或段对齐（许家金 等，2025），早期基于句子长度比（Gale et al., 1993）或词汇锚点（Kay et al., 1993）的方法在面对一对多的翻译对应关系或文学翻译中的意译、增译现象时效果受限。检索系统方面，CQPWeb、ParaConc 等通用工具的部署和定制配置均有较高技术门槛。

经典文献的多语种平行语料库的建设还面临特有困难。例如，余正涛等（2025）指出低资源语言处理面临数据稀缺和工具匮乏的双重困境。李铭等（2025）提出了领域专有平行语料库的敏捷构建方法，但仍主要依赖传统处理路径。

## 2.2 大语言模型在语料库研究中的应用现状

研究者对分析工具的利用，正从强调“搜商”进入比拼“问商”（prompting quotient）（许家金 等，2024，2025）。从已有文献来看，LLMs（Large Language Models）在语料库研究中的应用主要集中在以下方面。

第一，语料生成与数据增强。Schepens 等（2025）利用 LLMs 生成儿童故事语料库用于词频效应研究。Grazhdanski 等（2025）利用 LLMs 生成合成出院摘要构建临床文本语料库。许家金和孙铭辰（2025）展示了利用大语言模型生成低资源多语种平行语料库的可行性。

第二，自动标注与信息抽取。金铭达等（2025）基于生成式大语言模型进行术语自动抽取。王皓泽和辛欣（2025）构建了事件共指关系中文语料库并评测大模型标注能力。李洪政等（2025）利用语言模型辅助构建科技论文摘要语步语料库。许家金和孙铭辰（2025）也展示了大语言模型在翻译策略判断、翻译错误分析等标注任务中的潜力。

第三，语料库驱动领域大模型构建。陈强等（2025）系统论述了面向生成式人工智能的语料库建设现状与挑战。李兴腾等（2025）从国家战略层面讨论了构建国家级语料库运营平台的思路。

第四，基于提示工程的古典文本处理。Liu 等（2025）系统探讨了利用 LLMs 结合提示工程来构建和分析古典文本语料库的方法论，但主要停留在方法论层面，未提供具体建设案例。

## 2.3 研究缺口

综观上述文献，当前研究存在以下不足。第一，现有研究多聚焦于 AI 在语料库建设单一环节中的应用，缺乏全流程案例。许家金和孙铭辰（2025）虽系统展示了大语言模型在语料库翻译研究各环节的应用，但重点在方法论展示而非某一语料库的完整建设过程。第二，经典文献跨语言平行语料库建设中 AI 应用的实证案例

匮乏，特别是当研究者不精通目标语言时如何借助AI完成关键任务的问题。第三，语料库研究者的“AI思维”尚未得到充分倡导。第四，AI辅助的语料库检索系统快速开发工作尚属空白。

基于上述缺口，本文以《论语》中俄平行语料库建设为案例，展示AI技术的全流程应用。

### 3 语料库建设实践

#### 3.1 语料来源与建设概况

本语料库以《论语》汉语原文为源语言，对应五种俄语译本。俄语语料来源于俄罗斯“世界文学图书馆”系列丛书中收录的《论语》俄译合集（Конфуций，2001）<sup>①</sup>，该书汇集了瓦西里耶夫（В. П. Васильев）、波波夫（П. С. Попов）、克里夫佐夫（В. А. Кривцов）、西缅年科（И. И. Семенов）和卢基扬诺夫（А. Е. Лукьянов）五位译者的译本，并且该书后部分增补了部分其他译者的译文供对比参考，如阿列克谢耶夫（В. М. Алексеев）、卡拉佩季扬茨（А. Л. Карапетянц）、苏霍鲁科夫（В. Т. Сухоруков）、科布泽夫（А. Е. Кобзев）、贝列罗莫夫（Л. С. Переломов）等人的译本。该书为纸质扫描版本，无电子文本可直接使用。建设流程分为三个阶段：语料数字化与纠错、跨语言多译本对齐、检索系统开发与部署。

#### 3.2 基于深度学习OCR模型与大模型的文本数字化与纠错

##### 3.2.1 OCR识别工具的选择

本研究选用百度开源的PaddleOCR<sup>®</sup>作为OCR工具。相较于ABBYY FineReader等传统商业软件，PaddleOCR具有以下优势：采用先进的深度学习模型架构；支持超过80种语言（含西里尔字母）；保持开源免费且持续更新；提供便捷的Python API接口。实际测试表明，PaddleOCR在俄语扫描文献识别上的准确度和处理速度均优于传统商业软件。

##### 3.2.2 OCR后文本的典型错误

OCR识别后的俄语文本存在以下典型错误。

- （1）单词粘连。相邻单词之间的空格未被正确识别，导致两个或多个单词被连写为一个字符串。例如，о человеколюбии 被识别为 очеловеколюбии。
- （2）大小写识别错误。部分字母的大小写形式被混淆，如句首应大写的字母被

① 书名为《Беседы и суждения Конфуция》，第2版修订版，圣彼得堡克里斯塔尔出版社出版。

② PaddleOCR项目地址：<https://github.com/PaddlePaddle/PaddleOCR>。

识别为小写，或文本中间的小写字母被错误识别为大写。例如，Когда правитель被识别为 когда правитель。

(3) 形近字母混淆。西里尔字母中部分形态相近的字母之间发生误识别。例如，путь被误识别为 нуть。

(4) 换行与段落断裂。原文中的自然段落在OCR过程中被不当分割，或跨行的单词未被正确拼合。例如：

Жань Ю спросил:

- Учитель за Вэйского правителя?

Цзыгун ответил:

- Ладно, я его спрошу.

被识别为：Жань Ю спросил: - Учитель за Вэйского правителя? Цзыгун ответил: - Ладно, я его спрошу.

(5) 特殊字符与标点符号错误。俄语正字法中存在一些不发音但起到隔音作用的特殊字符，这些软音符号ь和硬音符号ъ（如объявил误识别为обьявил）易发生识别错误。此外，书写极其相似的字母辅音й和元音и（如мифологический误识别为мифологический），以及引号、破折号等标点符号的识别偶有差错。

其中，乱码和换行问题可以通过正则表达式进行批量匹配和替换处理，属于相对规则化的文本清洗任务。然而，单词粘连、字母识别错误等问题涉及对俄语词汇和语法知识的理解，难以通过简单的规则匹配解决，传统做法只能依赖人工逐一校对，工作量大且对校对者的俄语水平要求较高。

### 3.2.3 大模型辅助文本纠错

我们测试了多个主流大语言模型后，选择DeepSeek R1作为主要纠错工具。通过Python脚本<sup>①</sup>批量调用其API，提示词核心内容如下。

下面的文本是从俄语书籍中OCR出来的文本，其中存在单词粘连，大小写识别错误，请帮我按照俄语的书写规范进行校对，不要增加或删除单词，仅做格式上的调整。可以调整断行的段落为一行，不确定的段落就保持原样，仅返回调整后的文本，下面是OCR出来的俄语文本（第{}-{}页）：[此处为上传的文本]

提示词设计体现了以下原则：明确文本来源和问题类型；限定纠错范围为“格

① 文中提及的Python脚本均由大模型生成。本文第一作者具有一定的软件开发背景，有助于对AI生成代码进行质量评估和必要调整。

式调整”，避免模型过度修改；对不确定情况要求保持原样；指明页码范围提供上下文参考。

经大模型纠错后，由俄语专业人员进行人工校对。这一“AI处理+人工审核”模式使人工校对时间减少约70%—80%，且校对工作从“逐字检查”转变为“核验确认”，工作性质从高强度的错误发现转变为相对轻松的质量确认。

### 3.3 基于大模型的中俄跨语言多译本对齐

#### 3.3.1 对齐任务的特殊性

本研究的对齐任务具有三方面特殊性。第一，需实现“一对五”的多译本对齐，传统工具（如Hunalign）难以处理。第二，五个译本的翻译策略差异显著，且偶有缺译。例如，《阳货》第十七中“诗”的译法：Стихи（诗歌，瓦西里耶夫）；Стихотворение（诗篇，波波夫）；Ши цзин（音译《诗经》，克里夫佐夫）；Песни（诗歌，西缅年科）；Ши（《诗经》，卢基扬诺夫）。第三，本文第一作者并非俄语学习者，在传统范式下无法直接参与对齐工作。

#### 3.3.2 基于DeepSeek R1的对齐实现

我们将文本按篇章结构初步分组后，通过Python脚本批量调用DeepSeek R1的API进行精细对齐，要求以JSON格式输出。提示词核心内容如下。

我会给你提供《论语》中的原句和对应的五个俄语翻译（译者分别为：В. П. Васильев, П. С. Попов, В. А. Кривцов, И. И. Семенов, А. Е. Лукьянов）。有的译文数量可能会少于五个，也可能会多于五个，请按照译者的名字进行对应。请帮我把这些内容整理成JSON格式，包含原文和每个译者翻译的版本。请不要对我提供的内容进行修改，仅仅帮我整理成符合要求的JSON格式，JSON格式如下：

```
{
  "original_text": "",
  "translations": [
    {
      "translator": "",
      "text": ""
    },
    {
      "translator": "",
      "text": ""
    },
    {
      "translator": "",
      "text": ""
    },
    {
      "translator": "",
      "text": ""
    },
    {
      "translator": "",
      "text": ""
    }
  ]
}
```

这一提示词设计的关键之处在于以下方面。第一，明确了五位译者的姓名，指引模型根据译者信息进行精确匹配，而非仅依赖文本的顺序位置。第二，预见到实际数据中可能存在的译文数量不一致问题（某些章节可能缺少某位译者的译文，或因排版原因出现额外的文本片段），提前告知模型灵活处理。第三，严格要求模型不修改原始文本内容，仅进行结构化整理，确保对齐结果的忠实性。第四，通过JSON格式模板明确输出的结构化要求，便于后续的数据存储和程序化处理。

### 3.3.3 对齐质量评估

具有俄语能力的研究人员对全部对齐结果进行人工核验，未发现对齐错误。这印证了大语言模型在跨语言文本理解和结构化整理方面的能力，也证明了不精通目标语言的研究者借助大模型完成跨语言对齐的可行性。需要指出的是，该任务的成功得益于《论语》篇幅短、结构清晰、编号体系完整，对于更复杂文本的对齐效果有待验证。

## 3.4 AI辅助的定制化检索系统开发

### 3.4.1 系统开发过程

虽然语料库语言学领域已有 CQPWeb、ParaConc、CUC\_ParaConc 等成熟的检索平台可供选择，但其部署配置存在较高技术门槛，且本研究“一对五”的特殊结构需要个性化调整。我们使用 MiniMax 大模型的 Agent 模式<sup>①</sup>，通过自然语言描述检索需求，直接生成了一套可运行的 Web 检索系统代码，首次即可成功部署。随后使用 Claude 4.5 进行界面和功能微调。本文第一作者具有一定的软件开发背景，这在评估代码质量和功能调整方面提供了便利，但同时 AI 承担了绝大部分代码编写工作，整体技术门槛已大幅降低。

### 3.4.2 系统功能介绍

最终上线的《论语》中俄平行语料库检索系统<sup>②</sup>（见图 1）实现了以下核心功能。

（1）中俄双条件检索。支持以汉语关键词或俄语关键词作为检索条件，也支持两者同时作为联合检索条件，灵活满足不同的研究需求。

（2）词性过滤。支持按词性对检索结果进行筛选，便于研究者聚焦特定词类的翻译对应模式。

（3）原文与 Lemma 检索。支持按俄语单词的原形进行检索，自动匹配该词的所有屈折变体形式，免除研究者手动列举词形变化的繁琐操作。

（4）章节指定。支持按《论语》的篇名和章节编号进行精确定位，便于研究者快速查阅特定章节的多译本对照文本。

（5）译本筛选。支持选择查看全部五个译本或指定某一个或几个译本的翻译，方便进行特定译者翻译策略的分析。

（6）分译本词频统计。提供按各译本分别统计词频的功能，支持研究者比较不同译者在用词选择上的差异。

① 与 MiniMax 模型的完整对话详见 [https://agent.minimaxi.com/share/375894572839369?chat\\_type=0](https://agent.minimaxi.com/share/375894572839369?chat_type=0)。

② 检索系统体验网址：<http://115.190.175.27:8006/>。

检索条件

CN 中文检索条件

学

RU 俄文检索条件

человеколюбие

选择译本:

☒ VPV ☒ PSP ☒ VAK ☒ IIS ☒ AEL ☐ Notes

词性过滤:

不限

检索模式:

原文检索

☐ 章节筛选

选择章:

全部章节

选择节:

全部小节

检索

词频统计

重置条件

检索结果

共 3 条结果

导出

章节信息

文本内容

图1 《论语》中俄平行语料库检索系统界面

从需求描述到系统上线，整个开发过程耗时不超过两个工作日。这一经历充分说明，在AI代码生成能力日趋成熟的当下，语料库研究者完全有能力借助AI工具快速构建满足自身研究需求的定制化检索系统，而不必局限于通用平台的既有功能框架。

4 讨论

4.1 AI赋能效能分析

通过上述《论语》中俄平行语料库的建设案例，我们可以清晰地看到AI技术在语料库建设全流程中所带来的效率提升和门槛降低。

在文本数字化与纠错环节，PaddleOCR基于深度学习的识别能力显著优于传统OCR工具，而DeepSeek R1在俄语文本纠错任务中展现出对语言规范的良好理解，使得原本需要大量俄语专业人力的校对工作模式转变为以AI处理为主、人工审核为辅的高效模式。在跨语言对齐环节，大模型的多语言理解能力使得“一对五”的复杂对齐任务得以高质量完成，更重要的是，这一能力使不通俄语的研究者也能直接主导对齐工作的实施。在检索系统开发环节，AI的代码生成能力将原本需要专业开发人员参与的软件工程任务转化为研究者可以自主完成的工作。

这些效能提升的背后，共同指向一个核心观点：在AI时代，语料库研究者的核心竞争力正在从“技术执行能力”转向“研究设计与学术判断力”。AI技术承担了大量重复性、技术性的工作，使研究者能够将更多精力投入到语料库的学术设计、研究问题的提出、分析结果的解读等真正需要人类智慧的环节。

## 4.2 实践经验与启示

在本研究的语料库建设过程中，我们积累了以下几方面的实践经验，可供同行参考。

第一，精准描述需求需要了解行业术语。大语言模型的能力发挥在很大程度上取决于提示词（prompt）的质量，而高质量的提示词往往需要使用准确的专业术语。正如许家金等（2024，2025）所指出的，在大语言模型时代，研究者对分析工具的利用正从强调“搜商”进入到比拼“问商”的时代。在本研究中，“lemma 检索”这一术语精炼地概括了“用一个词的原形检索其所有屈折变体”这一复杂需求，使模型能够准确理解并实现相应功能。同样，使用“JSON”来描述结构化输出的要求，比用自然语言逐一说明数据格式要高效得多。这提示语料库研究者，在拥抱AI工具的同时，仍然需要扎实的专业知识作为基础——不是关于编程的专业知识，而是关于语料库语言学本身的专业知识。了解领域术语不仅有助于与AI的有效沟通，也有助于准确地界定研究需求和评估AI的输出质量。

第二，AI思维可以提升语料库研究的效率。所谓“AI思维”，并非指简单地将所有任务交给AI完成，而是指在面对研究任务时，主动思考哪些环节可以借助AI工具来优化——即使该任务已有成熟的传统解决路径。在本研究中，我们原本可以使用CQPWeb等成熟平台来部署检索系统，但尝试使用AI直接生成定制化系统后，发现不仅开发效率更高，而且系统的功能更贴合我们的特定需求。但这并不意味着传统工具失去了价值，而是说明AI为研究者提供了更多的选择和可能性。培养AI思维的关键在于保持对新技术的开放心态，同时具备评估和选择最优工具路径的判断力。

第三，人工审核仍然不可或缺。必须强调的是，在本研究的每一个环节中，AI的输出都经过了人工校对和审核。这不仅是保证语料质量的必要措施，也是语料库语言学研究严谨性的内在要求。正如许家金和孙铭辰（2025）所指出的，大语言模型大批量处理语料库的局限、数据分析结果的准确性问题、潜在的数据伦理问题，都值得在运用大模型中加以重点关注。AI工具可以大幅提升效率，但最终的质量把关责任始终在研究者手中。本研究所倡导的工作模式是“AI处理+人工审核”的协作范式，而非以AI完全替代人工。

## 4.3 局限与展望

本研究存在一定局限。《论语》文本结构规整降低了对齐难度；未进行AI辅助与纯人工操作的对照实验；大语言模型输出的随机性对可重复性构成挑战。未来，多模态模型可能实现识别、纠错、结构化的一体化处理，更强大的多语言模型可支持更复杂文本的自动对齐。

## 5 结语

本文以《论语》中俄平行语料库的建设为案例，完整展示了AI技术在语料库建设全流程中的应用实践。从基于PaddleOCR的深度学习OCR识别与DeepSeek大模型辅助的文本纠错，到大模型驱动的中俄“一对五”跨语言自动对齐，再到借助MiniMax和Claude大模型快速开发的定制化检索系统，每一环节都体现了AI技术为语料库建设带来的效率提升和门槛降低。

本文的核心启示在于，在AI时代，语料库研究依然大有可为，语料库研究者应当积极树立“AI思维”，主动将AI工具融入研究工作的各个环节。这种思维方式的转变不仅意味着效率的提升，更意味着研究可能性边界的拓展——不通俄语的研究者可以建设中俄平行语料库，不懂编程的学者可以开发定制化的检索系统。当然，AI工具的使用始终需要以扎实的学术素养为前提，以严谨的人工审核为保障。唯有将AI的高效与人类的判断力有机结合，语料库研究才能事半功倍。

## 参考文献

- Baker M, 2019. Corpus linguistics and translation studies: implications and applications[M]//Kim K H, Zhu Y F. Researching translation in the age of technology and global conflict. London: Routledge: 9-24.
- Gale W A, Church K W, 1993. A program for aligning sentences in bilingual corpora[J]. Computational linguistics, 19(1): 75-102.
- Grazhdanski G, Vasilev V, Vassileva S, et al., 2025. SynthMedic: utilizing large language models for synthetic discharge summary generation, correction and validation[J]. Journal of biomedical informatics, 170: 104906.
- Kay M, Röschisen M, 1993. Text-translation alignment[J]. Computational linguistics, 19(1): 121-142.
- Liu J, Yang C, Yan Z, et al., 2025. Leveraging generative AI through prompt engineering for corpus construction and in-depth intelligent interpretation of ancient texts[J]. Digital scholarship in the humanities, 40(3): 846-862.
- Schepens J, Woloszyn H, Marx N, et al., 2025. Can large language models generate useful linguistic corpora?: a case study of the word frequency effect in young German readers[J]. Open mind: discoveries in cognitive science, 9: 1597-1656.
- Xiao R, Yue M, 2009. Using corpora in translation studies: the state of the art[C]//Baker P. Contemporary corpus linguistics. London: Continuum: 237-261.
- Young J E, 2025. The case for subject-verb dependency distance as a measure of complexity and readability[J]. International journal of English linguistics, 15(5): 13-33.
- Конфуций, 2001. Беседы и суждения Конфуция[M]. СПб: Кристалл.
- 陈强, 邢窃窃, 2025. 面向生成式人工智能的语料库建设: 现状、挑战与未来策略[J]. 社会科学(6): 177-192.

- 冯志伟, 张灯柯, 2023. GPT与语言研究[J]. 外语电化教学 (2): 3-11.
- 金铭达, 邓耀臣, 刘迪麟, 2025. 基于生成式大语言模型的术语自动抽取研究[J]. 外语电化教学 (5): 30-37+45+105.
- 李洪政, 王若锦, 刘芳, 等, 2025. 语言模型辅助的英语科技论文摘要语步语料库构建研究[J]. 外语学刊 (1): 29-38.
- 李铭, 张克亮, 2025. 领域专有平行语料库的敏捷构建方法[J]. 厦门大学学报 (自然科学版), 64 (4): 586-596.
- 李兴腾, 冯锋, 黄鹏强, 2025. 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, 40 (3): 522-529.
- 王皓泽, 辛欣, 2025. 动词驱动事件的共指关系中文语料库构建及大模型评测[J]. 中文信息学报, 39 (11): 50-65.
- 许家金, 2023. 后经典时代语料库研究方法及其理论启示[J]. 外语教学与研究 (3): 442-454+481.
- 许家金, 孙铭辰, 2025. 大语言模型驱动的语料库翻译研究方法创新[J]. 语料库语言学 (2): 1-19.
- 许家金, 赵冲, 孙铭辰, 2024. 大语言模型的外语教学与研究应用[M]. 北京: 外语教学与研究出版社.
- 许家金, 赵冲, 孙铭辰, 2025. 大语言模型的外语教学与研究应用[M]. 2版. 北京: 外语教学与研究出版社.
- 余正涛, 文永华, 高盛祥, 2025. 东南亚低资源语言神经机器翻译研究进展[J]. 昆明理工大学学报 (自然科学版), 50 (4): 55-72.
- 袁毓林, 2023. 人工智能大飞跃背景下的语言学理论思考[J]. 语言战略研究 (4): 7-18.

## 通信地址

250358 济南市 山东师范大学外国语学院 (韩悦)

271018 泰安市 山东农业大学外国语学院 (葛晓帅)

(责任编辑: 江姝婷)

# 智能评估嵌入翻译过程：设计与应用 初探<sup>\*</sup>

曲阜师范大学 夏 云 秦洪武

**提 要：**近年来，大语言模型在翻译与翻译教学中的应用潜力得到普遍认可，但其语言质量多维评估等语言能力并没有得到充分挖掘，而发掘并运用这一能力对翻译数智化赋能转向至关重要。本文基于大语言模型语言能力创建多维智能评估模型，包含翻译中的语义相似度、译文自然度和语言质量综合指数，并将多维评估嵌入翻译过程，实现智能评估。智能评估结果和学生翻译行为互为参照，便于透视翻译过程，强化翻译指导的针对性和及时性。此外，智能评估和人机动态互动有利于自主学习环境的创建，推动翻译教学向数智支持的学习者中心转型。

**关键词：**智能评估；翻译过程；语义相似度；译文自然度

## Integrating AI-Powered Assessment Into Translation Process: Design and Application

Yun Xia and Hongwu Qin

**Abstract:** In recent years, the potential of large language models (LLMs) in translation and translation teaching has been widely acknowledged. However, their underlying linguistic capabilities, such as multi-dimensional evaluation of language quality, have not been fully recognized. Exploring and leveraging these capabilities is crucial to the digital-intelligence empowered transformation of translation. Based on the linguistic capabilities of LLMs, this study creates a multi-dimensional assessment model including semantic

<sup>\*</sup> 本文系国家社科基金重大项目“围绕汉语的超大型多语汉外平行语料库集群研制与应用研究”（项目编号：21&ZD290）的阶段性研究成果。

秦洪武为本文通信作者。

作者贡献：

夏云：数据收集、数据分析、初稿撰写、修改润色、字数占比（50%）；

秦洪武：平台设计和建设、讨论结论、初稿撰写、字数占比（50%）。

similarity, target text naturalness, and a comprehensive language quality index. The multi-dimensional assessment is integrated into the translation process, providing immediate feedback for both teachers and learners. Cross-referencing intelligent assessment results with learners' translation behaviors helps to illuminate the translation process and enhances the relevance and timeliness of translation teaching guidance. Furthermore, AI-powered intelligent assessment combined with dynamic human-AI interaction contribute to creating an autonomous learning environment, promoting the transformation of translation teaching toward a learner-centered model supported by digital and AI technologies.

**Keywords:** AI-powered assessment; translation process; semantic similarity; target text naturalness

## 1 引言

大语言模型 (large language models, 简称 LLMs), 尤其是具备跨语言泛化能力的多语言模型为语言互通提供了重要支持, 直接影响翻译行为的实现方式, 也影响着翻译研究的对象和趋势。近年来, 学界密切关注大语言模型与翻译之间的关联, 探讨大语言模型作用于翻译的方式以及翻译可能受到的影响。然而, 就目前研究内容和应用方式看, 大语言模型作用于翻译的方式局限于对话式交互, 模型主要用于译文产出, 其在语言质量多维评估方面的潜在能力并没有得到充分认识, 而这一能力对于翻译人才培养自主学习环境的创建至关重要。本文探索大语言模型作用于翻译行为和翻译教学的新方式, 为数智赋能翻译教育提供应用方案。

实现上述研究目标需要认识大语言模型的语言能力, 判断哪些语言能力可用于翻译评估, 探索大语言模型智能评估在翻译过程中发挥作用的方式和效果。在探讨大语言模型语言能力时, 需重点关注大语言模型通过计量对语言意义和语感的把握能力, 即运用语义相似度和困惑度对译文进行评价。将智能评估嵌入翻译过程需要使用多个多语大语言模型, 鼓励译者与大语言模型的动态互动, 同时将评估实时嵌入翻译过程。

## 2 大语言模型在翻译和翻译教学中的应用

人工智能时代, 大语言模型赋能的“机器翻译+译后编辑”成为主流翻译模式 (王华树 等, 2021; 肖志清 等, 2021), 不仅提高了翻译的准确性, 还增强了译文流畅性和一致性 (Kwok et al., 2025; Kruk et al., 2025)。然而, 涉及法律、中医药、文学等复杂文本时, 该翻译模式准确性略显不足 (Lee, 2023; Moneus et

al., 2024; 许明武 等, 2025)。此外, 视频字幕机器翻译也存在语义偏差、惯用语失当及音画不同步等问题, 易造成观众理解障碍 (Jin et al., 2023)。大语言模型虽具速度和成本优势, 但对文化内涵的“捕捉”仍逊于人工翻译 (李奉栖, 2022; Sawi et al., 2024), 它在文学翻译上最大的“缺陷”并非理解能力不足, 而是在同一文本内部的翻译表现差异巨大, 译文像是不同译者拼凑起来的“大杂烩” (袁筱一等, 2025)。

人工智能技术的蓬勃发展给传统翻译教学带来诸多挑战 (胡开宝 等, 2024), 也推动单向灌输式教学模式向智能化、交互式人机协同模式转变 (王律 等, 2024; 李锡阳 等, 2024)。大语言模型不仅推动了数字化教学资源建设, 在学习者翻译质量评估反馈方面也具有较高的可行性和有效性 (吕晓轩 等, 2024; 冯庆华, 2025; 王华树 等, 2024)。Xu 等 (2025) 创建了集成 ChatGPT 功能的翻译教学平台, 基于 Hurtado Albir (2015) 的评估量表设置提示词进行翻译评估, 发现智能评估提升了反馈的质量与时效性, 增强了学习者的认知投入。Cao 等 (2025) 发现, 大语言模型在译文词汇复杂性和指代连贯性方面的反馈优于教师反馈和自我反馈。在实践层面, 当前研究显示, 大语言模型赋能翻译与翻译教学的应用场景较为单一, 学习者多将其用作译文生成工具。而在“AI + 翻译教学”混合式教学模式中, 大语言模型则主要作为智能提取辅助翻译资源的工具参与教学环节 (舒晓杨, 2021)。

大语言模型应用于学习者翻译质量评估反馈的效度和合理性虽得到关注, 但当前的智能评估主要依赖提示词, 通过与大语言模型聊天问答方式对译文进行质性评价。这一模式有其不足之处。一是该模式缺少对译文忠实性与地道性的数值型评估, 不利于教师和学生直观比较原文与译文。二是学生虽能得到直接的生成式语言反馈, 却也容易过度依赖模型生成结果, 削弱主动思考与独立判断能力。在真实的教学情境中, 智能评估仅充当教师的“助手”, 辅助教师评价反馈, 这不仅限制学生在自主学习过程中获取实时质量评估, 也影响教师及时、有效诊断学情, 其培养学习者批判性思维和自我效能感的潜在教育价值有待进一步发掘。

综上, 目前缺少能够提供实时诊断性反馈的智能互动自主学习平台, 也缺乏有效的反馈机制构建动态交互场景。秦洪武等 (2024) 认为, 大语言模型代行教师评估角色, 可与学习者译文产出形成良性动态互动。根据这一场景设计理念, 本研究拟探讨基于语义计算的翻译质量多维智能评估方法, 创建 AI 赋能的翻译和翻译教学操作平台, 实现智能评估与自主学习有机结合, 强化学习者长效学习动机, 推动其翻译能力递进式发展。研究尝试回答以下问题: (1) 如何发挥大语言模型的语言能力, 为翻译增加智能质量评估维度? (2) 如何将智能评估深度嵌入翻译与翻译教学过程, 实现人与智能体的动态互动?

### 3 大语言模型的语言能力和评估维度

大语言模型支持词嵌入对应。跨语言词嵌入（cross-lingual word embeddings, 简称 CLWEs）是指在一个共有的空间里（shared space）表达两个或多个语言中的词语，不同语言里语义近似的词汇会彼此接近（Ormazabal et al., 2021）。嵌入是使用一个数值集合表达复合信息的方式，既可以表达文本信息，也可以表达图像信息。此外，嵌入还可以供机器学习或跨模型调用。语言学长期难解的语义计算问题，随着大语言模型和嵌入技术的发展获得了新的解决路径。语义计算可能会推动语言研究在坚守形式规则的同时尝试探索更深的语义层次。

生成式人工智能的核心构件是转换器（Transformer），包含自注意力机制<sup>①</sup>和位置编码<sup>②</sup>。此外，思维链（chain of thought）使得模型在学习大量语言数据后构建对语言结构和意义的解释。大语言模型可模拟人类推理过程，这种结构化推理使其语言产出更具连贯性和逻辑性。融合思维链的大语言模型所展示的推理过程和推断结果已经不是传统意义上的词汇概率逼近模型，而是更能逼近人类思维过程的计算模式（秦洪武等，2024）。

除了上述几点，还有几项关键技术推动了大语言模型文本生成能力的飞跃。共享标记机制（shared tokens）可以处理跨多个对话轮次或信息的上下文和信息，保持生成和响应的一致性和相关性。嵌入对应（embedding alignment）将来自不同嵌入空间或模型中的词嵌入关联起来，并在其表征间建立有意义的对应关系（秦洪武等，2024），实现知识迁移（广域知识提供、解释、翻译、内容生成和个性化学习支持），辅助人类进行智能判断。

#### 3.1 语义相似度测量

大语言模型将词、句子和语言片段投射到高维空间的向量，构成词、句和语段嵌入，意义相近的嵌入相近。多语模型如 mBERT 内涵的嵌入具备语言间共享属性，语言之间和语言内部的词、句或概念的关系可以通过相似度得分表达出来，如表 1。

表 1 显示，词（标记）的方向、量和聚类关系是词序列、词与词、句与句关系的数字化表达，既有关系的方向，也有关系的权重，这些数据构成可计算的语义。它不同于我们对意义的传统认识：按语义学传统，符号意义是人赋予的符号与所指之间的关系，一种由社会认知约定的象征性关系，语言意义是符号意义的组合。传统上的语义关系无法数字化，但语境决定语义的基本语言观念（Firth, 1935）可以通过数字化加以计算和表达，即一个词的意义由该词与其他词的高维向量关系决定。

① 允许模型权衡序列中不同词相对于某一给定词的重要性，提取序列中任意先前点的状态信息，解决长距离关系捕捉问题。

② 转换器在输入文本嵌入中添加位置编码，以提供有序标记位置的信息。

表1 基于句嵌入的句子相似度计算

| 源语句子                                                                           | 比较句                    | 句嵌入                                                                                                                                                  | 源语与比较句相似度         |
|--------------------------------------------------------------------------------|------------------------|------------------------------------------------------------------------------------------------------------------------------------------------------|-------------------|
| Writers make national literature, while translators make universal literature. | 句一：作家创作民族文学，而译者创作普世文学。 | <b>源语：</b><br>[-5.36693931e-02 -7.35808583e-03 -5.90024926e-02<br>-4.47580777e-02...9.57035366e-03 -1.92590524e-02<br>2.60802004e-02 3.69540695e-03] | <b>0.8464961</b>  |
|                                                                                |                        | <b>句一：</b><br>[-7.41371065e-02 2.66810488e-02 -2.67775636e-02<br>-3.99937555e-02...8.62570945e-03 -3.74894142e-02<br>3.32789198e-02 -8.07779189e-03] |                   |
|                                                                                |                        | <b>句二：</b><br>[-5.16062677e-02 -7.31601473e-03 -4.49571991e-03<br>-4.69454937e-02...-1.81823578e-02 3.16572934e-02<br>3.72979082e-02 4.43525583e-04] |                   |
|                                                                                | 句二：翻译是通过不同的方式产生类似的效果。  |                                                                                                                                                      | <b>0.34933448</b> |

3.2 自然程度测量

大语言模型对语言自然程度的度量来自模型提供的困惑度计算。困惑度衡量语言模型预测一个词语序列的准确程度，它能有效反映一个句子与特定模型所学的语言模式间的契合程度。新输入的句子在高维向量表达上越接近常态，模型困惑度越低，反之则越高，比如例（1）。

（1）Tom is an English teacher running a financial company.

这个句子的每一个词由大语言模型预测的对数概然率（Log probabilities）分别为：Tom: -6.4695；is: -0.1400；an: -0.0036；english: -0.2939；teacher: -5.3322；running: -6.0449；a: -0.0076；financial: -8.5376；company.: -2.4114。对数概然率越低，困惑度越高，由此可以简单判断Tom、teacher、running、financial用得都不太自然。类似于人类的语言直觉，大语言模型常态判断也认为“英语教师经营金融公司”不合常规。

我们进一步使用大语言模型掩码学习方式预测掩码位置的词，如例（2）。

（2）Tom is an English [MASK] running a financial company.

我们让模型提供5个概然率最高的词，分别为：businessman: 0.3029；man: 0.0902；billionaire: 0.0431；lawyer: 0.0377；accountant: 0.0344。模型预测的词首选businessman（商人），显然这也合乎人类基于世界知识对预测词的期待。这说明大

语言模型不仅具备语法能力，还具备世界知识。语言自然度判断是语言知识和世界知识共同作用的产物，大语言模型在此方面与人类判断高度趋同。根据例（1），我们将高困惑度词依次改为预测词（如例3），则成为一个符合大语言模型常态和人类语言直觉的低困惑度句子。

（3）He is an English businessman running a construction company. (Perplexity: 3.2234)

凭借高算力、大储量，生成式人工智能通过计算高维向量计算语义关系。不同于人类通过符号—概念—客体关系建立语义理解，但二者均胜任语言理解任务，可谓殊途同归。相较于传统人工智能和基于有限特征变量的语言分析，生成式人工智能可基于十分复杂的向量关系提供整体的语义和语言判断，是助力人类推理和判断的智能实体，能协同人类完成高层次的、基于语义和语用的翻译关系判断和语言运用质量分析任务。

## 4 大语言模型嵌入翻译与翻译教学过程

大语言模型可以提供初译译文，也可以对人工初译进行修改，修改后依然可以称之为初译，初译形成可能经历的翻译操作不在本研究探讨的范围。本文聚焦译后编辑阶段，大语言模型如何为译文质量提供智能评估与辅助支持。

### 4.1 适用环境：译后编辑

人工译后编辑的效率、可靠性受个体翻译水平、资源利用能力和工作环境等多种因素影响，编辑效果因人而异。目前采用的翻译自动评估主要关注译文间比较，很难为翻译文本提供忠实性、地道性以及微观层次（句子翻译）的翻译建议，因此，高质量的译后编辑目前过度依赖译者的翻译和语言水平，很难体现智能翻译的效率和优势。本文建立一种新的人类智能、人工智能协同模式，探讨新型智能能否辅助译者解决上述问题。

人工智能辅助译后编辑的方式包括：（1）提示语（Prompt）模式：主要为在线应用或者本地部署的小参数语言模型（如Ollama），但每次操作需重复输入提示语，或过于依赖预设的提示，效率和灵活性有限；（2）智能体（Agent）模式：通常借助大语言模型的扩展插件，如通过AI划词程序可快速响应，极大提升工作效率。该模式支持对译文或源语重点片段提供翻译和修改建议；（3）协同模式：人类调节代理，加快完成文档编辑。编辑结果经人类审定后定稿。

提示语模式已为译界熟知，这里不需赘言。智能代理模式已经常见于在线应

用，如Kimi、豆包、DeepL等应用，学习者方便使用，但无法进行学习数据管理。本文介绍的是可以进行本地部署，同时也便于管理翻译过程数据的智能协同模式，即人类翻译得到大语言模型的支持，协同完成译文修改、润色、编辑，其主要工作流程如图1所示。

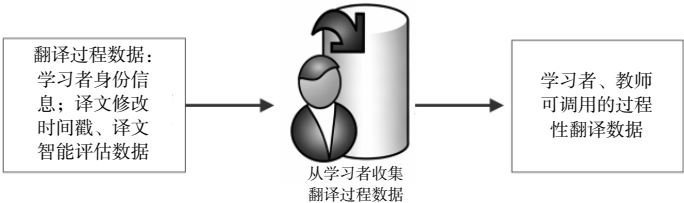


图1 AI赋能翻译过程数据管理

大语言模型可实时提供译文智能评估结果，包括原文-译文语义相似度和译文自然度/困惑度评估。该评估可以基于大语言模型如Chinese-bert-wwm-ext的困惑度本身，也可以通过困惑度以及与用户自制汉语参考句嵌入库相似度计算加权得出，如图2所示。

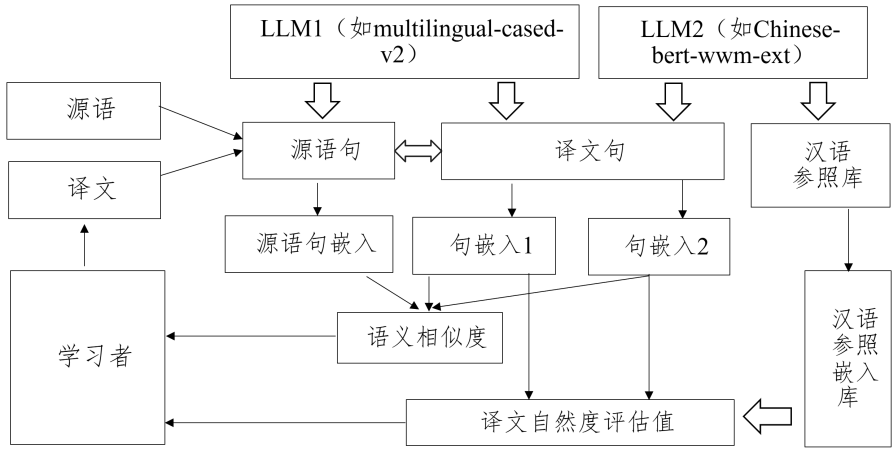


图2 AI赋能翻译过程评估的设计思路（秦洪武等，2024）

图2展示了AI赋能翻译评估的设计思路。源语和译文首先切分为句子，通过大语言模型的转化器转换成句嵌入。译文句包含原始译文句和新编辑译文句，分别通过转换器转换为句嵌入1和句嵌入2。源语句嵌入分别和译文句嵌入1、句嵌入2对比得到语义相似度；同时，句嵌入1和句嵌入2在目标语（如汉语）模型中得到困惑度值并转换成自然度；也可以通过汉语参照嵌入库通过相似度计算得到转换（proxy）的自然度。这些数值可以在工作界面上直接反馈给学习者，学习者可以判断评估值，再次对译文进行评估，如此反复，直到得到满意的译文。

## 4.2 智能评估嵌入翻译过程

由于智能评估表达为量化数值，将智能评估嵌入翻译过程需要构建平台予以支持。平台主要设计思路如下。

(1) 使用者友好的平台工作环境。工作界面有源语、译语、句级对齐（初译）、句级对齐（译后编辑），以及可即时更新的语义相似度、自然度和复合语言质量评估数值。

(2) 确定评估维度。我们选择语义相似度和译文自然程度：源语—译文的语义相似度评估译文在语义上的忠实性，译文自然程度计量译文语言是否符合目标语语言规范，反映译文的自然度。

(3) 确定显示方式。基于向量化语言数据显示语义相似度和译文自然程度数值，译文更新时可以即时更新数值。

(4) 数据存储和提取。支持存储译文、初译评估数据和译后编辑的译文评估数据，按时间戳记录多次译后编辑的译文评估数据；数据以半结构化方式存储在数据库（如MySQL）中，可以通过检索页面检索译后编辑信息。

设计时需要解决两个关键问题：一是外挂知识库，二是语言质量复合指数。外挂的知识库主要用于评估译文质量。现有大语言模型提供的困惑度是每一标记的平均预测损失函数，句子越短，预测损失平摊的标记就越少，越难以稀释损失；而长句能够有更多标记平摊预测损失，故困惑度低。鉴于此，在评估汉语译文质量时，BERT类模型不能为短译文提供理想的困惑度评估。这时，需要考虑自建外挂目标语参考库，如特定领域（高质量文学语言语料）的汉语语料库。该参考库是向量化汉语句嵌入库，用于汉语译文和原生汉语间的句嵌入比较。这样，通过嵌入比较句级译文，并与困惑度结合（困惑度70%，嵌入相似度得分30%），得到自然度综合评估值（naturalness）。自然度综合评估值关涉译文的“通顺”，和语义相似度（similarity）共同实现对句级汉语译文质量的智能自动评估。

图3显示，智能语言质量评估对译文改动敏感。在翻译过程中所做的改动会得到即时反馈，学生可以根据反馈数值审视所做修改，考虑是否需要进一步修改。通过这种方式，学生在与大语言模型的互动中寻求忠实与通顺平衡，在争取更高评估结果的挑战中更为专注和投入，更利于提升翻译水平。

此外，汉英翻译平台还针对英语译文使用了传统的语言质量评估指数。如句长、形符类符比、可读性等，形成综合语言质量评估指数，如图4。

综合评估指数基于语言的形式特征，主要构成如下：(1) TTR（句子的类符/形符比）；(2) 平均句长；(3) 小句数量；(4) Flesch\_Kincaid（以年级表达的可读性）；(5) 词汇重复率。将以上五个指标作归一化处理，得到5个数值，5个数值相加除以5然后乘以10即为可读性综合指标。该指数与人工智能的评估不同，具有可解释性，其作用是在相似度和困惑度出现较大偏差时提供一种数值参考。



图3 英译汉智能编辑与评估系统



图4 汉译英智能评估系统界面

### 4.3 翻译过程评估数据的存储和应用

翻译过程数据涉及学生身份信息、源语文本、每次修改润色的译文、译后编辑提交时间、智能评估（相似度、困惑度）和传统的语言质量评估数据（见表2）。数据以结构化形式在数据库中存储（如MySQL），可通过PHP生成的网页界面高效调用。该存储机制既支持学生查询自身翻译过程以促进学习反思，也为教师观测学生翻译学习行为提供了直接数据。

表2 译后编辑智能评估数据结构化存储

| Timestamp         | name | grade | class | level    | learner | Alignment | Source               | Target                                                                                     | Similarity | Perplexity | Comprehensive |
|-------------------|------|-------|-------|----------|---------|-----------|----------------------|--------------------------------------------------------------------------------------------|------------|------------|---------------|
|                   |      |       |       |          | id      | Type      |                      |                                                                                            |            |            | Index         |
| 2025/7/2<br>18:25 | MT   | 2023  | 1     | Graduate | 01      | Initial   | 河面上泛着夕阳的余晖，将水流染成金红色。 | The afterglow of the setting sun casted on the surface of the river, dyeing it golden red. | 0.818764   | 35.43274   | 3.196078431   |
|                   |      |       |       |          |         |           |                      |                                                                                            |            |            |               |
| 2025/7/3<br>16:12 | WH   | 2023  | 1     | Graduate | 02      | renewed   | 河面上泛着夕阳的余晖，将水流染成金红色。 | The river shimmered with the glow of sunset, turning the water golden red.                 | 0.896188   | 34.8355    | 3.612368024   |
|                   |      |       |       |          |         |           |                      |                                                                                            |            |            |               |
| 2025/7/2<br>18:25 | MT   | 2023  | 1     | Graduate | 01      | Initial   | 岸边芦苇随风轻摆，发出沙沙的声响。    | "shing-shing-", the reeds on the riverside were waving slightly in the wind.               | 0.520138   | 106.0736   | 2.740573152   |
|                   |      |       |       |          |         |           |                      |                                                                                            |            |            |               |
| 2025/7/3<br>16:12 | WH   | 2023  | 1     | Graduate | 02      | renewed   | 岸边芦苇随风轻摆，发出沙沙的声响。    | Reeds swayed along the bank, rustling in the wind.                                         | 0.57985    | 80.0884    | 3.854030501   |
|                   |      |       |       |          |         |           |                      |                                                                                            |            |            |               |

| Timestamp         | name | grade | class | level    | learner Alignment<br>_id      _Type | Source  | Target                                                      | Similarity                                                                                                                                      | Perplexity | Comprehensive<br>_Index |             |
|-------------------|------|-------|-------|----------|-------------------------------------|---------|-------------------------------------------------------------|-------------------------------------------------------------------------------------------------------------------------------------------------|------------|-------------------------|-------------|
| 2025/7/2<br>18:25 | MT   | 2023  | 1     | Graduate | 01                                  | Initial | 黑孩子的裤脚已经<br>被露水打湿，但<br>他浑然不觉，只<br>是专注地盯着水<br>面下银光闪烁的<br>小鱼。 | The black boy were<br>staring at the silver<br>fish in the river,<br>without noticing<br>that his trouser legs<br>were already wet<br>with dew. | 0.782096   | 40.98302                | 4.826086957 |
|                   |      |       |       |          |                                     |         | 黑孩子的裤脚已经<br>被露水打湿，但<br>他浑然不觉，只<br>是专注地盯着水<br>面下银光闪烁的<br>小鱼。 | The boy's trouser<br>legs damp with<br>dew, but he<br>remained<br>motionless, staring<br>intently at the<br>silvery fish darting<br>below.      | 0.811231   | 31.35503                | 6           |
| 2025/7/2<br>18:25 | MT   | 2023  | 1     | Graduate | 01                                  | Initial | 老铁匠走过来，<br>用粗糙的大手拍<br>了他的肩膀。                                | The old blacksmith<br>came, and patted<br>the boy's shoulders<br>with his gnarled<br>hands.                                                     | 0.79166    | 27.48215                | 5.008403361 |
|                   |      |       |       |          |                                     |         | 老铁匠走过来，<br>用粗糙的大手拍<br>了他的肩膀。                                | The old blacksmith<br>approached and<br>patted his shoulder<br>with a rough hand.                                                               | 0.888097   | 63.89995                | 5.058823529 |

续表

图5展示了智能评估嵌入翻译与翻译教学过程的思路。教师可以基于教学目标和问题，合理设计教学内容，使用大语言模型生成原文和有问题的干扰译文，由学生通过人机协作完成译后编辑任务，并在智能评估反馈的引导下反思修订译文。教师通过系统后台实施翻译过程监测，可直接观察学生译文质量，也可通过检索特定原文，对不同学生译文进行横向比较，或对学生翻译行为数据进行历时追踪，由此判断是否需要干预学生的翻译，或者捕捉典型例证在课堂教学中与学生分享、讨论，实现有针对性的教学，精准帮助学生学习。

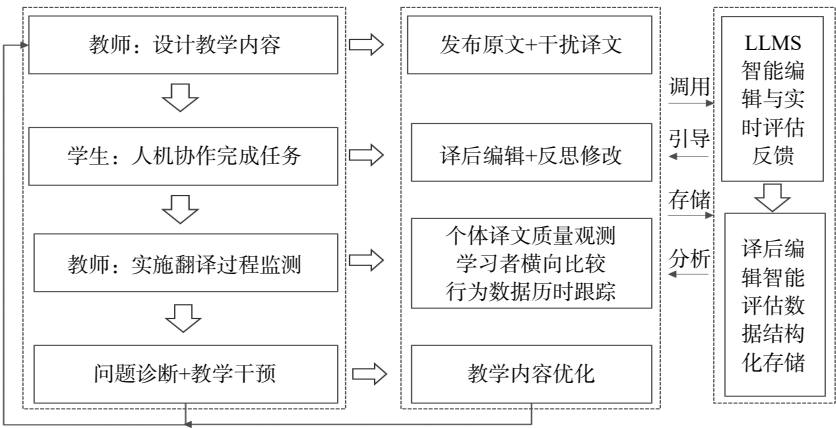


图5 智能评估嵌入翻译与翻译教学过程流程

在智能评估平台教学试用中发现，学生在处理拼写、语法以及比较明显的错译、漏译等问题时，使用译后编辑处理效率更高，翻译过程数据表现最佳。在应对语言地道性问题和比较细微的语义搭配问题时，学生需要付出较多认知努力，但总体表现仍属良好。以表3中的译文一（Initial1）为例，“20th century’s crisis”并不符合英语母语者的表达习惯。学生一般能够根据实时译文智能评估反馈，识别并纠正此类中式英语表达。修改后的译文语义相似度变化不大，但困惑度指数多数有明显改善。此外，学生WH敏锐地察觉到“appeared with”的搭配问题，并将其修改为更地道的“emerged with”，语言困惑度指标因之表现最佳（24.0886）。学生JX的译文因存在拼写、语法、语义逻辑等问题，困惑度分值异常偏高（531.2032）。针对这些问题，教师可以选择表现突出或存在不足的典型案例组织课堂分享和讨论，也可针对学生译文普遍存在的问题及时干预，并据此有针对性地设计教学内容。

表3 译后编辑多译文智能评估数据

| Timestamp            | Learner<br>_id | Alignment<br>_Type | Source                      | Target                                                                                                    | Similarity | Perplexity | Comprehensive<br>_Index |
|----------------------|----------------|--------------------|-----------------------------|-----------------------------------------------------------------------------------------------------------|------------|------------|-------------------------|
| 2025/7/2<br>17:55:00 | MT             | Initial1           | 二十世纪人类的信仰危机是随着全球化的逐渐兴起而出现的。 | 20th century's crisis of faith in humanity appeared with the rise of globalization.                       | 0.956049   | 141.1191   | 5.008403                |
| 2025/7/2<br>17:55:00 | WH             | Renewed            | 二十世纪人类的信仰危机是随着全球化的逐渐兴起而出现的。 | The crisis of faith in humanity emerged along with the gradual rise of globalization in the 20th century. | 0.949082   | 24.0886    | 3.357298                |
| 2025/7/3<br>16:47:00 | JX             | Renewed            | 二十世纪人类的信仰危机是随着全球化的逐渐兴起而出现的。 | Humans' faith crisis in twentieth century comes, engaging with the appearance of the globalization.       | 0.939572   | 531.2032   | 3.078431                |
| 2025/7/4<br>12:17:00 | LM             | Renewed            | 二十世纪人类的信仰危机是随着全球化的逐渐兴起而出现的。 | The belief crisis in the 20th century emerged with the rise of globalization.                             | 0.884241   | 69.11776   | 4.189542                |
| 2025/7/5<br>10:31:00 | XL             | Renewed            | 二十世纪人类的信仰危机是随着全球化的逐渐兴起而出现的。 | The faith crisis in humanity in the 20th century appeared with the gradual emergence of globalization.    | 0.952656   | 60.02284   | 3.986928                |
| 2025/7/5<br>11:46:00 | XL             | Renewed            | 二十世纪人类的信仰危机是随着全球化的逐渐兴起而出现的。 | The faith crisis in humanity in the 20th century appeared with the gradual rise of globalization.         | 0.95048    | 55.34      | 3.986928                |
| 2025/7/4<br>9:42:00  | MT             | Initial2           | 远处传来狗叫声，时断时续。               | From afar came intermittent barks of a dog.                                                               | 0.818808   | 323.6256   | 4.588235                |
| 2025/7/4<br>9:42:00  | JX             | Renewed            | 远处传来狗叫声，时断时续。               | The barks of a dog came from afar intermittently.                                                         | 0.811049   | 79.04845   | 4.705882                |
| 2025/7/5<br>11:46:00 | XL             | Renewed            | 远处传来狗叫声，时断时续。               | Intermittent barks of dogs came from the distance.                                                        | 0.810702   | 53.16677   | 4.588235                |
| 2025/7/4<br>12:17:00 | LM             | Renewed            | 远处传来狗叫声，时断时续。               | Intermittent barks of a dog came from afar.                                                               | 0.822473   | 64.0988    | 4.588235                |

学习者不仅能够识别译文错误或非地道表述，还能够根据反馈数据，反思并总结智能体的翻译偏好，及时调整译文表述。以表3中的译文二（Initial2）为例，因倒装句式的困惑度数值异常偏高（323.6256），学习者为了获取更理想的评分，会主动还原正常语序以降低困惑度值。在这种情况下，教师须引导学生正确认识智能评估分值的相对参考性，在人机协作过程中始终保持认知主体地位。

平台储存的历时数据还可以帮助教师有效追踪学生对译文的多轮修改过程，这不仅为教学提供了实证基础，也为学生翻译能力的提升提供参照。对学习行为数据跟踪观察显示，在每项翻译任务中，学习者普遍能够做到2—4轮译后编辑。若反馈数据不理想，学习者会快速定位问题并修正译文。例如，学生WH对初始译文进行了两轮译后编辑（见表4）。首轮修改因语法问题导致评估结果并不理想，困惑度值居高。第二轮修改后相似度有所提升，困惑度分值降至28.74。

表4 多轮译后编辑智能评估数据

| Timestamp            | Learner<br>_id | Alignment<br>_Type | Source                     | Target                                                                                 | Similarity | Perplexity | Comprehensive<br>_Index |
|----------------------|----------------|--------------------|----------------------------|----------------------------------------------------------------------------------------|------------|------------|-------------------------|
| 2025/7/2<br>17:55:00 | MT             | Initial1           | 他们对市场经济一无所知，他们公司因而受了很大的损失。 | They knew nothing about the market economy, which costs a lot to their company.        | 0.8145     | 64.13      | 5.29412                 |
| 2025/7/2<br>17:55:00 | WH             | Renewed1           | 他们对市场经济一无所知，他们公司因而受了很大的损失。 | Lacking knowledge of market economy, their company suffered a great loss.              | 0.8975     | 84.78      | 3.357298                |
| 2025/7/3<br>16:47:00 | WH             | Renewed2           | 他们对市场经济一无所知，他们公司因而受了很大的损失。 | Due to their lack of knowledge of market economy, their company suffered a great loss. | 0.9203     | 28.74      | 3.078431                |

5 结语

智能评估和人机动态互动有利于自主学习环境的创建。相对于传统教学的反馈滞后，智能评测反馈对提升学习者的学业表现有很大优势（吴智慧，2023）。有效的实时反馈可刺激学生在最短时间内了解自身的翻译能力和学习成效并加以修正。此外，数值型评估可以给学生更多思考和尝试空间，更易激发学生的挑战和探索欲望，从而增强学生在反思、比较、推理和总结等方面的认知投入，促进其深度学习。这可以在一定程度上避免提示语反馈代替学生思考、降低其自主学习能力的风险。

智能评估嵌入翻译教学，可推动翻译课堂教学向学生中心转型。教师可以利用数据库存储的学习行为数据，从多个侧面观察学生的学习状态，有针对性地设计教学内容和教学环节。这一场景中，教师角色转型为“教学引导者”和“内容创新者”（文旭 等，2024）。大语言模型不再局限于充当教师的“助手”，而是成为教师与学生的“合作者”，协同解决翻译和教学问题。需要注意的是，智能评估只具有参考性，教师应审慎分析评估结果，引导学习者合理看待评估值。此外，智能评估平台提供的是句级译文质量的智能评估，教师应引导学生关注上下文语境信息，重视语篇的衔接与连贯。

总而言之，大语言模型，特别是具备跨语言泛化能力的多语言模型，正在深刻改变语言互通与翻译实践的实现方式，并进一步影响翻译研究的对象、方法与发展趋势。大语言模型除聊天功能之外的多维语言能力，如多维评估能力，仍有待进一步认识、开发和应用，以创建自主学习环境，催生新的翻译和翻译教学范式，从而更有效地服务于翻译人才培养。

## 参考文献

- Cao S, Zhou T, 2025. Exploring the effectiveness of ChatGPT-based feedback compared with teacher feedback and self-feedback: evidence from Chinese to English translation[J]. *SAGE open*, 15(3): 1-18.
- Firth J R, 1935. The technique of semantics [J]. *Transactions of the philological society*, (1): 36-73.
- Hurtado Albir A, 2015. The acquisition of translation competence. Competences, tasks, and assessment in translator training. *Meta*, 60(2), 256-280.
- Jin H, Yuan Z, 2023. The application of machine translation in automatic dubbing in China: a case study of the feature film *Mulan*[J]. *Linguistica Antverpiensia, new series: themes in translation studies*, 22: 80-94.
- Kruk M, Kaluzna A, 2025. Investigating the role of AI tools in enhancing translation skills, emotional experiences, and motivation in L2 learning[J]. *European journal of education*, 60(1): 1-12.
- Kwok H, Shi Y, Xu H, et al., 2025. GenAI as a translation assistant? A corpus-based study on lexical and syntactic complexity of GPT-post-edited learner translation[J]. *System*, 130: 1-14.
- Lee T, 2023. Artificial intelligence and posthumanist translation: ChatGPT versus the translator[J]. *Applied linguistics review*, 15(6): 2351-2372.
- Moneus A M, Sahari Y, 2024. Artificial intelligence and human translation: a contrastive study based on legal texts[J]. *Heliyon*, 10(6): 1-14.
- Ormazabal A, Artetxe M, Soroa A, et al., 2021. Beyond offline mapping: learning cross-lingual word embeddings through context anchoring[C]//Association for Computational Linguistics. Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing. California: Association for Computational Linguistics: 6479-6489.

- Sawi I, Allam R, 2024. Exploring challenges in audiovisual translation: a comparative analysis of human- and AI-generated Arabic subtitles in Birdman[J]. Plos one, 19(10): 1-20.
- Xu S, Su Y, Liu K, 2025. Investigating student engagement with AI-driven feedback in translation revision: a mixed-methods study[J]. Education and information technologies, 30(12): 16969-16995.
- 冯庆华, 2025. DeepSeek在翻译教学与研究中的创新应用[J]. 中国翻译 (2): 58-67.
- 胡开宝, 李娟, 2024. 大语言模型背景下的翻译人才培养: 挑战与前景[J]. 外语电化教学 (6): 3-7.
- 李奉栖, 2022. 人工智能时代人机英汉翻译质量对比研究[J]. 外语界 (4): 72-79.
- 李锡阳, 彭丽华, 金慧, 2024. 技术赋能的人机协同翻译教学模式构建研究[J]. 外语界 (3): 43-50.
- 吕晓轩, 冯莉, 2024. 人工智能赋能的CSE笔译能力量表在教学评价中的应用[J]. 外语界 (6): 29-36.
- 秦洪武, 鲁艳芳, 2024. 大语言模型与外语教育: 基于语言能力的应用研探[J]. 外语界 (6): 37-44.
- 舒晓杨, 2021. AI环境下基于工作场所学习的递进式笔译教学工作坊实践探索[J]. 外语电化教学 (2): 65-72.
- 王华树, 刘世界, 2021. 人工智能时代翻译技术转向研究[J]. 外语教学 (5): 87-92.
- 王华树, 谢斐, 2024. 大语言模型技术驱动下翻译教育实践模式创新研究[J]. 中国翻译 (2): 70-78.
- 王律, 王湘玲, 2024. 人工智能时代的翻译教学研究:概念界定、逻辑框架与实践路径[J]. 外语界 (6): 83-90.
- 文旭, 田亚灵, 2024. 人类智能与人工智能在外语教育与研究中的融合[J]. 外语电化教学 (4): 18-24.
- 吴智慧, 2023. 基于智慧平台的英语混合式教学研究——以低成就英语学习者为例[J]. 中国电化教育 (12): 121-127.
- 肖志清, 魏光凤, 2021. 人机交互理念下CAT+MT+PE师生协同翻译模式探索[J]. 语言教育 (2): 69-74.
- 许明武, 张凯维, 2025. 中医药典籍智能翻译与多模态传播研究[J]. 外语电化教学 (1): 35-40.
- 袁筱一, 罗茜, 2025. 人工智能与文学翻译的未来——基于人机文学翻译对比研究的思考[J]. 外语与外语教学 (2): 1-9.

## 通信地址

276800 山东省日照市 曲阜师范大学翻译学院 (夏云)

273165 山东省曲阜市 曲阜师范大学外国语学院 (秦洪武)

(责任编辑: 杨瑞雪)

# 基于Praat内置神经网络的自适应LPC阶数共振峰提取\*

复旦大学 蒋博文

**提 要：**传统线性预测系数（linear prediction coefficient，简称LPC）法在提取共振峰时采用固定阶数，通常存在前高元音出现“虚假共振峰”，后高元音发生“共振峰合并”等问题。为此，本文提出一种基于Praat内置神经网络的自适应LPC阶数共振峰提取方法。该方法基于神经网络模型，以帧为单位，自动将语音帧分类为“清音/噪声”“前高元音”“后高元音”及“其他浊音”段，并据此自适应调整LPC阶数以进行鲁棒、准确的共振峰提取。实验表明，模型对语音帧分类的准确率达89%以上，显著降低了复元音等复杂音段的共振峰提取错误率。本方法直接在Praat平台实现，兼具界面友好、低成本与易复现的优势，为汉语方言韵母系统的基础声学研究提供了一个高效准确的自动化分析工具。

**关键词：**元音；共振峰；线性预测系数；神经网络

## Formant Extraction With Adaptive LPC Orders Based on the Praat-Embedded Neural Network Module

Bowen Jiang

**Abstract:** Conventional Linear Prediction Coefficient (LPC) methods use fixed orders for formant extraction, which often encounter issues such as “spurious formants” in front high vowels and “formant merging” in back high vowels. To address these problems, this paper proposes an adaptive LPC-order formant extraction method based on the Praat-Embedded neural network. Utilizing a neural network model on a frame-by-frame basis, this method automatically classifies speech frames into “voiceless/noise” “front high

\* 本文系河南省2025年科技攻关项目“说话人音视频深度伪造检测技术研发及应用”（项目编号：252102320061）和2026年度中央高校基本科研项目“基于物理特征分析和高性能模型优化策略的深伪音视频数据鉴别技术研究”（项目编号：2026ZJBKY033）的阶段性研究成果。

vowel” “back high vowel”, and “other voiced” segments. Based on these classifications, the LPC order is adaptively adjusted to achieve robust and accurate formant extraction. Experimental results demonstrate that the frame classification accuracy exceeds 89%, which further significantly reduces the formant extraction error rates for complex segments especially diphthongs. Implemented directly within the Praat platform, this method offers the advantages of a user-friendly interface, low cost, and high reproducibility, providing an efficient and accurate automated analysis tool for the fundamental acoustic research of vowel systems in Chinese dialects.

**Keywords:** vowel; formant; linear prediction coefficient; neural network

## 1 引言

元音音质的声学差异主要体现在前三个共振峰（formant）的频率分布上，无论是从产出还是感知的角度来看，共振峰频率的分布特征对于不同种类元音的判别有着关键作用。例如，对于普通话男性发音人而言，前高元音[i]的第一和第二共振峰（分别记为F1和F2）分别分布于250—350 Hz和2100—2300 Hz的频率范围内；后高元音[u]的F1和F2分别分布于300—400 Hz和400—500 Hz的范围内（吴宗济等，1989：96）。另外，圆唇化通常也会导致第三共振峰（记为F3）的降低（陈忠敏，2022）。随着计算机科学与信号处理算法理论的发展，实验语音学的方法在汉语方言的韵母系统研究中得到广泛应用，其实现的基础就在于能准确提取语音信号的共振峰。

现有的信号分析算法能实现绝大多数情况下的音频共振峰提取，但对于前高元音（如[i]）和后高元音（如[u]）的共振峰提取则存在问题。一方面，前高元音的F1和F2通常相隔较远，例如，[i]的F1一般在300 Hz以下，而F2通常高于2100 Hz，易出现“虚假共振峰”现象，即共振峰提取算法有可能在其间的某一频段误识别出一个并不存在的共振峰。另一方面，后高元音的F1和F2通常相距极近，如[u]的F1和F2往往只相隔100—200 Hz，易出现“共振峰合并”现象，即共振峰提取算法难以分辨出后高元音的F1和F2，而将两个峰合并为一个共振峰。“虚假共振峰”和“共振峰合并”错误会对汉语方言韵母系统的自动化分析造成严重后果。“虚假共振峰”会导致原本是F2的共振峰误识别成F3，而所计算出的F2是一个介于正确的F1和F2之间的虚假值；“共振峰合并”会导致原本是F2的共振峰与F1合为一体，而原本是F3的共振峰则被识别成F2。上述现象会造成共振峰次序混乱，导致提取出的共振峰频率值相对于真实值产生几百乃至上千Hz的误差，严重影响元音声学特征分析的可靠性，在实验语音学研究中往往难以采信。

目前,传统的共振峰提取算法通常基于线性预测系数(Linear prediction coefficient,简称LPC)实现,在应用LPC提取共振峰时,需要人工设定LPC阶数。理论上来说,LPC的阶数一般设置为待提取共振峰数的两倍以上。假设一段语音信号有5个共振峰,那么LPC阶数可设置为11(Ladefoged, 1996: 192-193)。然而这只是针对理想的语言信号而言,LPC阶数会影响信号识别,进而影响共振峰的提取结果。在实际操作过程中,LPC的阶数需要根据具体情况进行调整。当LPC阶数设置过高时,共振峰提取算法会对信号频谱中的谱包络峰值过于灵敏,导致“虚假共振峰”的出现;当LPC的阶数设置过低,共振峰提取算法又难以分辨相距较近的峰值,导致“共振峰合并”的现象。问题在于,LPC的阶数是人工设定的,传统方法无法自适应地调整用于共振峰提取的LPC阶数。当需要对复元音音节进行共振峰提取时,这个问题显得尤为突出。例如,汉语普通话的“有”字([iou]),其共振峰在音节内经历了从前高元音到后高元音的变化。此时,操作者必须先以人耳识别出前高元音段和后高元音段,将其手动标注出来,随后为不同的元音段人工设定LPC阶数。这一过程无疑是繁琐且低效的,不利于汉语方言韵母系统研究的自动化和高效化。另外,如果所有的语音样本都通过人工识别进行语音片段标注,标注出错的概率也会大大增加。

要解决上述问题,一个行之有效的策略就是将音频的标注过程通过一种自动化而非人工的方法实现。这一自动化任务可以通过神经网络模型完成。具体来说,神经网络的输入是音频或音频的声学特征,而神经网络的输出则是与输入数据相对应的音段类型分类结果,音段的类型包括清音段、前高元音段、后高元音段和其他浊音段等。神经网络是一种诞生于上世纪的非线性可训练的数学模型,随着2012年Alexnet(Krizhevsky et al., 2012)在图片分类任务中大放异彩,这一模型因其强大的自适应性、抗干扰性、可泛化性,以及不需要对输入数据进行任何统计学先验假设等优点,越来越多地被应用于各种自然语言处理任务,包括各种与汉语方言学语音研究相关的工作。例如,Jiang等(2024)通过一个采用自适应时序注意力机制和特征分离策略的神经网络,实现了针对汉语方言的音素自动化识别与标注;颜为之(2023)利用卷积神经网络和循环神经网络以及迁移学习策略,研发了一个基于语料库的江西境内客、赣方言的自动识别系统;赵泽彬等(2024)着眼于基于迁移学习的小样本语音识别研究,实现了神经网络模型从普通话到西南官话的知识迁移;付婧等(2020)和杨迪一(2023)也分别实现了基于神经网络的成都方言和陕北方言的语音识别。

上述研究表明,神经网络在汉语方言语音系统的研究中已经得到了广泛应用,并且取得了令人瞩目的成果。但是不难发现,现有研究对神经网络在汉语方言学中的应用往往停留在宏观地对方言自动识别或分类任务上,缺乏从物理和声学层面对语音信号本身更加细致而深入的分析。另外,现有研究对神经网络的搭建往往基

于Python编程语言和Pytorch机器学习框架。虽然这些机器学习框架具有功能强大、可编程性强的特点，也能够部署目前各种先进的神经网络模型，但对于广大汉语方言研究者来说，要熟练使用Pytorch框架，需要经过多年的程序设计训练以及开发经验的积累，该方法并不是一个易于上手、简单快捷的神经网络实现手段。此外，基于Python编程语言和Pytorch框架的神经网络开发需要配置软件运行的环境，而运行环境的配置过程不仅要下载大量的开发包，操作过程也较为繁琐。最后，结构复杂、可训练参数量庞大的神经网络虽然具有优越的性能，但也要求开发者配备合适且充足的硬件设备以提供足够的算力来训练并测试所搭建的神经网络模型，这不仅限制了设备的便携性<sup>①</sup>，也带来了高昂的经济成本。

针对上述问题，本文旨在构建一个准确、高效、能够自适应不同元音种类、易于上手和复现的语音信号共振峰自动化提取模型。首先，以帧为单位，用训练好的神经网络模型自动识别每一语音帧的语音类型。本文所规定的语音类型包括“清音段（含无语音信号的空白段）”“前高元音段”“后高元音段”和“其他浊音段”四类。随后，模型根据语音类型识别结果，以不同的LPC阶数对每一语音帧进行共振峰提取。具体来说，识别为“前高元音”的语音帧将在较低的LPC阶数下提取共振峰；识别为“后高元音”的语音帧将以较高的LPC阶数进行共振峰提取；若识别为“其他浊音”，将采用适中的LPC阶数；识别为“清音（或无语音信号）”则不进行共振峰分析，以此实现LPC阶数对不同语音种类语音帧的自适应。考虑到Praat具有操作界面友好、所占磁盘空间小、可移植性好、更新及时、功能强大等优点（尚春雨，2016），同时也是为广大汉语方言研究者熟悉和喜爱的语音学分析平台，本文直接在通用的Praat语音学实验平台而非Pytorch机器学习框架上实现上述神经网络模型以及后续的自适应LPC共振峰提取过程，以保证本文所提出的方法具有低成本、用户友好、容易上手和复现等优点。

为体现本文所提出的方法在汉语方言语音学研究特别是韵母系统研究的应用前景，本研究以北京话的流摄细音开口字、潮汕方言惠来县方言点下的效摄细音开口字和流摄细音开口字为例，展示本文的模型对相应语音样本的共振峰自动化提取结果。实验结果证明，本文所提出的方法相比传统的LPC法能够更加准确地提取前高元音和后高元音段的共振峰，通过该方法所绘制的音节内共振峰的变化轨迹也较为准确地呈现了北京话中流摄细音开口字（如“优”）和效摄细音开口字（如“妖”）在一个音节内的动态声学特点。

① 例如，某些硬件设备的体积和功耗都较为庞大，无法安装在笔记本电脑上供方言田野调查者随身携带。

## 2 神经网络模型的原理

本章将简要介绍神经网络模型的原理，并展示运用该模型自动识别某一语音帧语音类型的操作流程。

神经网络也是数学模型的一种，旨在利用其中的非线性运算和可通过优化算法（如梯度下降法）不断迭代训练的参数，对任何复杂且在多数情况下无解析解的映射关系进行建模。利用数学模型来描述一组数据到另一组数据的映射关系是一种常见的数据分析方法。例如，当输入数据 $x$ 为 $\{1, 2, 3, 4\}$ ，输出数据 $y$ 为 $\{3, 5, 7, 9\}$ 时，采用线性回归模型即可描述 $x$ 到 $y$ 的映射关系，该线性回归模型在数学上有解析解，即“ $y=2x+1$ ”。这个解析解不仅能用于描述已知数据的映射关系，还能够预测未知数据的映射结果。例如，输入数据 $x$ 为“5”，根据线性回归模型，输出结果 $y$ 为“11”。对于本任务而言，输入数据为某个语音帧的声学特征（如LPC或MFCC（Mel frequency cepstral coefficients，梅尔频率倒谱系数）等），输出数据为表示该语音帧所属语音类型的一个类别标签。在这种情况下，输入与输出间的映射关系难以通过有解析解的数学模型来刻画，而神经网络模型更适合处理此类问题。一个神经网络模型通常包含多层以矩阵乘法运算为基础的线性运算模块，每一层之间通过非线性函数连接。非线性函数使得神经网络模型可以拟合任意复杂和抽象的映射关系，但该功能的实现要求神经网络的层数足够深、参数量足够大，且结构和优化方法设计合理。更详细的神经网络原理介绍可参看Alpaydin（2021）。

一个神经网络模型包含一系列与输入数据或中间运算结果相作用的参数。本文所构建的神经网络以某一语音帧的声学特征为输入，输出该语音帧的语音类型识别结果。神经网络的初始参数值往往是随机的，如果以一个未训练的神经网络对输入数据进行数学运算，无法得到符合预期的结果。因此，需要通过一个适当的优化算法来不断调整神经网络中的参数，使神经网络能够适应各种情况下的输入特征，为每一个输入数据输出尽可能正确的语音类型预测值。具体的优化算法通常采用梯度下降法和Adam算法（Kingma et al., 2014）等。神经网络模型完成训练后，需要对模型进行测试以评估所开发的神经网络模型在某特定任务上的性能。总的来说，一个神经网络模型的开发与实现至少要包含训练和测试两个步骤，这就要求数据集至少要被划分为训练集和测试集两部分。训练集用于优化神经网络的参数值，使得神经网络能够较好地拟合训练集中的数据。测试集中的数据不能参与神经网络的训练，因为一个训练良好的神经网络应当具有良好的泛化能力，不仅能较好地拟合已知数据，还要能以较高的正确率预测未知数据的分类结果。由于测试集中的输入数据和训练集通常是同分布的，即测试集和训练集中的数据具有基本相同或相似的统计学特征，一个性能良好的神经网络模型只要能在训练集上输出较好的预测结果，在测试集上也应当预测出基本正确的语音类型识别结果。另外，为保证神经网络

络具有足够的泛化能力，应确保神经网络在训练阶段学习到尽可能多样且复杂的数据。这与人类的学习过程类似，比如，考生在考试前接触和掌握的学习资料越丰富，越有可能在考试中取得较好的成绩。因此，在神经网络的开发与训练开始之前，获取充分的相关数据十分重要。

### 3 汉语方言音频数据来源和预处理

#### 3.1 数据来源

如上文所述，开发一个神经网络模型需要大量的数据样本。本文旨在通过神经网络自动识别每一语音帧的语音类型，以合适的 LPC 阶数完成共振峰提取，最终服务于汉语方言的语音学研究。本文所采集的语音数据以汉语方言为主，包括来自全国各地不同方言分区的 10 个方言点下 12 名不同性别发音人的单字音频。录音材料以《中国语言资源调查手册·汉语方言卷》（2015）中的声母表和韵母表为主。方言音频数据的详细信息列于表 1。

表 1 汉语方言音频数据来源与样本量

| 方言点 | 方言分区 | 发音人性别 | 单字样本量 | 语音帧样本量 |
|-----|------|-------|-------|--------|
| 北京  | 北京官话 | 女     | 200   | 6,405  |
| 苍南  | 吴方言  | 男     | 194   | 6,008  |
|     |      | 男     | 220   | 8,045  |
| 重庆  | 西南官话 | 女     | 220   | 9,153  |
|     |      | 女     | 220   | 10,720 |
| 惠来  | 闽方言  | 女     | 201   | 6,931  |
| 茂名  | 粤方言  | 女     | 192   | 5,828  |
| 宁波  | 吴方言  | 女     | 198   | 6,419  |
| 瑞安  | 吴方言  | 女     | 184   | 6,592  |
| 石林  | 西南官话 | 女     | 194   | 5,070  |
| 运城  | 晋方言  | 女     | 194   | 6,294  |
| 漳州  | 闽方言  | 女     | 169   | 4,522  |
| 合计  |      |       | 2,386 | 81,987 |

注：单字样本主要参考《中国语言资源调查手册·汉语方言卷》（2015）

#### 3.2 音频录制、标注和声学特征提取

##### 3.2.1 音频录制和标注

本文的汉语方言音频在隔音良好、回声较小的录音棚中采集。录制设备包括 AKG C544 L 头戴式麦克风和 Komplete Audio 6 外置声卡，使用 Audacity 作为录音软件，声道设置为单声道，采样率设置为 22050 Hz。录制的音频首先在 Praat 中切

分为单字文件，以国际音标符号按照严式记音原则对各音段类型进行标注，无语音信号的静音段标注为“silent”或空白字符串。随后以 25 ms 帧长和 12.5 ms 帧移和对音频进行分帧，并将分割出的各语音帧单独保存为 wav 文件。各方言点不同发音人的有效语音帧样本量同样列于表 1。

### 3.2.2 声学特征提取

针对自然语言处理的神经网络模型不直接输入语音帧的波形信号，因为时域的波形信号包含大量冗余信息，不仅会降低神经网络的训练效率，还可能影响神经网络的预测准确率。因此，本文将神经网络模型的输入设置为每一语音帧的声学特征，具体表征为当前语音帧的 MFCC、LPC，以及相对于整个音节内语音信号的平均强度。其中，MFCC 的阶数设置为 16，LPC 阶数为 16、24、32。此时，输入到神经网络的数据就是一个维数为  $16+16+24+32+1=89$  的特征向量。另外，考虑到语音信号的时序性，即当前语音帧所蕴含的特征信息不仅与其本身的声学特征有关，还可能受到之前时刻语音帧的影响。因此，本文将当前帧的前两个语音帧的声学特征向量一并拼接到当前帧的特征向量中，此时输入的数据就变成了一个维数为  $89 \times 3=267$  的特征向量，神经网络模型将据此输出当前帧语音类型的预测结果。最后，使用 z-score 对每一语音帧的特征向量进行标准化处理，将处理后的值平移到  $[0, 1]$  的区间内，提升神经网络训练效果。

## 4 神经网络的结构及基于 Praat 实验平台的模型

### 4.1 神经网络的结构

常见的神经网络结构包括全连接网络、卷积神经网络、循环神经网络，以及近年来受到越来越多关注的基于 Transformer (Vaswani et al., 2017) 的网络结构。复杂的神经网络结构和更大的模型参数量有利于提升模型的学习能力，使其能够拟合更加复杂的映射关系。但这通常会带来更高的计算复杂度及更长的模型训练时间，往往需要成本高昂的硬件设备提供足够的算力，不利于开展实时、低成本的汉语方言实验语音学研究。为解决这一问题，本文直接采用 Praat 内置的神经网络模块搭建了一个可用于识别各语音帧语音类型的神经网络模型。Praat 中内置的神经网络模块实际上是一个包含 3 层全连接层的网络结构，本文将 3 层全连接层的特征数分别设置为 2048、267、4。每一全连接层对输入到该层的特征向量先进行线性变换，再施加非线性激活函数 (Non-linear activation function)，用数学公式可以表示为

$$y_i = \text{Nonlinear}[Wx_i] \quad (1)$$

其中， $i$  为神经网络的层数序号； $x_i$  和  $y_i$  分别为第  $i$  层网络的输入和输出特征向量；

$W$ 为权重矩阵，可以通过优化算法在神经网络的训练阶段动态调整，使得神经网络的输出尽可能接近预期结果； $\text{Nonlinear}[\cdot]$ 为非线性函数，通常以Sigmoid函数或ReLU函数（Pedamonti, 2018）实现。

整个神经网络最终输出一个维数为4的特征向量，其中的4个特征值分别对应所输入语音帧4种语音类型的预测概率，取概率最大的类别作为神经网络对当前输入语音帧的语音类型预测结果。另外，本文还规定，如果一个语音帧被神经网络预测为“清音段”的概率大于0.2，该帧语音信号就属于清音段。这是因为受到发音生理因素的限制，静音段到有声段之间的语音信号是不稳定的，就汉语方言实验语音学研究的一般情况而言，通常也不会对不稳定的音频段作声学分析。本文所构建的神经网络模型展示于图1。

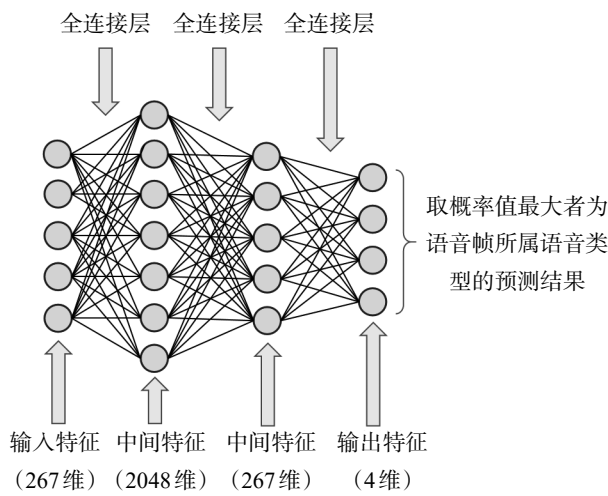


图1 本文构建的神经网络模型

## 4.2 基于Praat构建神经网络模型

Praat是一个易于上手且功能强大的语音学实验平台，已被广大研究者应用于许多实验语音学相关的研究工作中。新版本的Praat平台<sup>①</sup>内置了一个神经网络模块，可以搭建一个至多3层的全连接神经网络。如图2所示，打开Praat软件后，首先依次点击New > Feedforward neural networks > Advanced > Create PatternList，再依次点击New > Feedforward neural networks > Advanced > Create Categories，即可生成两个对象“PatternList”和“Categories”。其中，“PatternList”用于存储神经网络的输入特征，其行数为输入数据的样本量，列数在本文中为267，即每一行代

<sup>①</sup> 本文所用Praat版本为6.4.34。

表的是某项输入样本的声学特征。“Categories” 储存了各数据样本对应的正确分类标签，在本文中为4类可能的语音类别：“清音”“前高元音”“后高元音”和“其他浊音”。

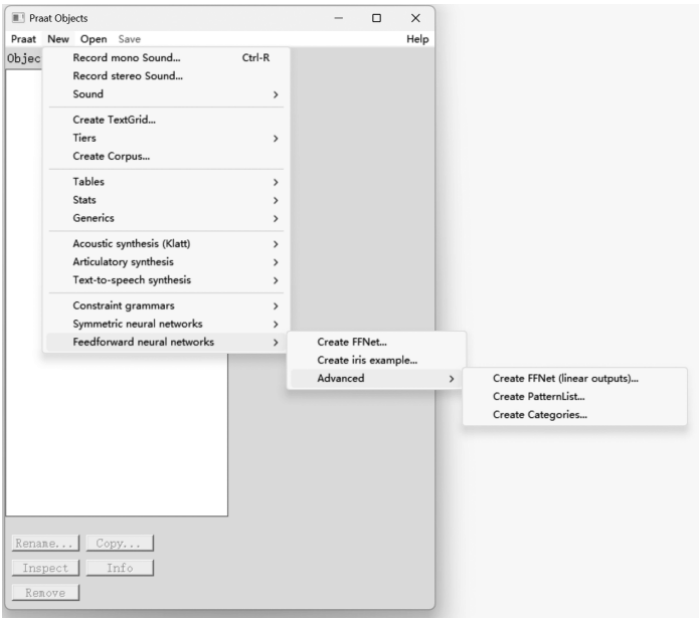


图2 创建“PatternList”和“Categories”对象

随后，如图3所示，同时选中刚刚生成的“PatternList”和“Categories”对象，并点击右侧的“To FFNet...”，在弹出窗口中设置好参数，即可实现基于Praat平台的神经网络模型的搭建。

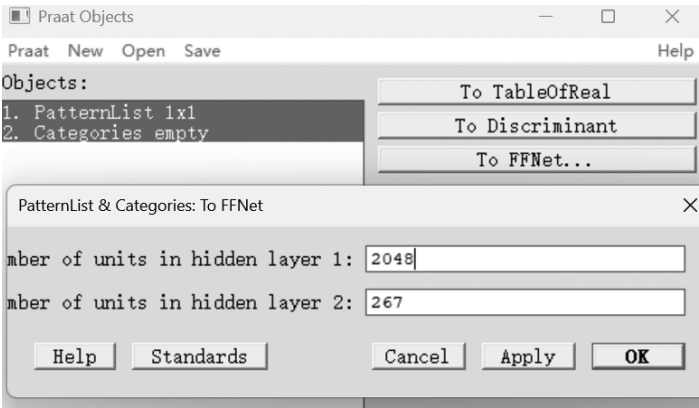


图3 创建“FFNet”对象后完成神经网络模型的搭建

## 5 神经网络的训练和测试

### 5.1 数据集划分

如前文所述，神经网络模型的开发至少需要训练和测试两个步骤，原始数据集要被划分为两部分，一部分用于神经网络的训练，称为训练集；一部分用于测试，称为测试集。被划分到测试集的数据不能参与神经网络的训练。一般而言，训练集和测试集样本的划分比例为 8:2。本文将来自北京（北京官话）和惠来（闽方言）两个方言点的样本划分到测试集中，将来自其余方言点的样本划分为训练集。

### 5.2 训练参数设置

如图 4 所示，在 Praat 中，同时选中上文所生成的“PatternList”“Categories”和“FFNet”对象，再点击右侧的“Learn...”，即可开始神经网络的训练。首先需在弹出窗口中设置 3 个参数。第一个参数指最大训练周期数。神经网络的训练是一个需要进行多次迭代的过程，迭代轮数通过“epoch”（周期）这一概念来量化。具体而言，训练集中所有数据样本均被送入神经网络进行训练，完成一次训练称为一个 epoch。本文将最大 epoch 数设置为 2,400。第二项参数表示在相邻两次迭代训练过程中，代价函数（cost function）下降值低于该参数的设定值时，需要提前结束训练，本文保留 Praat 预设的默认值。第三项参数是指需要选择哪一个函数作为代价函数。由于本文搭建的神经网络是一个分类而非回归任务，因此选择交叉熵（cross-entropy）作为代价函数。代价函数描述神经网络的预测结果与正确预测结果之间的偏差，而神经网络训练的目的正是使代价函数值最小化，尽可能减小预测值与真实值之间的误差。有关代价函数（通常也称为“损失函数”，loss function）更详细的介绍可参看 Mao et al. (2023)。

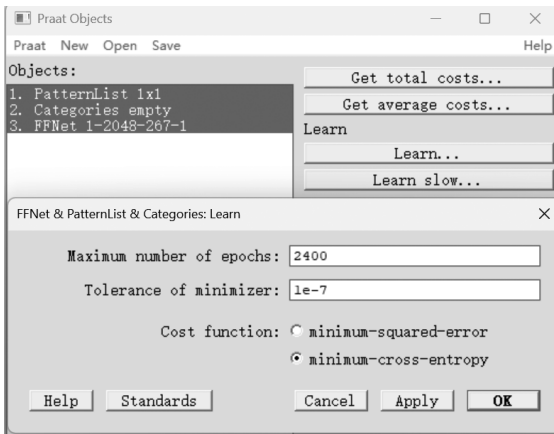


图 4 神经网络的训练参数设置

5.3 神经网络在测试集上的分类准确率

完成神经网络训练后，需要使用测试集进行测试，以检验训练模型是否具备足够的泛化能力，并考察其在语音帧语音类型自动识别任务上是否达到预期水平。本文将神经网络的分类准确率定义为正确分类样本数与测试集样本总数的百分比。经过计算，本文设计并训练的神经网络模型在测试集样本（即北京和惠来两个方言点的音频数据）上的语音帧语音类型识别准确率为89.67%，表明训练模型能够基本完成对一段语音信号中各语音帧语音类型的自动识别任务。表2所示为北京话“优”字音频中各语音帧的语音类型识别结果，其中字段“voiceless”“frontHighVowel”“backHighVowel”和“otherVowel”分别对应于“清音段”“前高元音段”“后高元音段”和“其他浊音段”。从表2所示的实验结果可以看出，本文的神经网络对该音频中55个语音帧的语音类型预测结果与人工标注的语音类型基本吻合，只有个别语音帧在语音类型之间的过渡段上存在偏差，如从最开始的静音段到音节起始的有声段之间的过渡段，以及从前高元音段到后高元音段之间的过渡段等。另外，根据实际用法，北京话中“优”字的读音可记为[iou]，该音节是以前高元音[i]起头，后高元音[u]收尾的复元音音节。本文训练的神经网络对该音节中各语音帧语音类型的预测结果在时间上从前往后依次是静音段、前高元音段、后高元音段和静音段，符合实际发音情况。

表2 本文训练的神经网络对北京话“优”字各语音帧的语音类型自动识别结果

| 语音帧序号 | 神经网络预测的语音类型    | 人工标注的语音类型      |
|-------|----------------|----------------|
| 1     | voiceless      | voiceless      |
| 2     | voiceless      | voiceless      |
| 3     | voiceless      | voiceless      |
| 4     | voiceless      | voiceless      |
| 5     | voiceless      | voiceless      |
| 6     | voiceless      | voiceless      |
| 7     | voiceless      | voiceless      |
| 8     | voiceless      | voiceless      |
| 9     | voiceless      | voiceless      |
| 10    | voiceless      | frontHighVowel |
| 11    | voiceless      | frontHighVowel |
| 12    | voiceless      | frontHighVowel |
| 13    | frontHighVowel | frontHighVowel |
| 14    | frontHighVowel | frontHighVowel |
| 15    | frontHighVowel | frontHighVowel |
| 16    | frontHighVowel | frontHighVowel |
| 17    | frontHighVowel | frontHighVowel |

续表

| 语音帧序号 | 神经网络预测的语音类型    | 人工标注的语音类型      |
|-------|----------------|----------------|
| 18    | frontHighVowel | frontHighVowel |
| 19    | frontHighVowel | frontHighVowel |
| 20    | frontHighVowel | frontHighVowel |
| 21    | frontHighVowel | frontHighVowel |
| 22    | backHighVowel  | backHighVowel  |
| 23    | backHighVowel  | backHighVowel  |
| 24    | backHighVowel  | backHighVowel  |
| 25    | backHighVowel  | backHighVowel  |
| 26    | backHighVowel  | backHighVowel  |
| 27    | backHighVowel  | backHighVowel  |
| 28    | backHighVowel  | backHighVowel  |
| 29    | backHighVowel  | backHighVowel  |
| 30    | backHighVowel  | backHighVowel  |
| 31    | backHighVowel  | backHighVowel  |
| 32    | backHighVowel  | backHighVowel  |
| 33    | backHighVowel  | backHighVowel  |
| 34    | backHighVowel  | backHighVowel  |
| 35    | backHighVowel  | backHighVowel  |
| 36    | backHighVowel  | backHighVowel  |
| 37    | backHighVowel  | backHighVowel  |
| 38    | backHighVowel  | backHighVowel  |
| 39    | backHighVowel  | backHighVowel  |
| 40    | backHighVowel  | backHighVowel  |
| 41    | backHighVowel  | backHighVowel  |
| 42    | backHighVowel  | backHighVowel  |
| 43    | backHighVowel  | backHighVowel  |
| 44    | backHighVowel  | backHighVowel  |
| 45    | backHighVowel  | backHighVowel  |
| 46    | voiceless      | backHighVowel  |
| 47    | voiceless      | backHighVowel  |
| 48    | voiceless      | backHighVowel  |
| 49    | voiceless      | voiceless      |
| 50    | voiceless      | voiceless      |
| 51    | voiceless      | voiceless      |
| 52    | voiceless      | voiceless      |
| 53    | voiceless      | voiceless      |
| 54    | voiceless      | voiceless      |
| 55    | voiceless      | voiceless      |

## 6 基于自适应LPC阶数的共振峰提取

上文展示的神经网络模型成功实现了以帧为单位对所输入的单字音频在不同时刻下语音类型的自动化标注任务，但未提取音节各语音帧的共振峰，而基于神经网络模型的语音类型自动化标注正是为了服务后续共振峰的提取。被识别为前高元音的语音帧，需要在LPC阶数较低的情况下提取共振峰，以防“虚假共振峰”；被识别为后高元音的语音帧，需要在LPC阶数较高的情况下提取共振峰，避免出现“共振峰合并”；被识别为其他浊音的语音帧，LPC阶数则应设置在适中范围内。具体来说，本文在对前高元音段、后高元音段、其他浊音段的语音帧进行基于LPC算法的共振峰提取时，LPC阶数分别设置为16、30和22，而对于被识别为“清音段”的语音帧则不提取共振峰。

图5展示了不同阶数对北京话“一”“五”“优”“妖”单字音频的共振峰提取结果。共振峰提取结果以带箭头轨迹图的形式呈现在二维坐标系中，其中横轴为F2，纵轴为F1，原点位于右上角，轨迹上的点对应某语音帧的共振峰计算结果。F1大于1300 Hz或小于150 Hz、F2大于3100 Hz或小于250 Hz的数据点视为噪声点，在绘制轨迹图之前移除。前三行表示基于burg算法（Andersen, 1974）设定LPC阶数分别为16、24、30时，计算得到的共振峰轨迹，第四行的共振峰轨迹则基于Robust Linear Prediction（RLP）算法（Lee, 1988）计算，最后一行为本文提出的神经网络与自适应LPC阶数法的共振峰提取结果。

由图5可见，基于传统LPC共振峰提取算法的计算结果对所选取的LPC阶数非常敏感。就单元音音节“一”而言，当LPC的阶数较低时（16阶），如图5第一行所示，算法尚且能够输出基本正确的共振峰。“一”在北京话中读作前高元音[i]，共振峰应该出现在声学元音图的左上角，意即F1较低，集中在400 Hz以下，同时F2较高，女性可达将近3000 Hz。但随着LPC阶数的增加，介于F1和F2之间的“虚假共振峰”逐渐显现，如24阶LPC、30阶LPC和RLP算法都错误地提取出介于1000—2000 Hz之间的虚假共振峰F2，这显然与客观事实不符，前高元音[i]的F2通常不可能低于2000 Hz。相反，如图5最后一行所示，本文所提出的神经网络识别语音类型与自适应LPC阶数的算法能够准确地提取北京话“一”字音频的共振峰。北京话“一”字的读音是典型的单元音而非复元音，所提取出的共振峰不仅在频率值上符合客观事实，而且在声学元音图上分布也非常集中。当LPC阶数较低时（16阶），如图5第一行所示，对于F1和F2相隔较远的前高元音，基于该LPC阶数的共振峰提取算法能得到基本正确的结果。但如果输入后高元音，通过比较“五”字共振峰提取结果发现，后高元音的F1和F2相隔太近（仅相差100—200 Hz），容易出现“共振峰合并”现象。此时，算法无法分辨某些语音帧中的F1和F2，而将其合并为一个共振峰，随后将频率更高的F3识别为F2，导致共振峰计算

错误。如图5第一行第二列所示，不少语音帧的F2计算结果高于2000 Hz，这显然不符合客观事实。北京话单元音音节中后高元音[u]的F2一般不超过600 Hz，不可能达到2000 Hz。增大LPC阶数能够减少“共振峰合并”错误的发生，但会增加“虚假共振峰”的出现。因此，传统的LPC算法很难找到合适的LPC阶数，既适合前高元音也适合后高元音的共振峰提取。相较之下，本文所提出的方法在前高元音和后高元音音节上都实现了较为准确的共振峰提取，北京话单字音节“一”和“五”的共振峰轨迹点在声学元音图上的分布都较为集中，很好地展现了两个单元音特性。

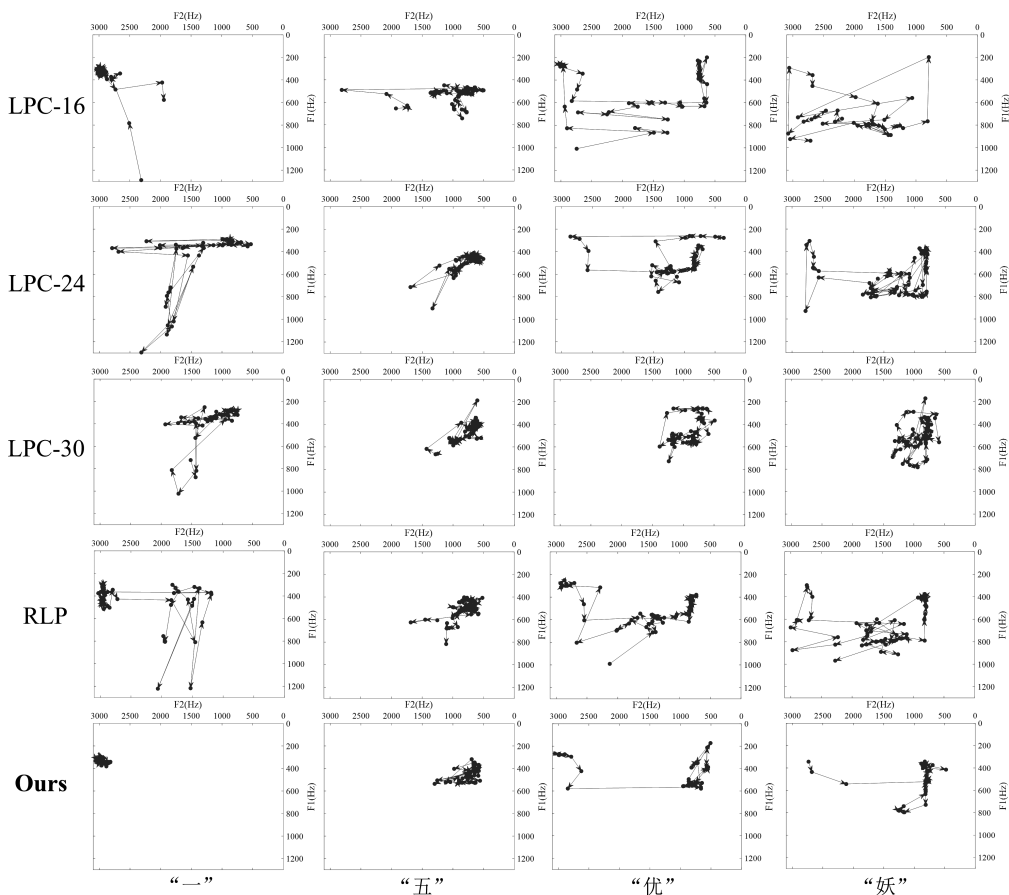


图5 不同算法对北京话“一”“五”“优”“妖”单字音频的共振峰轨迹提取结果

当音节的韵母为结构更加复杂的复元音，尤其是复元音组合中既包含前高元音，又包含后高元音（如北京话“妖”“优”）时，传统算法对共振峰识别的准确率严重下降。如图5第三、四列所示，由于“虚假共振峰”和“共振峰合并”现象

频繁发生，传统算法提取出的共振峰频率值不仅与真实值产生几百乃至上千 Hz 的误差，而且在声学元音图上描绘出的共振峰轨迹也较为混乱，研究者难以辨别复元音音节中韵母的共振峰变化情况。相较之下，本文所提出的方法在处理复元音共振峰提取时依然能得出较为准确的结果，描绘出的共振峰轨迹也较为清晰。其原理在于，传统的基于固定阶数 LPC 的共振峰提取算法难以同时兼顾前高元音和后高元音的提取，但本文首先以帧为单位，通过神经网络模型自动识别输入音频每一语音帧的语音类型，根据识别结果动态调整适用于不同语音帧共振峰提取的 LPC 阶数，从而实现适应各种元音类型的共振峰提取。

根据本文对北京话“优”和“妖”的共振峰提取轨迹可见，共振峰频率点先相对集中分布在声学元音图的左上角，即 F2 较高，F1 较低的区域，随后逐渐向右上角（即 F2 和 F1 都较低的区域）移动。还可以观察到，前高元音区域的共振峰轨迹点较为稀疏，后高元音区域较为密集。上述实验结果不仅与北京话“妖”和“优”的实际音节结构相吻合，也很好体现了汉语方言音节中介音的一般特点，即介音[i]的声学强度和稳定度在整个音节中占比通常较低。

最后，观察发现，北京话中“优”和“妖”单字音节内靠近音节起始和音节末尾的共振峰轨迹点分布较为集中，而在音节起始的前高元音逐渐向后高元音变化的过渡段中，共振峰轨迹点不稳定。一方面，这种现象可以通过言语交际的“清晰原则”（陈忠敏，2019）来解释，即从功能主义的角度来看，语言是传达信息的工具，只有听者清楚地听到说者所发出的语音信号，才可能实现清晰、无障碍、无歧义的言语交际。一般来说，位于声学元音格局图周边的元音（即 1—8 号标准元音）最容易被人类清晰感知，因为它们在特定语言或方言的音系结构里具有最大的声学距离。这种感知优势可能进一步引起“掩蔽效应”，即强音征遮蔽弱音征（陈忠敏，2022）。具体来说，在一个韵母为复元音结构的音节里，若韵母的起始音段和结尾音段都是周边元音，在声学上是强音征，那么这两个音段之间的过渡段则是相对而言的弱音征。根据掩蔽效应，具有弱音征的音段更难被听者感知，在言语交际中的重要性相对较低。因此，说话者倾向于将有限的发声能量集中在位于声学元音图周边的元音上，发出具有相对稳定的声学特征的语音。而对于不那么重要的过渡段，说话者通常不会投入额外的发音控制来维持其声学稳定性。另一方面，语音的量子理论（Quantal theory）（Stevens, 1989）同样可以解释这一实验现象。简单来说，量子理论揭示了语音生理和声学的不对应性。在一些情况下，发音生理上的剧烈变化并不会在声学上带来明显的改变；但在另一些情况下，发音生理上的微小变化就能引起声学上的明显改变。前者为“稳定的发音段”，后者为“不稳定的发音段”。人类语言需要选择位于稳定发音段的音素作为其音系组成部分。Stevens 根据发音部位和声学参数推导得出，元音[u]正好属于稳定发音段（陈忠敏，2019）。本

文所考察的北京话“妖”“优”正好包含前高元音[u]和后高元音[u],因此它们相对于其间过渡段的音段应当具有稳定的声学表现,这与图5所示的实验结果吻合。

## 7 结语

元音的声学特征主要体现在共振峰这一项声学特征上,准确地提取语音信号的共振峰是开展汉语方言韵母系统实验语音学研究的基础。然而,对于实际的语音信号而言,传统的基于固定阶数的LPC共振峰提取算法往往会遇到诸如“虚假共振峰”或“共振峰合并”等问题,更易错误识别前高元音和后高元音的共振峰。究其原因,LPC阶数会影响传统共振峰算法的提取结果。具体来说,LPC阶数较低时,有利于前高元音的共振峰提取,但在后高元音上会产生“共振峰合并”现象;LPC阶数较高时,有利于后高元音的共振峰提取,但在前高元音上会出现“虚假共振峰”错误。如果一个音节同时包含前高元音和后高元音,情况则更为复杂。此时,难以找到通用的LPC阶数使其既能实现前高元音又能适应后高元音的共振峰提取。针对这一困难,本文设计并训练了一个神经网络模型,以帧为单位,识别输入音频中每一语音帧的语音类型,根据所识别出的语音类型,为每一语音帧动态地设定用于共振峰提取的LPC阶数,由此实现自适应LPC阶数的共振峰提取。这不仅极大地提高了共振峰提取的准确率,也提升了汉语方言语音学研究效率。本文的神经网络模型以及后续的共振峰提取在学界通用的Praat实验平台上实现,避免了昂贵的硬件设备和繁琐的运行环境配置过程,该方法有望在业内进行普及和推广。

实验证明,本研究开发的神经网络模型在汉语方言语音学研究领域有着广阔的应用前景和发展潜力。该模型不仅适用于诸如汉语方言语种识别等宏观的汉语方言学自动化分析任务,对于底层的物理、声学层面的汉语方言语音信号分析也大有裨益。高效、准确的共振峰提取有助于研究者清楚地观察到复音节内共振峰的动态变化轨迹及其展现的语音学现象。此外,文章从言语交际的清晰原则和量子语音的角度对本文所观察到的语音现象进行了较为合理的解释。

## 参考文献

- Alpaydin E, 2021. Machine learning[M]. Cambridge: MIT Press.
- Andersen N, 1974. On the calculation of filter coefficients for maximum entropy spectral analysis[J]. Geophysics, 39(1): 69-72.
- Jiang B, Dong Q, Liu G, 2024. A method of phonemic annotation for Chinese dialects based on a deep learning model with adaptive temporal attention and a feature disentangling structure[J]. Computer speech & language, 86: 101624.

- Kingma D P, Ba J, 2014. Adam: a method for stochastic optimization[J]. arXiv preprint, arXiv: 1412.6980.
- Krizhevsky A, Sutskever I, Hinton G E, 2012. ImageNet classification with deep convolutional neural networks[C]//Advances in neural information processing systems 25 (NIPS 2012). Red Hook: Curran Associates: 1097-1105.
- Ladefoged P, 1996. Elements of acoustic phonetics[M]. 2nd ed. Chicago: The University of Chicago Press.
- Lee C H, 1988. On robust linear prediction of speech[J]. IEEE transactions on acoustics, speech, and signal processing, 36(5): 642-650.
- Mao A, Mohri M, Zhong Y, 2023. Cross-entropy loss functions: theoretical analysis and applications [C]//International conference on machine learning. Honolulu: PMLR: 23803-23828.
- Pedamonti D, 2018. Comparison of non-linear activation functions for deep neural networks on MNIST classification task[J]. arXiv preprint, arXiv:1804.02763.
- Stevens K N, 1989. On the quantal nature of speech[J]. Journal of phonetics, 17(1-2): 3-45.
- Vaswani A, Shazeer N, Parmar N, et al., 2017. Attention is all you need[C]//Advances in neural information processing systems 30 (NIPS 2017). Red Hook: Curran Associates: 5998-6008.
- 陈忠敏, 2019. 论言语发音与感知的互动机制[J]. 外国语, 42 (6): 2-17.
- 陈忠敏, 2022. 音类的音征与语音演变[J]. 语言科学, 21 (6): 561-579.
- 付婧, 李艳梅, 陶卫国, 等, 2020. 基于深度神经网络的成都话识别研究[J]. 西华师范大学学报 (自然科学版), 41 (4): 440-444.
- 教育部语言文字信息管理司, 中国语言资源保护研究中心, 2015. 中国语言资源调查手册·汉语方言卷[M]. 北京: 商务印书馆.
- 尚春雨, 2016. 英语语音课可视化研究: 以 Praat 软件为例[J]. 无锡职业技术学院学报, 15 (2): 35-37.
- 吴宗济, 林茂灿, 1989. 实验语音学概要[M]. 北京: 高等教育出版社.
- 颜为之, 2023. 基于语料库的江西境内客、赣方言自动识别研究[D]. 南昌: 江西师范大学.
- 杨迪一, 2023. 基于深度学习的陕北方言语音识别系统设计[D]. 延安: 延安大学.
- 赵泽彬, 兰亮, 姜丹, 等, 2024. 基于迁移学习的小样本语言语音识别研究[J]. 北京印刷学院学报, 32 (6): 27-34.

## 通信地址

200433 上海市 复旦大学中国语言文学系

(责任编辑: 杨瑞雪)

# 中外学术期刊英文摘要认知难度历时变化研究：基于信息熵与依存距离的量化分析<sup>\*</sup>

河南师范大学 刘国兵 侯佳欣

**提 要：**本研究基于自建语料库，以信息熵与依存距离为测量指标，分别评估摘要的认知编码难度与认知解码难度，对比分析2000—2024年中外应用语言学期刊英文摘要认知难度的历时变化。研究发现，在认知编码难度层面，国内外期刊均呈现上升趋势，国际期刊的认知编码难度显著高于国内期刊，二者之间存在显著正相关但不存在格兰杰因果；在认知解码难度层面，国内外期刊均未呈现统计学意义上的显著变化，但两者在考察期内的变化趋势存在差异。国内期刊呈波动上升态势，国际期刊则平稳中略有下降，二者之间存在显著负相关，但未发现格兰杰因果关系。本研究揭示了2000—2024年中外应用语言学期刊英文摘要中认知编码难度与认知解码难度的历时演变特征，为进一步了解中外学术论文摘要的认知难度差异提供了新的实证依据。

**关键词：**认知编码难度；认知解码难度；信息熵；依存距离

## A Diachronic Study Based on Information Entropy and Dependency Distance for Cognitive Difficulty of English Abstracts in Chinese and International Journals

Guobing Liu and Jiaxin Hou

**Abstract:** Based on a self-built corpus, this study uses information entropy and dependency distance as measurement indicators to evaluate the cognitive encoding

<sup>\*</sup> 本文系河南省哲学社科项目“汉英学术论文评价性语言使用特征对比研究”（项目编号：2024BYY012）及2025年度河南省高等学校哲学社会科学创新团队支持计划“人工智能时代的语料库与语言大数据研究”（项目编号：2025-CXTD-06）的阶段性研究成果。

侯佳欣为本文通信作者。

作者贡献：

刘国兵：选题构思、研究方法、讨论结论、修改润色、字数占比（60%）；

侯佳欣：数据收集、数据分析、初稿撰写、字数占比（40%）。

difficulty and cognitive decoding difficulty of abstracts. It conducts a comparative analysis of the diachronic changes in the cognitive difficulty of English abstracts in Chinese and international applied linguistics journals from 2000 to 2024. The findings reveal that, in terms of cognitive encoding difficulty, both domestic and international journals exhibit an upward trend, and the cognitive encoding difficulty of international journals is significantly higher than that of domestic journals. While a significant positive correlation exists between the two, they do not constitute a Granger causality. In terms of cognitive decoding difficulty, neither domestic journals nor international journals show statistically significant changes, but their trends during the study period differ. Domestic journals show a fluctuating upward trend, while international journals show a slight decline with relative stability. There is a significant negative correlation between the two, but it does not constitute a Granger causality. This study reveals the diachronic evolution characteristics of cognitive encoding difficulty and cognitive decoding difficulty in English abstracts of Chinese and international applied linguistics journals from 2000 to 2024, and provides new empirical evidence for further understanding the differences in cognitive difficulty between Chinese and international academic paper abstracts.

**Keywords:** cognitive encoding difficulty; cognitive decoding difficulty; information entropy; dependency distance

## 1 引言

摘要是学术论文的核心组成部分，是读者了解研究内容的首要窗口。作为一种兼具信息性与劝说性的学术体裁，摘要不仅承担准确概括研究内容的功能，更在吸引读者兴趣和促进学术传播中发挥关键作用（Jiang et al., 2017）。英文摘要作为国际学术交流与合作的桥梁和媒介，近年来受到学者的广泛关注（曹雁等，2017）。以往研究多集中于摘要的修辞结构（Martín, 2003；牛桂玲，2013；鞠玉梅，2020；陈庆斌，2021）和元话语（Jiang et al., 2017；李芝等，2020）等语言表层特征，对认知维度的探讨略显不足。事实上，摘要的撰写与理解分属不同的认知过程，作者需进行高效的信息编码，读者则面临解码复杂性带来的认知挑战。这一认知双过程视角为摘要研究提供了新的切入点。

本研究以2000—2024年中外应用语言学期刊英文摘要为语料，选取信息熵和依存距离作为测量指标，分别用于评估摘要的认知编码难度与认知解码难度。研究重点聚焦于该时期内中外应用语言学期刊英文摘要认知编码难度与认知解码难度的

历时演变特征，旨在揭示该学科英文摘要认知难度的动态发展规律，为学术语篇计量研究提供实证依据与有益借鉴。

## 2 文献综述

论文摘要作为学术交流的重要媒介，其认知难度直接影响读者的理解效率和学术传播效果。现有关于摘要难度的研究经历了从表层语言特征分析到深层认知机制探索的发展过程。早期研究主要依赖传统可读性公式（如FRE、SMOG等），通过词长、句长等表层指标评估文本难度（Flesch, 1948），普遍发现学术摘要阅读难度较高（Ante, 2022; Gazni, 2011; Lei et al., 2016; Wang et al., 2022）。然而，以上研究仅关注语言的表层特征，难以捕捉学术语篇特有的认知复杂性。随着研究推进，学者们逐渐转向关注摘要的语言复杂度，如词汇丰富度、句法复杂度等语言特征（宋瑞梅等, 2019; 刘国兵等, 2023）。该类研究主要从读者理解的角度分析摘要的认知解码难度，尚未建立系统的摘要认知难度评估框架。

在此背景下，计量语言学的发展为认知难度研究提供了新的突破点。方昱等（2021）系统总结了两类认知指标：一类基于工作记忆负荷，如依存距离（Dependency Distance）、存储成本（Storage Cost）和整合成本（Integration Cost）；另一类基于经验预测，如熵（Entropy）、惊异值（Surprisal）和概率配价（Probabilistic Valency）。在这些指标中，依存距离和熵因其独特的理论价值而备受关注。具体而言，依存距离作为衡量句法复杂度的重要指标，能够测量文本的认知解码难度，有效揭示读者在理解句子过程中面临的认知挑战（Liu, 2008; Liu et al., 2017）。依存距离越大，文本认知解码难度越大。熵作为信息论的核心概念，通过量化词汇分布的不确定性，可以反映作者在文本生成和信息编码中所需的认知加工努力（Juola, 2013; Sayood, 2018），测量文本的认知编码难度。熵值越大，文本认知编码难度越大。这两类指标分别从理解和生成两个维度为文本认知难度测量提供了理论基础和方法支持。

值得注意的是，现有研究多聚焦第一类指标，对基于经验预测的认知难度测量指标缺乏深入探讨（方昱等, 2021），将两类指标整合应用的研究更为少见。Zhao等（2023）提出一个具有开创性的二分框架，将摘要认知难度解构为认知编码难度与认知解码难度两个维度，并分别采用信息熵和依存距离进行量化，深入探讨摘要认知难度的学科差异和历时变化。这一框架为全面理解摘要认知加工的双重过程提供了重要理论支持。随后，Wu等（2025）应用该框架考察了两个时间段（1981—1985年和2011—2015年）和两种研究范式（定量和定性）应用语言学论文认知编码难度和认知解码难度的历时变化。

尽管已有研究推动了该领域的发展，摘要认知难度的研究仍存在以下几点局

限。一是多数研究侧重于认知解码难度，对认知编码难度的关注相对不足；二是方法上多局限于静态分析，缺乏长期历时追踪；三是学科内跨文化对比研究较为匮乏，难以揭示不同学术传统下认知难度的差异特征。这些不足凸显出进一步整合认知双过程视角、系统考察摘要认知难度历时演变与中外差异的必要性与紧迫性。

对此，本研究基于自建语料库，分析2000—2024年中外应用语言学学术论文摘要认知难度的历时演变与中外差异，以期为深入理解学术论文摘要认知难度变化提供新的理论视角。

### 3 研究设计

#### 3.1 研究问题

本研究自建中外应用语言学期刊摘要语料库，分析摘要认知难度的历时演变，试图回答以下两个问题：

- (1) 中外应用语言学期刊摘要认知编码难度变化呈现出怎样的历时特征？两者之间存在何种差异？
- (2) 中外应用语言学期刊摘要认知解码难度变化表现出怎样的发展趋势？两者之间存在何种差异？

#### 3.2 语料来源

本研究选取中外应用语言学领域具有代表性的国际期刊 *Applied Linguistics* (AL) 与国内期刊 *Chinese Journal of Applied Linguistics* (CJAL) 作为语料来源，自建两个语料库。为确保语料的可比性，本研究选取2000—2024年两种期刊的英文摘要，每个期刊每年选取10篇英文摘要，且长度相近（100—200词），分别建立AL语料库和CJAL语料库，语料库具体信息详见表1。

表1 语料库基本信息

| 语料库      | 文本数 | 形符数    |
|----------|-----|--------|
| AL 语料库   | 250 | 40,416 |
| CJAL 语料库 | 250 | 35,821 |

#### 3.3 研究测量指标

##### 3.3.1 信息熵

信息熵 (information entropy) 由 Shannon (1948) 提出，亦称香农熵 (Shannon entropy)，表示系统 (信源) 的不确定性，用于量化和测算文本中的信息量。系统不确定性越大，熵值越大，确定该文本所需的信息量也就越大，反之亦然。Mead

(2005) 提出, 信息量既可以表征认知加工强度, 也能反映认知负荷水平。文本信息量的增加会线性提高认知负荷, 进而提升文本处理难度。相关实证研究表明, 熵值与文本处理难度呈显著正相关 (Koch, 2009; Searle, 2013), 熵值升高直接反映信息量的增加和认知编码难度的提升 (Sayood, 2018)。作为一种在信息论中广泛使用的指标, 信息熵已被引入语言学多个研究领域, 如文化复杂性研究 (Juola, 2013; Liu, 2016; Zhu et al., 2018)、语言复杂度分析 (Juola, 2008)、词汇丰富度测量 (Rajput et al., 2018; Shi et al., 2020; Zhang, 2015) 以及文本类型特征研究 (Chen et al., 2017) 等, 显示出跨领域测量的适应性与解释力。

信息熵的基本原理是通过概率统计方法测量文本信息的不确定性, 该公式计算各种可能事件发生概率的对数与其概率乘积之和, 计算方法详见公式 (1)。

$$H(T) = - \sum_{i=1}^n p(x_i) \log_2 p(x_i) \quad (1)$$

公式 (1) 中,  $p(x_i)$  表示文本中第  $i$  个单词类型出现的概率,  $n$  是文本中单词类型的总数,  $H(T)$  为香农熵, 表示单词类型的平均信息量 (Bentz et al., 2017)。

由于本研究语料规模有限且文本较短, 直接使用公式 (1) 计算熵值可能导致结果出现偏差。为此, 引入 Zhao 等 (2023) 所采用的 Miller-Madow 熵 (Hausser et al., 2014) 来计算摘要熵值。该算法通过对香农熵计算结果进行校正来抵消文本较短造成的结果偏差, 可以有效提升计算结果的可靠性, 计算方法详见公式 (2)。

$$H(MM) = H(T) + \frac{V-1}{2N} \quad (2)$$

其中,  $H(MM)$  为 Miller-Madow 熵,  $V$  代表单词类型数,  $N$  代表单词形符数。

### 3.3.2 平均依存距离

平均依存距离 (Mean Dependency Distance, 简称 MDD) 是指句子中所有词语之间依存距离的平均值。其中, 依存距离表示句子中存在句法依存关系的两个词之间的线性距离 (Heringer et al., 1980; 刘海涛, 2009)。Liu 等 (2009) 提出了基于词序位置的依存距离计算方法: 支配词序号减去从属词序号所得之差, 如例 (1) 中 The 和 cat 的依存距离为  $3-1=2$ ; caught 和 mouse 的依存距离为  $5-8=-3$ 。

(1) **The** clever **cat** quickly **caught** a small **mouse**.

研究表明, 较长的依存距离会增加工作记忆负担, 提升语言处理认知难度 (Gibson, 1998; Hawkins, 2004)。工作记忆理论及语言加工模型进一步证实, 依存距离与认知负荷之间存在正相关关系 (Cowan, 2010; Gildea et al., 2010)。

MDD 值越大，文本认知解码难度越大 (Zhao et al., 2023)。本研究采用 Liu 等 (2009) 提出的方法计算 MDD 值，在计算时排除词根与标点符号的依存关系，具体计算方法如公式 (3) 所示。

$$MDD = \frac{1}{N-1} \sum_{i=1}^{N-1} |DD_i| \tag{3}$$

$N$  代表句子中单词的数量， $|DD_i|$  代表句子中第  $i$  个依存对依存距离的绝对值。  
 为使计算过程更加清晰，表 2 展示了例 (1) 中的依存关系数据，包括该句的词性标注和依存关系类型标注。

表 2 依存关系数据

| Dependency relation | Governor | Position of the governor | Dependent | Position of the dependent |
|---------------------|----------|--------------------------|-----------|---------------------------|
| root                | Root     | 0                        | caught    | 5                         |
| det                 | cat      | 3                        | The       | 1                         |
| amod                | cat      | 3                        | clever    | 2                         |
| nsubj               | caught   | 5                        | cat       | 3                         |
| advmod              | caught   | 5                        | quickly   | 4                         |
| det                 | mouse    | 8                        | a         | 6                         |
| amod                | mouse    | 8                        | small     | 7                         |
| obj                 | caught   | 5                        | mouse     | 8                         |
| punct               | caught   | 5                        | .         | 9                         |

例 (1) 的 MDD 值如下，结果保留三位小数。

$$MDD = \frac{|2| + |1| + |2| + |1| + |2| + |1| + |3|}{7} = 1.714$$

### 3.4 数据处理与分析

本研究使用 R (4.4.1) 计算熵值。首先对摘要文本进行预处理，清除其中的文献引用、序号及标点符号，随后统计各单词在摘要中的出现频率及其概率。在此基础上，依据公式 (1) 和公式 (2) 分别计算每篇摘要的信息熵，并进一步汇总得到中外应用语言学期刊摘要每年的信息熵均值。

本研究采用 R (4.4.1) 和 Stanford CoreNLP (3.9.2) 分析器 (Manning et al., 2014) 计算 MDD 值。鉴于依存距离计算需以标点符号作为句法边界，本研究调整了 R 脚本的预处理流程，仅剔除摘要中的文献引用与序号。随后，运用 Stanford CoreNLP (3.9.2) 对摘要进行句法分析和依存关系标注，并依据公式 (3) 计算每篇摘要的 MDD 值，最终汇总得到中外应用语言学期刊摘要每年的 MDD 均值。

为量化考察期内熵值与 MDD 值的变化趋势，本研究分别对两组数据进行了曲

线拟合，绘制了包含95%置信区间的趋势线，以直观呈现国内外期刊的变化趋势并增强统计推断的可靠性。随后，采用线性回归分析考察两类期刊认知难度的历时变化趋势及其相关性，同时通过独立样本t检验判断其年度均值是否存在显著差异。若线性回归分析结果表明两类期刊的认知难度之间存在显著相关性，则继续进行格兰杰因果检验，探究其相互影响的方向性。

## 4 结果与讨论

### 4.1 中外应用语言学期刊摘要认知编码难度历时变化对比

AL 和 CJAL 英文摘要在2000—2024年的熵值年均值变化如图表所示。图1与图2分别为AL熵值与CJAL熵值的曲线拟合图，表3为熵值线性回归分析结果，表4为熵值独立样本t检验结果，表5为熵值格兰杰因果检验结果。

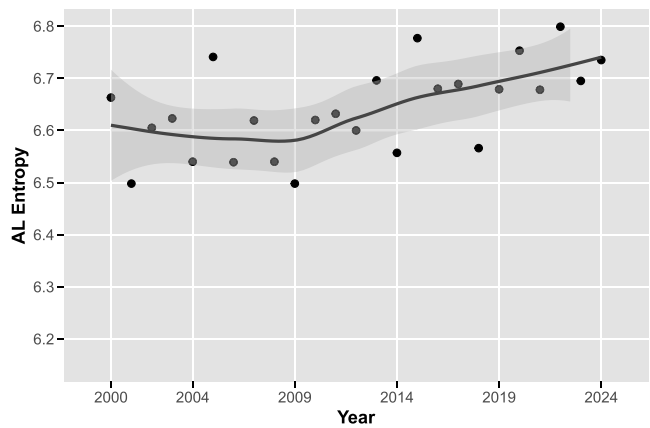


图1 AL熵值曲线拟合

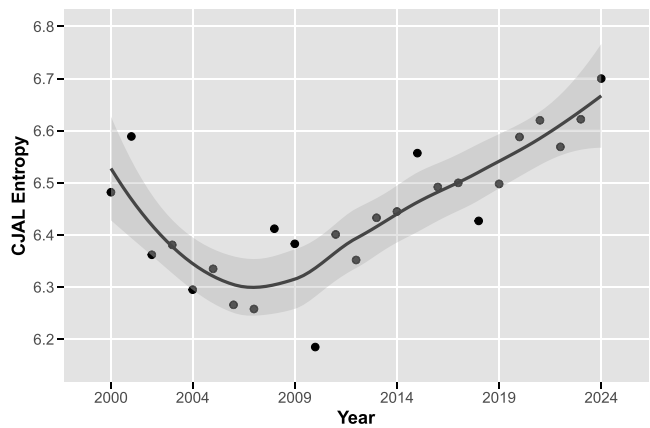


图2 CJAL熵值曲线拟合

表3 熵值线性回归分析结果

| 期刊         | 调整R <sup>2</sup> | F      | 自由度     | 截距    | Beta  | p       |
|------------|------------------|--------|---------|-------|-------|---------|
| AL         | 0.291            | 10.870 | (1, 23) | 6.555 | 0.007 | 0.003** |
| CJAL       | 0.363            | 14.650 | (1, 23) | 6.304 | 0.011 | 0.001** |
| CJAL与AL的关系 | 0.208            | 7.038  | (1, 23) | 1.528 | 0.741 | 0.013*  |

注：\* $p < 0.05$ ，\*\* $p < 0.01$ ，\*\*\* $p < 0.001$

表4 熵值独立样本t检验结果

| 期刊   | M     | SD    | t     | DF | p     | Cohen's d [95% CI] |
|------|-------|-------|-------|----|-------|--------------------|
| AL   | 6.641 | 0.086 | 6.275 | 48 | <.001 | 1.775[1.11, 2.426] |
| CJAL | 6.446 | 0.129 | 6.275 | 48 | <.001 | 1.775[1.11, 2.426] |

注：M=平均值，SD=标准差，DF=自由度，CI=置信区间

由图1、图2、表3和表4可知，考察期内，AL熵值（ $F(1, 23) = 10.87$ ， $\beta = 0.007$ ， $p = 0.003$ ）与CJAL熵值（ $F(1, 23) = 14.65$ ， $\beta = 0.011$ ， $p = 0.001$ ）均呈显著上升趋势，CJAL熵值增长速度快于AL熵值（ $\beta = 0.011 > 0.007$ ），两者之间存在显著差异（ $t = 6.275$ ， $p < 0.001$ ），AL熵值（ $M = 6.641$ ， $SD = 0.086$ ）显著高于CJAL熵值（ $M = 6.446$ ， $SD = 0.129$ ）。此外，AL熵值与CJAL熵值之间存在显著正相关关系（ $F(1, 23) = 7.038$ ， $\beta = 0.741$ ， $p = 0.013$ ）。

表5 熵值格兰杰因果检验结果

| 方向      | 原假设（ $H_0$ ）       | F     | p     | 结论         |
|---------|--------------------|-------|-------|------------|
| AL→CJAL | AL熵值不是CJAL熵值的格兰杰原因 | 0.704 | 0.411 | 不能拒绝 $H_0$ |
| CJAL→AL | CJAL熵值不是AL熵值的格兰杰原因 | 3.110 | 0.093 | 不能拒绝 $H_0$ |

表5显示，AL熵值与CJAL熵值之间不存在格兰杰因果（AL→CJAL： $F = 0.704$ ， $p = 0.411 > 0.05$ ；CJAL→AL： $F = 3.11$ ， $p = 0.093 > 0.05$ ），即两者之间不存在显著的“领先—滞后”预测关系。

结果表明，近二十五年来，中外应用语言学期刊摘要认知编码难度均不断增加，国内期刊摘要认知编码难度增长速度略快于国际期刊，两者之间存在显著差异，国际期刊摘要认知编码难度显著高于国内期刊。需要注意的是，两者之间存在显著正相关关系，但并不存在明显的“领先—滞后”预测关系。

中外应用语言学期刊摘要的认知编码难度均呈增加趋势，该结果与Wu等（2025）针对该学科国际期刊论文所观察到的结果高度吻合，也与Zhao等（2023：5）“近二十年来，摘要熵值呈上升趋势”的宏观结论相符。该一致性表明，学术文本信息密度的提升不仅是一个跨体裁的普遍现象，也已成为学科层面的普遍趋势。由于熵值与信息量呈正相关（Sayood, 2018），其增长趋势可主要归因于写作者学术知识的持续积累（Dolnicar et al., 2015）。随着学科知识体系的不断扩展，新近发表的摘要需要整合更多理论概念、方法细节与研究主题，从而导致其信息密度显

著高于以往 (Biber et al., 2010; Zhou et al., 2023)。作为一门蓬勃发展的交叉学科,应用语言学的理论范式与研究方法持续革新,学科体系日趋复杂 (Hyland et al., 2021)。这种学科特性促使研究者在有限篇幅内整合更多元的知识要素,进一步增加了摘要的信息负荷。此外,学术发表环境的竞争压力也在一定程度上强化了这一趋势 (Lambert et al., 2011)。作为研究成果的“窗口”,摘要承担着向期刊编辑与审稿人展示研究价值的重要功能 (Nicholas et al., 2003)。为提升研究的学术影响力,突出论文亮点,作者往往策略性地浓缩更多信息,这些内容客观上进一步提高了摘要的信息密度,加大了其认知编码难度。

然而,本研究的发现与 Zhao 等 (2023) 关于人文学科 (涵盖语言学) 摘要认知编码难度无显著变化的结论形成鲜明对比,这一差异可能源于分析层级的不同。Zhao 等 (2023) 以宏观的人文学科为观察单位,样本涵盖哲学、历史等以阐释为核心的传统领域,其知识积累模式呈线性、平稳特征 (Becher et al., 2001); 本研究聚焦应用语言学这一具体分支,该领域具有跨学科属性和社会科学取向,知识更新与理论迭代的速度要快于传统人文学科 (Lillis et al., 2010)。该对比表明,在考察学术文本的历时演变时,宏观学科层面的整体趋势可能掩盖其内部子领域的动态发展轨迹,忽视层次差异可能导致对真实演化模式的误判。

对中外期刊的对比分析进一步揭示了认知编码难度演变中的差异化路径。数据显示,国际期刊摘要始终具有更高的熵值水平,这与其作为全球学术话语前沿阵地的定位相符,也体现了对理论创新与方法严谨性的更高要求 (Loi et al., 2010)。相较之下,国内期刊摘要展现出更快的熵值增长率,表明国内学术界正系统性地推进其研究表达与国际规范接轨 (Li et al., 2020)。尤为重要的是,格兰杰因果检验结果显示,AL 与 CJAL 的熵值变化之间不存在显著的“领先—滞后”关系。这一发现表明,国内期刊认知编码难度的提升并非单向受国际期刊影响,而更可能是学科内在发展需求与中国特有学术评价体系共同驱动下形成的自主演进路径。这种“殊途同归”的演变模式提示我们,学术全球化背景下的知识传播既存在共性趋势,也保持着显著的地域特色与发展自主性 (Canagarajah, 2002)。

#### 4.2 中外应用语言学期刊摘要认知解码难度历时变化对比

国际期刊和国内期刊英文摘要在 2000—2024 年的 MDD 值年均值变化如以下图表所示。图 3、图 4 分别为 AL 与 CJAL 的 MDD 值曲线拟合图,表 6 为 MDD 值线性回归分析结果,表 7 为 MDD 值独立样本 t 检验结果,表 8 为 MDD 值格兰杰因果检验结果。

由图 3、图 4、表 6 和表 7 可知,AL 与 CJAL 的 MDD 值在考察期内均未达到统计学上的显著变化,但两者在变化方向和变化速度上存在明显差异。具体而言,AL 的 MDD 值相对平稳,呈弱下降趋势 ( $F(1, 23) = 0.242, \beta = -0.002, p = 0.627$ ); CJAL 的 MDD 值波动幅度较大,呈上升趋势 ( $F(1, 23) = 2.871, \beta = 0.007, p =$

0.104)，且其变化速度略快于AL ( $\beta=0.007>-0.002$ )。此外，AL的MDD值与CJAL的MDD值之间存在显著负相关关系 ( $F(1, 23)=5.211, \beta=-0.701, p=0.01$ )。

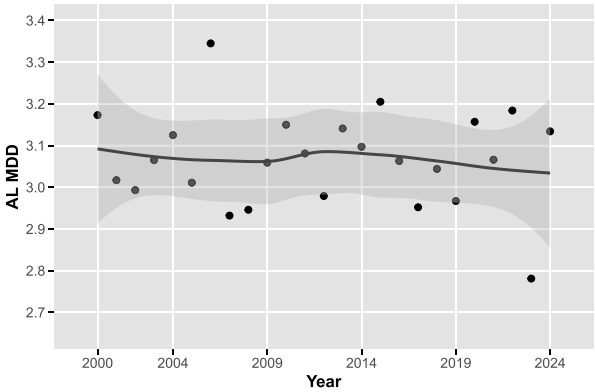


图3 AL的MDD值曲线拟合

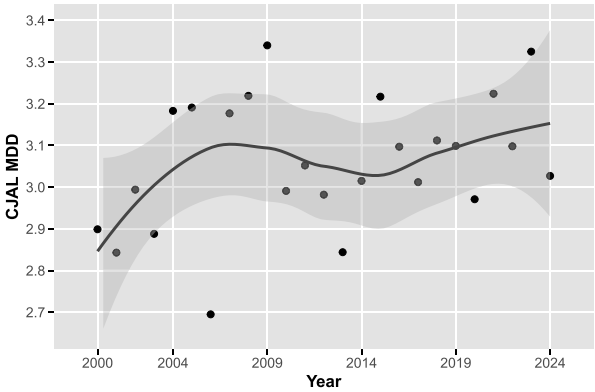


图4 CJAL的MDD值曲线拟合

表6 MDD值线性回归分析结果

| 期刊         | 调整R <sup>2</sup> | F     | 自由度     | 截距    | Beta   | p       |
|------------|------------------|-------|---------|-------|--------|---------|
| AL         | -0.033           | 0.242 | (1, 23) | 3.087 | -0.002 | 0.627   |
| CJAL       | 0.072            | 2.871 | (1, 23) | 2.968 | 0.007  | 0.104   |
| CJAL与AL的关系 | 0.225            | 7.956 | (1, 23) | 5.211 | -0.701 | 0.010** |

注：\*  $p < 0.05$ ，\*\*  $p < 0.01$ ，\*\*\*  $p < 0.001$

表7 MDD值独立样本t检验结果

| 期刊   | M     | SD    | t      | DF | p     | Cohen's d [95% CI]   |
|------|-------|-------|--------|----|-------|----------------------|
| AL   | 3.067 | 0.113 | -0.178 | 48 | 0.860 | -0.05[-0.605, 0.504] |
| CJAL | 3.060 | 0.157 | -0.178 | 48 | 0.860 | -0.05[-0.605, 0.504] |

注：M=平均值，SD=标准差，DF=自由度，CI=置信区间

表8 MDD值格兰杰因果检验结果

| 方向      | 原假设 ( $H_0$ )            | F     | $p$   | 结论         |
|---------|--------------------------|-------|-------|------------|
| AL→CJAL | AL的MDD值不是CJAL的MDD值的格兰杰原因 | 1.376 | 0.254 | 不能拒绝 $H_0$ |
| CJAL→AL | CJAL的MDD值不是AL的MDD值的格兰杰原因 | 0.204 | 0.656 | 不能拒绝 $H_0$ |

由表8可知，AL的MDD值与CJAL的MDD值之间不存在格兰杰因果（AL→CJAL：F=1.376， $p=0.254>0.05$ ；CJAL→AL：F=0.204， $p=0.656>0.05$ ），即两者之间不存在显著的“领先—滞后”预测关系。

研究结果显示，过去二十五年间，中外应用语言学期刊摘要认知解码难度均未达到统计学上的显著变化，但其演变模式存在明显差异。具体而言，国际期刊摘要认知解码难度呈弱下降趋势，而国内期刊摘要认知解码难度表现出一定的上升趋势，且国内期刊摘要认知解码难度的变化速度略快于国际期刊。进一步分析显示，两者之间存在显著的负相关关系，但并未表现出明确的“领先—滞后”预测关系。

在考察期内，国际期刊摘要的认知解码难度整体保持稳定，并呈现出轻微下降趋势。这一趋势不仅与Zhao等（2023）关于自然科学和人文学科期刊摘要认知解码难度下降的发现相一致，也符合依存距离最小化这一被广泛验证的人类语言普遍规律。作为高度功能化和传播导向的学术体裁，期刊摘要承担着在有限篇幅内高效传递研究核心信息的任务，过高的认知加工负荷反而会削弱其信息传播效率。从语言认知和信息加工的角度来看，较短的句法依存距离有助于降低工作记忆负担，提高读者对信息的即时整合能力（Gibson，1998；Liu，2008）。随着国际学术共同体对摘要写作规范和读者导向意识的不断强化，写作者在句法选择上趋向于使用结构更为紧凑、线性化程度更高的表达方式，从而使摘要在历时维度上呈现出认知解码难度稳定甚至微弱下降的趋势。这在一定程度上表明，国际期刊摘要写作已形成相对成熟且制度化的表达范式，其语言复杂度受到体裁规范与认知效率的双重约束。

相比之下，国内期刊英文摘要的认知解码难度整体呈现出波动中上升的趋势，且其变化速度略快于国际期刊难度下降的速度。这一结果与Wu等（2025）关于应用语言学定量研究全文认知解码难度上升的发现相一致，这说明该现象并非局限于摘要层面，而可能反映了应用语言学领域更为普遍的写作发展特征。其背后成因可从两方面理解。一方面，全球学术竞争日益激烈，国内学者为增强研究的可见度与论证说服力，往往在摘要中策略性地引入更多促销性元话语资源，如态度标记、强化词以及作者立场表达。这类资源在Hyland（2005）的学术话语模型中被视为构建学术身份和增强互动性的关键手段，但其使用往往伴随着更复杂的修饰结构和更长的句法跨度，从而在客观上拉大了文本的依存距离，增加了读者的认知解码难度。另一方面，国内学术写作规范正处于持续调整与快速发展的阶段，在积极向国际学术标准靠拢的过程中，部分写作者可能出于增强语篇制度化特征或适应第二语言学术写作中的结构约束的考虑，开始倾向于使用复杂名词化结构等高密度信息表

达方式。这种对国际学术话语形式的“过度模仿”虽有助于对标国际学术文本标准，却也可能在短期内推高文本的认知加工负荷，这一现象在第二语言学术写作研究中并不少见（Biber et al., 2011; Lu et al., 2015）。

进一步分析发现，国际期刊与国内期刊摘要认知解码难度变化之间虽存在显著的负相关关系，但并未呈现出格兰杰因果关系。这意味着，两者在时间序列上并不存在明确的单向预测关系，国内期刊英文摘要的发展路径并非简单地追随国际期刊的变化轨迹。相反，其变化更可能是在全球化学术规范压力、本土评价体系以及学科内部话语实践之间不断协商与调适的结果。从这一意义上看，国内期刊摘要认知解码难度的上升不宜被简单解读为“偏离国际标准”，而应理解为一种转型期的写作策略选择，其背后反映的是学术话语在不同语境中对说服力、复杂性与可理解性之间关系的重新权衡。

## 5 结语

本研究基于自建历时语料库，对2000—2024年中外应用语言学期刊英文摘要的认知难度演变进行了系统性考察，分别从认知编码难度与认知解码难度两个维度对其发展轨迹展开了对比分析。研究发现，随着时间推移，摘要的认知编码难度在国内外期刊中均呈现逐步增强的趋势，其中国际期刊的认知编码难度持续高于国内期刊，两者演变趋势虽呈现一定程度的正向关联，但并未表现出显著的时间预测关系。另一方面，认知解码难度在两类期刊中呈现出不同的发展方向，国内期刊摘要的认知解码难度逐渐增加，而国际期刊则保持相对稳定，甚至呈略微下降趋势，两者在解码维度上显示出一定的负向关联，但未发现格兰杰因果关系。上述结果表明，国内外应用语言学期刊英文摘要在认知难度的演变趋势与调节机制上既存在一定的差异性，又保持着相对的演变独立性。

这些发现不仅揭示了应用语言学领域英文学术摘要在不同认知维度上的历时演变路径，也为理解中外学术交流中的语言适应机制与认知负荷分配提供了实证依据。国际期刊摘要较高的认知编码难度，可能与其对理论创新、研究问题复杂性以及方法论多样性的持续强调有关，这在一定程度上推高了信息组织与概念编码的复杂度；而国内期刊摘要认知解码难度的上升，则可能反映出其在融入国际学术话语体系的过程中，对复杂句法结构、名词化表达及规范化学术修辞的逐步吸纳与适应。这一过程体现了学术写作在追求学术权威性、表达规范性与可理解性之间的动态权衡。

尽管本研究在一定程度上揭示了应用语言学期刊英文摘要认知难度的变化特征，但仍存在一定局限性。当前分析仅基于近25年间国内外两本具有代表性的应用语言学英文期刊摘要文本，样本规模与覆盖范围相对有限。未来研究可在此基础上进一步拓展期刊类型与数量，构建更大规模、多层次的历时语料库，并适当延长

观测时间跨度，以更全面、细致地刻画学术文本认知难度的演变规律及其背后的学科话语机制。

## 参考文献

- Ante L, 2022. The relationship between readability and scientific impact: evidence from emerging technology discourses[J]. *Journal of informetrics*, (16): 101252.
- Becher T, Trowler P, 2001. *Academic tribes and territories*[M]. Buckingham: SRHE/Open University Press.
- Bentz C, Alikaniotis D, Cysouw M, et al., 2017. The entropy of words—learnability and expressivity across more than 1000 languages[J]. *Entropy*, 19(6): 275.
- Biber D, Gray B, 2010. Challenging stereotypes about academic writing: complexity, elaboration, explicitness[J]. *Journal of English for academic purposes*, 9(1):2-20.
- Biber D, Gray B, Poonpon K, 2011. Should we use characteristics of conversation to measure grammatical complexity in L2 writing development?[J] *TESOL quarterly*, 45(1): 5-35.
- Canagarajah S, 2002. Multilingual writers and the academic community: towards a critical relationship [J]. *Journal of English for academic purposes*, 1(1): 29-44.
- Chen R, Liu H, Altmann G, 2017. Entropy in different text types[J]. *Digital scholarship in the humanities*, 32(3): 528-542.
- Cowan, N, 2010. The magical mystery four: how is working memory capacity limited, and why? [J] *Current directions in psychological science*, 19(1): 51-57.
- Dolnicar S, Chapple A, 2015. The readability of articles in tourism journals[J]. *Annals of tourism research* (52):161-166.
- Flesch R, 1948. A new readability yardstick[J]. *Journal of applied psychology* (32): 221-233.
- Gibson E, 1998. Linguistic complexity: locality of syntactic dependencies[J]. *Cognition*, 68(1): 1-76.
- Gildea D, Temperley D, 2010. Do grammars minimize dependency length?[J] *Cognitive science*, 34(2): 286-310.
- Gazni A, 2011. Are the abstracts of high impact articles more readable? Investigating the evidence from top research institutions in the world[J]. *Journal of information science* (37): 273-281.
- Hausser J, Strimmer K, 2014. Entropy: estimation of entropy, mutual information and related quantities [EB/OL]. (2014-11-04) [2026-04-12]. <https://CRAN.R-project.org/package=entropy>.
- Hawkins J A, 2004. *Efficiency and complexity in grammars*[M]. Oxford: Oxford University Press.
- Heringer H J, Strecker B, Wimmer R, 1980. *Syntax: Fragen, Lösungen, Alternativen*[M]. München: Wilhelm Fink Verlag.
- Hyland K, 2005. Stance and engagement: a model of interaction in academic discourse[J]. *Discourse studies*, 7(2): 173-192.
- Hyland K, Jiang F, 2021. A bibliometric study of EAP research: who is doing what, where and when?[J] *Journal of English for academic purposes*, 49: 100929.
- Jiang F K, Hyland K, 2017. Metadiscursive nouns: interaction and cohesion in abstract moves[J]. *English for specific purposes*, (46):1-14.

- Juola P, 2008. Assessing linguistic complexity[M]//Miestamo M, Sinnemäki K, Karlsson F. Language complexity: typology, contact, change. John Benjamins Press: 89-108.
- Juola P, 2013. Using the Google N-Gram corpus to measure cultural complexity[J]. *Literary and linguistic computing*, 28(4): 668-675.
- Koch C A, 2009. A theory of consciousness[J]. *Scientific American mind*, (20): 16-19.
- Lambert V A, Lambert C E, 2011. Writing an appropriate title and informative abstract[J]. *Pacific Rim international journal of nursing research*, 15: 171-172.
- Lei L, Yan S, 2016. Readability and citations in information science: evidence from abstracts and articles of four journals (2003 – 2012) [J]. *Scientometrics*, (108): 1155-1169.
- Li Y, Flowerdew J, 2020. Teaching English for research publication purposes (ERPP): a review of language teachers' pedagogical initiatives[J]. *English for specific purposes*, 59: 29-41.
- Lillis T, Curry M J, 2010. Academic writing in a global context: the politics and practices of publishing in English[M]. London: Routledge.
- Liu H, 2008. Dependency distance as a metric of language comprehension difficulty[J]. *Journal of cognitive science*, 9(2): 159-191.
- Liu H, Hudson R, Feng Z, 2009. Using a Chinese treebank to measure dependency distance[J]. *Corpus linguistics and linguistic theory*, 5(2): 161-174.
- Liu H, Xu C, Liang J, 2017. Dependency distance: a new perspective on syntactic patterns in natural languages[J]. *Physics of life reviews*, (21): 171-193.
- Liu Z, 2016. A diachronic study on British and Chinese cultural complexity with Google Books N-grams [J]. *Journal of quantitative linguistics*, 23(4): 361-373.
- Loi C K, Evans M S, 2010. Cultural differences in the organization of research article introductions from the field of educational psychology: English and Chinese[J]. *Journal of pragmatics*, 42(10): 2814-2825.
- Lu X, Ai H, 2015. Syntactic complexity in college-level English writing: differences among writers with diverse L1 backgrounds[J]. *Journal of second language writing*, 29: 16-27.
- Martín P M, 2003. A genre analysis of English and Spanish research paper abstracts in experimental social sciences[J]. *English for specific purposes*, 22(1): 25-43.
- Manning C D, Surdeanu M, Bauer J, et al., 2014. The Stanford CoreNLP natural language processing toolkit[C]//Association for Computational Linguistics. Proceedings of 52nd annual meeting of the Association for Computational Linguistics: System Demonstrations. Baltimore, Maryland: Association for Computational Linguistics: 55-60.
- Mead P, 2005. Methodological issues in the study of interpreters' fluency[J]. *The interpreters' newsletter* (13): 39-63.
- Nicholas D, Huntington P, Watkinson A, 2003. Digital journals, big deals and online searching behaviour: a pilot study[J]. *Aslib proceedings: new information perspectives*, 55(1/2): 84-109.
- Rajput N K, Ahuja B, Riyal M K, 2018. A novel approach towards deriving vocabulary quotient[J]. *Digital scholarship in the humanities*, 33(4): 894-901.
- Sayood K, 2018. Information theory and cognition: a review[J]. *Entropy*, 20(706): 1-19.
- Searle J R, 2013. Can information theory explain consciousness?[M]. Cambridge: MIT Press.

- Shannon C E, 1948. A mathematical theory of communication[J]. Bell system technical journal, 27: 379-423.
- Shi Y, Lei L, 2020. Lexical richness and text length: an entropy-based perspective[J]. Journal of quantitative linguistics, 29(1): 62-79.
- Wang S, Liu X, Zhou J, 2022. Readability is decreasing in language and linguistics[J]. Scientometrics, 127: 4697-4729.
- Wu J, Chen H, Chen L, 2025. The impact of research paradigms on encoding and decoding difficulties of applied linguistics articles[J]. Humanities and social sciences communications, (12): 416.
- Zhao X, Li L, Xiao W, 2023. The diachronic change of research article abstract difficulty across disciplines: a cognitive information-theoretic approach[J]. Humanities and social sciences communications (10): 194.
- Zhang Y, 2015. Entropic evolution of lexical richness of homogeneous texts over time: a dynamic complexity perspective[J]. Journal of language modelling, 3(2): 569-599.
- Zhou Y, He P, Liu H, 2023. Diachronic changes in lexical complexity of biology research articles: a century-long study. Journal of quantitative linguistics, 30(3): 267-289.
- Zhu H, Lei L, 2018. British cultural complexity: an entropy-based approach[J]. Journal of quantitative linguistics, 25(2): 190-205.
- 陈庆斌, 2021. 学术期刊论文摘要中作者立场标记的对比研究[J]. 外语学刊 (2): 41-47.
- 曹雁, 邹智杰, 2017. 科技期刊英文摘要核心语义系统的三维空间——从主题词到主题概念[J]. 中国外语, 14 (4): 62-71.
- 方昱, 刘海涛, 2021. 句法结构认知难度的计算指标分析[J]. 南京师大学报 (社会科学版) (6): 126-137.
- 鞠玉梅, 2020. 中外学者英语学术论文摘要修辞劝说机制比较研究[J]. 解放军外国语学院学报 (1): 85-92+160.
- 刘国兵, 张茹昕, 2023. 国内硕博学位论文摘要语言复杂度发展特征研究[J]. 中国 ESP 研究 (2): 66-76+120-121.
- 刘海涛, 2009. 依存语法的理论与实践[M]. 北京: 科学出版社.
- 李芝, 成晓敏, 2020. 中外学术期刊英文摘要立场标记语对比研究[J]. 西安外国语大学学报, 28 (2): 6-10.
- 牛桂玲, 2013. 学术期刊论文摘要研究的新视角[J]. 河南大学学报 (社会科学版), 53 (5): 150-156.
- 宋瑞梅, 汪火焰, 2019. 中国博士学位论文摘要句法复杂度研究[J]. 解放军外国语学院学报, 42 (1): 84-91.

## 通信地址

453007 河南省新乡市 河南师范大学外国语学院

(责任编辑: 杨瑞雪)

# 中国英语学习者与英语本族语者指物关系代词选择的多因素研究<sup>\*</sup>

北京外国语大学 郝美佳

西南大学 唐瑞梁

**提 要：**本研究基于中国学生万篇英语作文语料库（Ten-thousand English Compositions of Chinese Learners，简称TECCL）和CLOB语料库，考察中国英语学习者与英语本族语者在限定性关系从句中指物关系代词which与that选择的异同及其制约因素，并探析关系代词选择的认知加工机制。研究发现，本族语者与学习者对指物关系代词的选择是与先行词、从句、主句相关的多种词汇—语法、语义因素共同作用的结果，且这些因素在特定水平上存在显著的交互效应，这反映了语境在近义构式选择中的关键作用。同时，本族语者与学习者的关系代词选择在若干因素水平上存在显著差异，表现为本族语者与学习者在先行词词性、先行词复杂度等变量水平上对which与that的选择倾向不同；二者的关系代词选择虽受共性因素制约，但是这些因素对两个语言群体的影响权重不同；与本族语者相比，学习者对变量引发的认知负载变化更为敏感。

**关键词：**关系代词差异；制约因素；认知加工机制；多因素分析

## A Multifactorial Analysis of the Choice of Relativizers “which” and “that” Between Chinese EFL Learners and Native English Speakers

Meijia Hao and Ruiliang Tang

**Abstract:** Drawing on the Ten-thousand English Compositions of Chinese Learners (TECCL)

<sup>\*</sup> 唐瑞梁为本文通信作者。

本文系重庆市社科规划一般项目“汉语语用标记之构式演变与构式化研究”（项目编号：YBYY04）和中央高校基本科研重点项目“构式语法视域下英汉语用标记演变的对比研究”（项目编号：SWU1909318）的阶段性研究成果。

作者贡献：

郝美佳：选题构思、研究方法、数据收集、数据分析、讨论结论、初稿撰写、字数占比（70%）；

唐瑞梁：选题构思、研究方法、讨论结论、论文撰写、修改润色、字数占比（30%）。

and the CLOB corpus, this study examines similarities and differences between Chinese EFL learners and native speakers in their choice between the relativizers “which” and “that” in restrictive relative clauses with inanimate antecedents, and explores the conditioning factors and the cognitive processing mechanisms underlying relativizer selection. The analysis reveals that relativizer choice among native English speakers and Chinese learners is jointly conditioned by a series of lexico-grammatical and semantic factors associated with the antecedent, the relative clause, and the matrix clause. There are significant interactions among these factors at specific levels, which reflects the crucial role of context in the selection of near-synonymous constructions. On the other hand, native speakers and Chinese learners differ significantly in their relativizer choices at several levels of the conditioning factors. Specifically, they show different preferences for “which” versus “that” across variable levels such as the antecedent’s part of speech and its complexity. Additionally, although the same factors condition both groups’ choice of relativizer, their relative weights differ between the groups. Furthermore, compared with native speakers, learners are more sensitive to change in cognitive load induced by contextual factors.

**Keywords:** relativizer variation; conditioning factors; cognitive processing mechanisms; multifactorial analysis

## 1 引言

无论在句法分析和语言处理领域,还是在语言习得研究中,关系从句结构因其独特的句法特点及其在语法理论中的重要地位,一直受到广泛关注(李金满等,2007:198)。长期以来,针对限定性关系从句的研究主要集中于对比不同类型关系从句的认知加工难度,针对关系代词的研究则相对零散,对关系代词之间用法差异的论述多散见于语法著作中(房印杰,2016:1)。Biber等(1999:615)指出,关系代词which与that在语法潜能方面颇为相似,二者均可用于多种空位结构中,然而,它们在实际使用模式上却存在诸多重要差异。在房印杰(2016)关于关系代词取舍研究的启示下,本研究基于中国学生万篇英语作文语料库(Ten-thousand English Compositions of Chinese Learners,简称TECCL)(薛熙哲,2015)和CLOB语料库(Xu et al.,2013),采用多因素分析方法和复杂统计模型,探究中国英语学习者与英语本族语者对限定性关系从句中指物关系代词which与that选择的相似性和差异性,探讨词汇—语法环境对语言选择的影响,并尝试对两个群体不同的语言选择作出理论阐释。

## 2 文献综述

长期以来,国内外针对关系从句的研究主要集中于对比不同类型关系从句的认

知加工难度,对关系从句的习得过程提出假设并进行检验。针对关系代词的研究则相对零散,对关系代词之间用法差异的论述多散见于语法著作中(房印杰 2016: 1)。早在1933年,Jespersen (1933: 295)在语法著作《英语语法纲要》中就讨论了限定性关系从句中wh-型关系代词与that的选择:wh代词最初源于疑问代词,而在中古英语时期,which、whom和who逐渐用作关系代词,并在中世纪最后几百年间,尤其是在较为正式的文学体裁中,逐步取代that的部分用法。然而,自中世纪以来,that一直是文学语体中最常用的关系词。Quirk等(1985)所著《英语综合语法》、Biber等(1999)所著基于语料库的《朗文英语口语和笔语语法》以及Huddleston等(2002)所著《剑桥英语语法》也对关系词选择进行了系统、详尽的讨论。在上述语法研究基础上,本文归纳了在以无生命先行词为中心的限定性关系从句中,which与that选择所受制约的因素。一般而言,在以下情境中更倾向于使用that。

- a. 先行词为不定代词(如anything、everything、nothing、something);
- b. 先行词为非人称融合限定结构(如all、little、much、most、few、some、any);
- c. 先行词受最高级或first、last、next、only等后置限定词修饰;
- d. 先行词结构较为简单,仅为“限定词+中心词”,见例(1);
- e. 先行词为疑问代词who或which;
- f. 关系成分充当表属性的表语补语,见例(2)。

在以下情境中,关系从句更倾向于使用which。

- a. 先行词为指示代词(this, that, these, those);
- b. 先行词中心词与关系代词之间存在较复杂的短语或从句插入,见例(3)。

(1) I'll take you to [the building [that all elderly university teachers prefer]].

(2) He's no longer the trustworthy friend [(that) he was in those days].

(3) I have [[interests outside my immediate work and its problems] which I find satisfying].

上述语法研究表明,which与that的选择主要受先行词的复杂性与定指性、关系词在从句中的句法功能、先行词与关系从句之间距离等因素的制约。此外,文体特征、方言差异以及中心名词的语法类型对which与that的选择也具有重要影响。在构式语法理论框架下,Wiechmann (2015)运用关联规则挖掘、构型频率分析与聚类等多因素分析方法,系统考察并归纳了不同类型关系从句的典型结构特征。该研究特别关注关系词省略现象,发现当结构高度固化且可预测性较强时,更容易出现省略现象;而在歧义或认知负荷较高的语境中,则倾向于使用显性结构标记。在此基础上,房印杰(2016)考察了中国英语学习者和英语母语者产出的非主句型关

系从句, 聚焦两个语言群体在关系代词取舍上的异同, 揭示了其背后共有的认知加工机制, 并特别探讨影响中国学习者的特定机制。

房印杰(2016)在先导研究中发现, 中国英语学习者在关系代词的选择中倾向使用 *which*, 而英语本族语者倾向使用 *that*。基于这一典型倾向性差异, 本研究尝试对中国学生在英语限定性关系从句中关系代词指代指物的先行词时 *which* 与 *that* 的选择, 开展基于大规模语料的多因素统计分析。既有研究所识别出的影响关系词选择的因素为本研究提供了理论基础, 并指导了标注体系的构建。然而, 现有研究仍存在若干不足。首先, 既往研究多依赖内省方法, 虽识别了相关因素, 但各因素的相对权重尚不明确, 即哪些因素具有更显著的影响仍有待探讨。其次, 该领域整体上仍偏重描述性分析, 解释性研究相对不足。最后, 文献中提到的“影响关系代词选择的因素之间存在交互作用”的论述可以被看作多因素分析理念的萌芽, 但对这一论述的验证缺乏大规模真实语料的支持。

鉴于此, 本研究通过比较中国英语学习者与英语本族语者在限定性关系从句中指物关系代词 *which* 与 *that* 的使用情况, 旨在揭示造成两个语言群体关系代词选择差异的内在影响因素及潜在认知机制。本研究尝试回答以下两个研究问题。其一, 在限定性关系从句中, 当关系代词指代指物的先行词时, 中国英语学习者与英语本族语者对于关系代词 *which* 与 *that* 的选择存在哪些共同的制约因素? 其二, 在限定性关系从句中, 当关系代词指代指物的先行词时, 中国英语学习者与英语本族语者对于关系代词 *which* 与 *that* 的选择有何显著差异?

### 3 研究设计

#### 3.1 语料选取

本研究选取的本族语者语料与学习者语料来自同一时期。学习者语料来自 TECCL 语料库, 约 180 万词。语料产生于 2011—2015 年, 语料取样分布代表性较好, 产生语料的写作任务类型多样。本族语者语料来自 CLOB 语料库, 收录了 2008—2011 年共 100 万词的英语语料, 文本类型取样均衡。

将学习者与本族语者语料分别导入 AntConc 3.5.8, 以 *which* 和 *that* 为检索词进行检索。参照 Wiechmann (2015) 与房印杰 (2016: 62-66) 关于关系词选择研究中对限定性关系从句的筛选标准, 我们对检索结果进行人工筛选, 剔除所有非限定性关系从句, 以及先行词受多重或嵌套从句修饰等不适合纳入分析的限定性关系从句, 从而确保最终数据集仅包含以无生命先行词为中心, 且在语法上 *that* 和 *which* 均可替换的限定性关系从句。在中国学习者语料中, 仅保留关系词使用正确的关系从句。对于与关系从句本身无关的语法错误, 予以修正后纳入整体研究语料。对于影响关系词选择的语法错误句, 则予以剔除, 如关系词误用、关系词冗余、关系从句中缺失必要空位。

经过人工筛选，我们共从CLOB与TECCL语料库中提取出5,650个限定性关系从句，其在两个语料库中的分布情况见表1。

表1 限定性关系从句在语料库中的分布情况

| 语料类型  | 语料库（库容）          | which | that  | 合计    |
|-------|------------------|-------|-------|-------|
| 母语者语料 | CLOB（1,023,466）  | 553   | 2,304 | 2,857 |
| 学习者语料 | TECCL（1,817,335） | 1,442 | 1,351 | 2,793 |
|       | 合计               | 1,995 | 3,655 | 5,650 |

3.2 变量标注

经过对限定性关系从句的检索、抽取和筛选，依据前人对wh-型关系代词与that用法差异的研究选择潜在影响因素对语料进行特征标注。本研究共包括14项预测因素，反应变量为指物关系代词which与that的选择，标注框架如表2所示。

表2 标注框架

| 变量类型 | 变量（标签）              |                              | 变量水平（标签）                                                               |
|------|---------------------|------------------------------|------------------------------------------------------------------------|
| 预测变量 | 先行词                 | 词性（Head.POS）                 | 名词（Noun）、不定代词（Indef.Pro）、指示代词（Dem.Pro）                                 |
|      |                     | 修饰词（Head.determiner）         | 非人称后位限定词（Post.det）、序数词（Ordinal）、最高级（Superlative）、其他（Na）                |
|      |                     | 修饰语数量（Head.modi）             |                                                                        |
|      |                     | 复杂度：是否是“限定词+名词”的结构（Det.Noun） | 限定词+中心词（Det.N）、其他结构（Not）                                               |
|      |                     | 在从句中句法成分（Head.RC）            | 主语（Subject）、宾语（Object）、介词宾语（Prep.Object）、系词补语（Predicative）             |
|      |                     | 确定性（Head.def）                | 确定（Definite）、不确定（Not）                                                  |
|      | 从句                  | 在主句中嵌入位置（RC.embed）           | 中央嵌入（Central）、右侧嵌入（Right）                                              |
|      |                     | 从句与先行词之间有无分隔（RC.separation）  | 有插入（Sep）、无插入（No.Sep）                                                   |
|      |                     | 谓语动词类型（RC.predicate）         | 系表动词（Copular）、及物动词（Transitive）、双宾语动词（Ditransitive）、不及物动词（Intransitive） |
|      |                     | 从句长度（RC.length）              |                                                                        |
|      |                     | 从句语态（RC.voice）               | 主动（Active）、被动（Passive）                                                 |
|      |                     | 从句肯定/否定（RC.nega）             | 肯定（Affirmative）、否定（Negative）                                           |
|      | 主句                  | 主句语态（MC.voice）               | 主动（Active）、被动（Passive）                                                 |
|      |                     | 主句肯定/否定（MC.nega）             | 肯定（Affirmative）、否定（Negative）                                           |
| 反应变量 | 指物关系代词（Relativizer） |                              | which、that                                                             |

### 3.3 统计分析

标注完成后，将英语本族语者与中国英语学习者的数据分别导入统计编程软件R，使用通用线性模型glm函数，以所标注因素及因素间的交互为预测变量，以指物关系代词which与that的选择为反应变量，构建二分类逻辑斯蒂回归模型。通过逐步回归stepAIC函数对数据进行拟合，剔除不显著的因素及因素间的交互，对模型进行优化，得到回归结果。

## 4 数据结果

### 4.1 制约英语本族语者指物关系代词选择的因素

根据回归结果，制约英语本族语者指物关系代词选择的因素按统计显著性由高到低排列如表3所示。

表3 英语本族语者指物关系代词选择的显著性制约因素

| 序号 |                                          | Estimate   | Std. Error | z value | Pr(> z )    |
|----|------------------------------------------|------------|------------|---------|-------------|
| 0  | (Intercept)                              | 1.279e+00  | 2.462e-01  | 5.194   | 2.06e-07*** |
| 1  | Head.modi:Det.NounDet.N                  | -7.351e-01 | 1.646e-01  | -4.467  | 7.93e-06*** |
| 2  | Det.NounDet.N                            | 1.069e+00  | 2.508e-01  | 4.264   | 2.01e-05*** |
| 3  | RC.length                                | -6.449e-02 | 1.951e-02  | -3.305  | 0.000951*** |
| 4  | RC.embedCentral:RC.separationSep         | 2.098e+00  | 7.643e-01  | 2.745   | 0.006051**  |
| 5  | Head.POSIndef.Pro                        | 3.037e+00  | 1.154e+00  | 2.631   | 0.008505**  |
| 6  | Head.POSIndef.Pro:RC.length              | -2.331e-01 | 9.509e-02  | -2.451  | 0.014230*   |
| 7  | Head.defDefinite:RC.embedCentral         | 9.143e-01  | 3.882e-01  | 2.355   | 0.018514*   |
| 8  | RC.predicateIntransitive:MC.voicePassive | -1.082e+00 | 4.649e-01  | -2.328  | 0.019921*   |
| 9  | RC.length:RC.voicePassive                | -1.209e-01 | 5.386e-02  | -2.244  | 0.024819*   |
| 10 | Head.defDefinite:Head.RCObject           | 8.072e-01  | 3.740e-01  | 2.158   | 0.030922*   |
| 11 | MC.negaNegative:RC.voicePassive          | 2.663e+00  | 1.264e+00  | 2.107   | 0.035137*   |
| 12 | Head.modi:RC.embedCentral                | -2.995e-01 | 1.423e-01  | -2.105  | 0.035321*   |
| 13 | Head.modi:RC.predicateCopula             | -1.813e-01 | 8.958e-02  | -2.024  | 0.042963*   |

表3中显示的因素水平与英语本族语者对指物关系代词的选择均显著相关。回归系数的正负反映因素的预测方向，本研究中系数为正代表关系代词倾向选择that，系数为负代表关系代词倾向选择which。回归系数绝对值的大小反映因素对关系代词选择的影响力。由表3可知，制约英语本族语者指物关系代词选择的主效应包括先行词的复杂度（“限定词+中心词”的结构）、从句长度以及先行词词性。当先行词由“限定词+中心词”构成或先行词由不定代词充当时，本族语者显著倾向于选择that作为指物关系代词。而当从句长度增大时，显著倾向于选择which作为关系代词。

有些因素虽然没有作为主效应因素独立参与关系代词选择,但是这些因素在特定水平上存在交互效应,共同影响本族语者对指物关系代词的选择。例如,当从句以中央嵌入的方式插入主句中或先行词在从句中作宾语,即从句的认知加工负载较高时,先行词对于语义更为确定的预测概率升高;先行词的修饰语数量增多对从句中系表动词使用的预测概率升高。在这些因素中,与先行词相关的因素包括先行词的修饰语数量、先行词的确定性、先行词在从句中的句法成分;与从句相关的因素包括从句在主句中的嵌入方式、从句与先行词之间的距离(即短语插入到先行词与从句之间)、从句谓语动词类型、从句语态;与主句相关的因素包括主句语态与肯定/否定。

#### 4.2 制约中国英语学习者指物关系代词选择的因素

根据回归结果,制约中国英语学习者指物关系代词选择的因素按统计显著性由高到低排列如表4所示。

表4 中国英语学习者指物关系代词选择的显著性制约因素

| 序号 | 因素水平                                      | Estimate | Std. Error | z value | Pr(> z )    |
|----|-------------------------------------------|----------|------------|---------|-------------|
| 0  | (Intercept)                               | 0.60022  | 0.32715    | 1.835   | 0.066553.   |
| 1  | Head.modi                                 | -0.72189 | 0.17052    | -4.234  | 2.3e-05***  |
| 2  | RC.voicePassive                           | -0.88024 | 0.24339    | -3.617  | 0.000299*** |
| 3  | Head.modi:RC.separationSep                | 0.53647  | 0.16594    | 3.233   | 0.001225**  |
| 4  | Head.RCObject:RC.separationSep            | -3.68042 | 1.15340    | -3.191  | 0.001418**  |
| 5  | Head.defDefinite:RC.separationSep         | -1.70044 | 0.61539    | -2.763  | 0.005724**  |
| 6  | RC.embedCentral:RC.predicateCopula        | -2.04578 | 0.75402    | -2.713  | 0.006665**  |
| 7  | Head.defDefinite:Det.NounDet.N            | -0.77959 | 0.29140    | -2.675  | 0.007466**  |
| 8  | RC.voicePassive:RC.negaNegative           | 1.99586  | 0.84466    | 2.363   | 0.018131*   |
| 9  | Head.modi:Head.RCObject                   | 0.34766  | 0.15056    | 2.309   | 0.020938*   |
| 10 | Head.defDefinite:RC.predicateIntransitive | 0.92613  | 0.42521    | 2.178   | 0.029404*   |
| 11 | MC.negaNegative:RC.voicePassive           | 1.67895  | 0.80920    | 2.075   | 0.038002*   |
| 12 | RC.length                                 | -0.07910 | 0.03849    | -2.055  | 0.039874*   |

表4中显示的因素水平与中国英语学习者对指物关系代词的选择均显著相关。由表4可知,制约中国英语学习者指物关系代词选择的主效应包括先行词的修饰语数量、从句语态、从句长度。当先行词的修饰语数量增多,从句长度增大或学习者在从句中使用被动语态时,即认知加工负载较高时,学习者显著倾向于选择 which 作为指物关系代词。

因素水平之间的交互效应也影响着学习者对指物关系代词的选择。例如,当从句与先行词之间有分隔,即认知加工负载较高时,先行词对于语义更为确定的预测概率升高;当从句以中央嵌入的方式插入到主句中,即认知加工负载升高时,对从句中系表动词使用的预测概率升高。

由因素的回归系数可知,学习者对从句被动语态的敏感程度远大于从句长度,这可能与被动语态在表层语法形式上更为明显,更容易被学习者注意到有关。此外,学习者对于从句的否定、从句与先行词之间的分隔、从句以中央嵌入的方式插入到主句中等从句认知加工负载升高的情况更为敏感。从句的肯定/否定这一因素虽然并未独立或与其他因素交互参与到本族语者的关系代词选择中,但这一因素通过与其他因素的交互显著影响了学习者对于关系代词的选择。从句与先行词之间的分隔这一因素在学习者关系代词选择中与其他多个因素存在显著交互,从句与先行词的分隔带来的认知加工负载升高,这表明学习者对认知加工负载升高的情况更为敏感。

## 5 讨论

### 5.1 中国英语学习者与英语本族语者在指物关系代词选择中的共性制约因素

英语本族语者和中国英语学习者对指物关系代词的选择不是由单一因素决定的,而是由与先行词、从句、主句相关的多种词汇—语法、语义因素共同作用的结果,且这些因素在特定水平上存在显著的交互效应,共同影响本族语者与学习者对指物关系代词的选择,这反映了语境在近义构式选择中的决定作用。影响本族语者与学习者指物关系代词选择的共性因素包括先行词的复杂度、确定性、修饰语数量、在从句中的句法成分,从句在主句中的嵌入方式、从句与先行词之间的距离、从句长度、从句谓语动词类型、从句语态以及主句的语态、肯定/否定。本族语者和学习者数据模型中与先行词和主句相关的因素和多个从句因素之间存在显著交互效应,这表明先行词和主句的复杂度对从句的复杂度具有一定程度的预测性。

认知加工负载对本族语者与学习者指物关系代词的选择具有显著影响。关系从句中某些部分认知加工负载增高时,对其他部分的低认知加工负载具有预测性,如在本族语者模型中,当从句以中央嵌入的方式插入主句中或先行词作从句中的宾语,即从句的认知加工负载较高时,先行词对于语义更为确定,即先行词低认知加工负载的预测概率升高。在学习者模型中,当从句与先行词之间有分隔,即认知加工负载较高时,先行词对于语义更为确定,即先行词低认知加工负载的预测概率升高;当从句以中央嵌入的方式插入到主句中,即认知加工负载升高时,对从句中系表动词使用,即低认知加工负载的预测概率升高(系表动词的认知加工负载低于其他类型的动词)。

研究发现,当语境变得复杂,即语境中的认知加工负载增大时,指物关系代词对于 *which* 的预测概率升高,如当从句长度增大时,本族语者与学习者都显著倾向于选择 *which* 作为关系代词。这验证了 Rohdenburg (1996: 149) 提出的复杂度原

则：当语言环境变得复杂时，语言使用者倾向于使用形式上更为明确的语言形式，复杂语境包括各种不连续、间断的构式，被动构式以及主语、宾语及从句的长度。Rohdenburg (1996: 172) 明确指出，对于现代英语，在关系从句中，wh-型关系代词在语法形式上更为显著。因此，基于复杂度原则，当从句长度增大，即语境变得复杂时，语言使用者倾向于选择形式更为明确的 which 作为指物关系代词。笔者认为，相较于 that，which 作为指物关系代词在形式上更为明确，这可能与 that 能承担更多样的语法功能有关。that 不仅可以用作连接词，还可以用作代词、限定词；作为连接词，不仅可以引导关系从句，还可以引导名词性从句。

## 5.2 中国英语学习者与英语本族语者在指物关系代词选择中的差异性制约因素

总体而言，英语母语者在指物关系代词选择中更倾向于使用 that，这一趋势在美国英语中尤为显著。在百万词规模的 CLOB 语料库中，我们共提取出 553 例由 which 引导的限定性关系从句，以及 2,304 例由 that 作为指物关系代词引导的限定性关系从句。在这些语境中，which 与 that 在语法上均可以换用。而在同样库容的美语语料库 Crown 中，which 引导的限定性关系从句仅有 132 例。相比之下，中国英语学习者并未表现出明显偏好。在 TECCL 语料库中，共有 1,442 例由 which 作为指物关系代词引导的限定性关系从句，以及 1,351 例由 that 引导的对应结构。

英语本族语者和中国英语学习者对指物关系代词的选择在若干因素水平上存在显著差异，这种差异性表现在以下三方面。

第一，本族语者与学习者在某些变量水平上对 which 与 that 的选择倾向不同，学习者在某些因素水平上做出不符合本族语者表达习惯的关系代词选择。例如，当先行词结构为“限定词+中心名词”时，母语者明显偏好使用 that，这一点与权威语法书的描述一致，而学习者则更倾向于选择 which。当先行词为不定代词时，母语者显著偏好使用 that，而学习者则未表现出明确倾向。这可能与部分学习者不了解本族语者的语言使用规范有关，因此，在教学中应通过对比示例与引导性产出训练，帮助学习者形成更接近母语者的地道表达。

第二，尽管本族语者与学习者在指物关系代词选择上受到相同因素的制约，但这些因素对两个语言群体的影响权重不同。例如，从句长度对本族语者与学习者而言均为显著预测因素，但其在母语者关系代词选择的显著性影响因素中排序靠前，在学习者关系代词选择的影响因素中排序靠后，这表明同一因素对两个语言群体的影响权重不同。

第三，与英语本族语者相比，中国英语学习者对由语境变量引发的认知加工负载变化更为敏感。学习者对从句中的否定结构、由于插入成分导致的先行词与从句

之间距离增加、从句以中央嵌入的方式插入到主句中等从句认知加工负载升高的情况更为敏感。从句的肯定/否定虽未显著影响本族语者的关系代词选择，但其通过与其他因素的交互，对学习者的关系代词选择产生了显著影响。当关系从句以中央嵌入方式插入主句，从而导致认知加工负荷较高时，本族语者倾向于使用 *that* 作为关系词。相比之下，学习者更倾向于选择语法标记更为显性的 *which* 作为指物关系代词，并偏好使用认知加工负荷较低的系动词充当从句谓语，通过降低语境中其他部分的加工负担，以应对中央嵌入结构所带来的高认知负荷。

### 5.3 对指物关系代词选择的构式语法解读

本文对影响指物关系代词选择的多种词汇—语法与语义因素及其交互效应的分析表明，关系代词 *which* 与 *that* 的选择是多重语境因素共同作用的结果。与两种关系代词共现的先行词、关系从句及主句在语义与语用层面所呈现出的系统性差异，有力支持了 Goldberg (1995, 2019) 提出的“无同义构式原则”和“统计优先假设”：形式不同的语法构式必然在语义或语用功能上存在差异，这凸显了语境在近义构式选择中的决定性作用。正如 Goldberg (2019: 86-87) 所指出，当近义构式并存时，语言使用者会逐步习得其适用语境的差异。在本研究所涉及的指物关系代词引导的限定性关系从句中，语言使用者在 *which* 与 *that* 之间的选择，依赖于来自先行词、关系从句及主句的多重语境线索。例如，当先行词为不定代词时，本族语者显著偏好使用 *that*；而当关系从句长度增加时，无论本族语者还是学习者均更倾向于选择 *which*。

此外，Goldberg (2019: 8) 指出，语言使用者在表达过程中需要在表达性 (expressiveness) 与高效性 (efficiency) 之间取得平衡。在语境不确定性较高或结构复杂度增加时，语言形式之间的差异会被放大，语言使用者倾向于通过选择更为显性的表达形式来降低歧义。因此，当关系从句长度增加、语境复杂度提升时，本族语者与学习者均倾向于选择形式更为明确的 *which*，以增强表达的清晰度。这一发现亦与 Rohdenburg (1996) 提出的“复杂度原则”相一致，即在复杂语境中，语言使用者更倾向于采用形式上更为显性的变体 (more explicit variants)。

### 5.4 教学启示

关系从句一直是中学英语语法教学中的重点与难点。在实际教学与使用过程中，学生乃至部分教师在选择关系代词时，往往较大程度上依赖语言习惯，对近义关系代词 *which* 与 *that* 的差异缺乏系统认识。在近义结构辨析方面，当前英语教学中普遍存在理解不深入的问题，不少中国学习者误认为近义结构之间不存在本质差别，可以在任意语境中互换使用。这种误解在一定程度上亦源于词典及语法工具书对相关结构所作的概括性说明，缺乏对其使用条件的细致区分。

本文通过对近义关系代词使用差异的系统分析,旨在为我国英语教学,尤其是限定性关系从句的教学提供一定的启示。首先,教师应逐步培养并强化“构式意识”,以构式为基本单位组织教学,引导学生关注语言形式与意义、功能之间的对应关系,即语言的“形-义-用”一体性。需要明确的是,不同语言形式之间的差异必然对应着意义或语用功能上的差异,因此,近义构式并非本质等同,亦不能在所有语境中任意替换。其次,近义构式的选择并非由单一因素决定,而是词汇—语法、语义及语用等多重语境因素综合作用的结果。因此,在教学过程中,教师应引导学习者,尤其是高水平学习者,增强语境意识,关注符合母语者使用习惯的地道表达。可借助语料库等工具,为学生提供真实语境中的语言实例与针对性练习,促进其对语言形式的归纳与类化,从而形成更为系统的语言知识结构。

## 6 结语

本研究基于多因素分析方法,系统考察了中国英语学习者与英语本族语者在指物关系代词选择上的共性与差异。研究结果表明,无论是本族语者还是学习者,其对指物关系代词的选择均并非由单一因素决定,而是由与先行词、关系从句及主句相关的多种词汇—语法与语义因素共同作用的结果,且这些因素在特定条件下存在显著的交互效应。这一发现进一步印证了语境在近义构式选择中的决定性作用。与此同时,从整体分布来看,本族语者在指物关系代词选择上表现出对that的明显偏好,而中国英语学习者则未呈现出稳定的选择倾向。本族语者与学习者在关系代词选择上表现出多方面的显著差异,主要体现在以下三个方面。第一,在先行词词性、先行词复杂度等变量层面,两者对which与that的选择倾向存在差异,学习者在部分语境中所作选择偏离本族语者的常规使用模式。第二,尽管两类语言群体在关系代词选择上受到相似因素的制约,但各因素的影响权重存在差异。例如,从句长度对两者均具有显著影响,但在本族语者的影响因素排序中位居前列,而在学习者中则相对靠后,表明同一因素在不同语言群体中的作用强度存在差别。第三,与本族语者相比,中国英语学习者对由语境变量引发的认知加工负荷变化更为敏感,尤其体现在从句否定、先行词与从句之间的距离增加,以及关系从句以中央嵌入方式出现在主句中等情境。

本研究仍有进一步拓展的空间。其一,可采用双重回归差异分析(MuPDAR)对学习者和本族语者之间的差异程度进行更精细的量化考察,深入探讨各因素在不同变量水平上对两类群体语言选择差异的影响程度及其权重分布。其二,本研究结论可通过心理语言学实验加以验证,从而检验其心理现实性并增强解释力。同时,本研究在语料选择方面亦存在一定局限。本族语者语料采用平衡语料库CLOB,与学习者作文语料在体裁与语域上的可比性相对有限,而既有研究表明,语体正式程

度会影响关系代词 which 与 that 的使用，这一因素在未来研究中有必要进一步加以控制与考察。

## 参考文献

- Biber D, Johansson S, Leech G, et al., 1999. Longman grammar of spoken and written English[M]. Harlow: Longman.
- Goldberg A E, 1995. Constructions: a construction grammar approach to argument structure[M]. Chicago: University of Chicago Press.
- Goldberg A E, 2019. Explain me this: creativity, competition, and the partial productivity of constructions[M]. Princeton and Oxford: Princeton University Press.
- Huddleston R, Pullum G, 2002. The Cambridge grammar of the English language[M]. Cambridge: Cambridge University Press.
- Jespersen O, 1933. Essentials of English grammar[M]. London: Allen and Unwin.
- Quirk R, Greenbaum S, Leech G, et al., 1985. A comprehensive English grammar of the English language[M]. London: Longman.
- Rohdenburg G, 1996. Cognitive complexity and increased grammatical explicitness in English[J]. Cognitive linguistics, 7(2): 149-182.
- Wiechmann D, 2015. Understanding relative clauses: a usage-based view on the processing of complex constructions[M]. Berlin: Mouton de Gruyter.
- Xu J, Liang M, 2013. A tale of two C's: comparing English varieties with Crown and CLOB (The 2009 Brown family corpora)[J]. ICAME journal, 37: 175-183.
- 房印杰, 2016. 中国学生英语限定性关系从句中关系代词取舍的多因素分析[D]. 北京: 北京外国语大学.
- 李金满, 王同顺, 2007. 当可及性遇到生命性: 中国学习者英语关系从句使用行为研究[J]. 外语教学与研究 (3): 198-205.
- 薛熙哲, 2015. 中国学生万篇英语作文语料库 (V1.1) (Ten thousand English Compositions of Chinese Learners, Version 1.1, 简称 The TECCL corpus) [CP/OL]. (2019-09-24) [2026-05-28] <https://corpus.bfsu.edu.cn/info/1070/1449.htm>.

## 通信地址

100089 北京市 北京外国语大学中国外语与教育研究中心 (郝美佳)  
400715 重庆市 西南大学外国语学院 (唐瑞梁)

(责任编辑: 刘若冰)

# 俄罗斯主流媒体对李白形象的建构研究： 基于道琼斯语料库的话语历史分析

南开大学 杨玉琦

**提 要：**在全球化的时代背景下，李白作为中国古代最具影响力和代表性的诗人之一，其形象日益成为中外文明互鉴的重要载体。本研究选取2000—2025年道琼斯语料库中含有关键词“李白”的俄文报道为研究语料，采用话语历史分析与语料库相结合的研究方法，从高频实词、索引行和话语策略三个切入点出发，剖析俄罗斯主流媒体话语建构的李白形象。研究发现，俄罗斯媒体主要关注李白的文学造诣、诗歌风格、政治生涯等主题内容，并通过提名策略、述谓策略和视角化策略，建构了一个集卓尔不群的诗仙、纵饮不羁的酒仙、才华横溢的文士、潇洒随性的逸士、孤高傲然的狂士于一身的李白形象。此外，俄罗斯媒体话语建构的李白形象，本质上是文化符号在国际传播场域中的策略性转化，其学术价值已超越文学研究范畴，成为观察两国关系演变的重要棱镜。

**关键词：**李白形象；话语建构；话语策略；历史话语分析；俄罗斯媒体

## Constructing the Image of Li Bai in Russian Mainstream Media: A Historical Discourse Analysis Based on the Dow Jones Corpus

Yuqi Yang

**Abstract:** In the context of globalization, Li Bai as one of the most influential and representative poets of ancient China, has increasingly become a vital carrier for mutual civilizational dialogue between China and foreign cultures. This study selects Russian-language reports containing the keyword “Li Bai” from the Dow Jones Corpus between 2000 and 2025 as research material. Employing a combined methodology of discourse-historical approach and corpus-based analysis, it examines the constructed image of Li Bai in Russian mainstream media through three analytical dimensions: high-frequency content words, concordance lines, and discursive strategies. The findings reveal that

Russian media predominantly focus on themes such as Li Bai's literary achievements, poetic style, and political career. Through nomination strategies, predication strategies, and perspectivization strategies, they construct a multifaceted image of Li Bai that amalgamates the "Poetry Immortal", the "Wine Immortal", the erudite scholar, the free-spirited recluse, and the aloof maverick. Furthermore, the discourse-constructed image of Li Bai in Russian media essentially reflects the strategic transformation of cultural symbols within international communication fields. Its academic significance transcends the scope of literary studies, serving as a critical observational prism for analyzing the evolution of Sino-Russian relations.

**Keywords:** image of Li Bai; discursive construction; discursive strategies; historical discourse analysis; Russian media

## 1 引言

李白，字太白，号青莲居士，是中国古代最具影响力和代表性的浪漫主义诗人之一。1880年，俄国汉学家王西里（Васильев）在其编著的《中国文学史纲要》中，首次使用俄文 Ли Тайбо（李太白）译名，向俄语读者引介这位中国诗坛巨擘，开启了李白在俄国的接受序幕。随后，经阿列克谢耶夫（Алексеев）、吉托维奇（Гитович）、孟列夫（Меньшиков）和休茨基（Щуцкий）等多位汉学家的译介与耕耘，进一步促进了李白及其诗歌在俄罗斯的传播与接受（Ракова，2017：6）。

在当前全球化的时代背景下，中国文学的海外传播已超越单纯的文本译介层面，转向更深层的文化形象建构与国家话语权博弈。文记东（2016：144）指出，域外中国文学话语既是践行文化走出去战略的重要途径，也是国际社会解码中国文明基因的关键符码。作为中国古代诗歌美学的巅峰代表，李白的海外形象研究具有双重价值，既关涉古典文学当代传播的路径创新，亦折射文化他者认知中国的深层逻辑。目前国内外学界关于李白形象的研究主要体现为从传播学视角探讨李白形象的建构与国际传播（徐臻，2015；景若冰，2021），从比较文学视角考察国外文学作品中刻画的李白形象（熊晓霜，2013；李玉箫，2022），以及从译介学视角分析翻译作品中塑造的李白形象（Пороль，2019）。既有研究已从多学科对李白形象进行探索，但大多依赖定性分析，缺乏大规模文本数据的支撑，导致结论的客观性与解释力受限。鉴于此，本研究借鉴话语历史分析和语料库语言学的最新研究成果（Шилихина，2014；Лобанова，2019），构建语料库辅助话语历史分析的研究框架，以道琼斯语料库与“李白”相关的俄文话语为语料，聚焦俄罗斯媒体话语中的李白形象建构及其实现的话语策略，以期从“他者”视角探究李白形象如何跨越文化和历史的界限，在俄罗斯社会文化背景下传播、解读与接受。

## 2 研究设计

### 2.1 语料收集

道琼斯是世界一流的信息数据库之一，其整合了全球159个国家和地区发行的近3.6万种权威信息资源，囊括了俄罗斯大部分主流媒体电讯和报刊报道的相关信息，如«Известия»（《消息报》）、«Комсомольская правда»（《共青团真理报》）和«Российская газета»（《俄罗斯报》）等。本研究以道琼斯数据库为语料检索载体，在其语料检索页面将时间跨度设置为2000年1月1日至2025年6月30日，语料来源国家设置为俄罗斯，以“李白”的俄译名“Ли Бо”为关键词进行检索，共得到237条搜索结果。经人工删除26篇重复报道，以及与本研究无关的87篇报道，最终将124篇有效报道作为研究文本，以此构建俄罗斯媒体李白话语语料库，库容共计32,422个字符。

### 2.2 理论依据

话语分析的兴起是对传统结构主义语言学的补充，其不仅关注语法、语义和语用原则，同时深入探讨话语的产生过程、话语的使用情景以及话语与社会的关系（许家金，2019：11）。话语历史分析是奥地利批评话语分析家露丝·沃达克（Ruth Wodak）及其团队在探究欧洲反犹太主义语篇过程中形成的一种研究范式，旨在通过以下三个维度解析话语中个体或群体的身份表征：识别特定语篇话语的内容及主题、分析语篇使用的话语策略与探究话语生成的社会历史语境。本研究拟从话语历史分析的“建构”视角出发，将俄罗斯媒体表征的李白话语内容嵌入历史语境，从话语主题、话语策略和话语语境三个向度探究俄罗斯媒体话语所塑造的李白形象。

传统的话语分析大多基于语料进行定性考察，难以克服语料代表性不足和结论主观性过强的缺陷。1995年，哈特·莫特纳（Hardt-Mautner）首次将语料库研究范式引入批评话语分析中，语料库语言学与话语分析相结合的研究方法逐渐被相关学者所了解，如今已成为社会科学领域较为前沿的研究范式（钱毓芳，2010：198）。语料库辅助的话语研究建立在大规模语料抽样的基础之上，其定量与定性相结合的研究范式既强调运用语料库检索工具对大规模语料进行定量分析，又注重对具体文本语境中的语言现象进行定性话语解读，一定程度上能够弥补传统话语分析文本数量少以及研究者主观价值评判的缺陷，从而增强话语分析的客观性和可信度（刘立华等，2023：15）。

### 2.3 研究问题

本文聚焦俄罗斯媒体话语刻画的李白形象，主要围绕以下三个问题展开。

- (1) 俄罗斯媒体建构了怎样的李白形象？
- (2) 俄罗斯媒体使用何种话语策略建构李白形象？
- (3) 俄罗斯媒体建构李白形象所用话语反映了怎样的社会历史语境？

2.4 研究框架

基于上述研究问题，本研究借助语料库检索工具 AntConc，以话语历史分析法为理论框架，开展语料库辅助话语分析的实证研究，整体研究框架如图 1 所示。

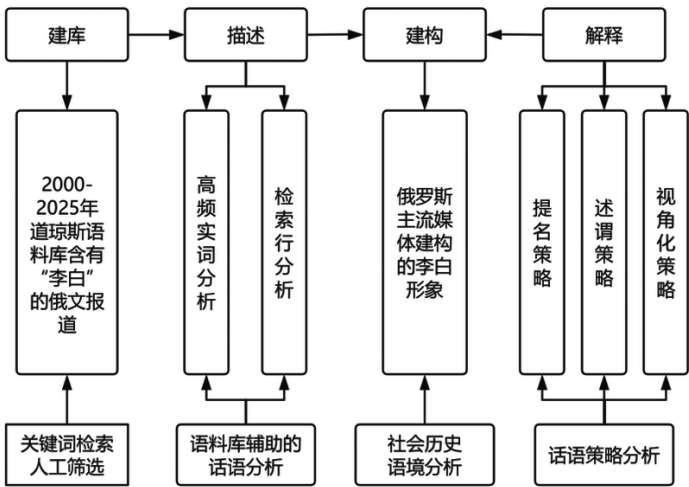


图 1 语料库辅助话语研究的多维分析框架

首先，本研究对自建俄罗斯媒体李白话语语料库中高频实词、索引行进行统计分析和可视化呈现，归纳关键话语信息，描述俄罗斯媒体话语所表征的李白形象。其次，分析俄罗斯媒体话语建构李白形象所用的话语策略，聚焦语料文本中的提名策略、述谓策略和视角化策略，考察俄罗斯媒体建构李白形象所使用的话语策略。最后，将研究发现置于中俄两国社会历史语境下，讨论话语建构的深层原因。

3 数据分析与研究发现

3.1 词汇频率分析

词频统计是语料库辅助话语分析研究的出发点。许家金（2019：64）认为，分析话语文本中反复出现的词汇，可揭示隐含在话语中的意识形态和价值取向。此外，在媒体话语中，实词可体现媒体聚焦的重点内容。对语料库实词进行统计分析，能够快速了解人物核心形象在公共话语空间中的角色定位，进而从多维度、多层次对人物形象进行描述和解读。本研究借助语料库检索软件 AntConc 的高频词统

计功能，对自建俄罗斯媒体李白话语语料库进行词频检索，去除检索结果中前置词、连接词、语气词等功能虚词，制作排名前30位的高频实词表（见表1）。

表1 俄罗斯媒体李白话语语料库高频实词（前30位）

| 次序 | 词目            | 频数 | 次序 | 词目            | 频数 |
|----|---------------|----|----|---------------|----|
| 1  | китайский     | 94 | 16 | вино          | 37 |
| 2  | поэт          | 81 | 17 | произведение  | 34 |
| 3  | быть          | 72 | 18 | каллиграфия   | 33 |
| 4  | китай         | 68 | 19 | язык          | 28 |
| 5  | стих          | 63 | 20 | писатель      | 28 |
| 6  | поэзия        | 62 | 21 | пиво          | 27 |
| 7  | великий       | 59 | 22 | книга         | 26 |
| 8  | литература    | 56 | 23 | презентация   | 24 |
| 9  | культура      | 55 | 24 | придворный    | 23 |
| 10 | пить          | 52 | 25 | стихотворение | 21 |
| 11 | гений         | 52 | 26 | мир           | 18 |
| 12 | романтический | 48 | 27 | год           | 18 |
| 13 | перевод       | 45 | 28 | писать        | 12 |
| 14 | лирический    | 42 | 29 | пьяный        | 10 |
| 15 | политик       | 38 | 30 | скиталец      | 8  |

俄罗斯语言学家埃加姆纳扎罗夫（Эгамназаров）（2018：185）认为，语言系统中的词汇并非彼此孤立，而是在某种支配性概念的聚合作用下，共同构成一个语义场。从表1可以看出，俄罗斯媒体李白话语语料库中高频词相互关联、相互作用，共同构成人物评价、文学造诣、诗歌风格、政治生涯和饮酒作诗五大核心语义场，对李白形象进行建构与呈现。

以“人物评价”语义场为核心，可以发现 поэт（诗人）、политик（政治家）、гений（天才）展现了李白在俄罗斯媒体话语中的崇高地位以及媒体对其多角度形象塑造。从“文学造诣”语义场来看，стих（诗歌）、стихотворение（诗句）、каллиграфия（书法）反映出俄罗斯媒体对李白在诗歌、散文、书法等文学领域作品的高度关注。就“诗歌风格”语义场而言，романтический（浪漫主义的）、лирический（抒情的）等高频词展示出李白诗歌的创作风格，构建了一个深情且极具浪漫主义色彩的诗人形象。观察与“政治生涯”语义场相关的高频词可知，придворный（朝廷的）、презентация（干谒）、скиталец（流浪者）揭示了俄罗斯媒体对李白形象的多维关注，既有其诗人的身份，也有其在政治与社会背景下的命运与选择。围绕“饮酒作诗”语义场，пить（喝）、вино（酒）、пьяный（醉酒的）等高频词揭示了李白在其创作过程中，通过饮酒寻求灵感、创作诗句。接下来，本研究将以高频实词提供的核心语义场为切入点，从索引行的搭配情况及其共现语境分析俄罗斯媒体话语建构的具体李白形象。

3.2 索引行分析

语料库索引借助检索功能，逐行呈现检索词及其在文本中的语言实例，进而帮助研究者考察检索词上下文语境间所传达话语发出者的情感倾向（黄昌宁 等，2007：194）。俄罗斯媒体对李白形象的话语建构必然围绕着“Ли Бо”一词进行刻画，故本研究以“Ли Бо”为检索词，在俄罗斯媒体李白话语语料库进行检索，随后点击索引行，回溯到检索词出现的具体语篇，并根据高频词所呈现的五大核心语义场，提取形成表2。

表2 俄罗斯媒体李白话语语料库“Ли Бо”索引行

| 左索引行                         | 检索词   | 右索引行                                                               | 主题 |
|------------------------------|-------|--------------------------------------------------------------------|----|
| /                            | Ли Бо | является <b>выдающимся представителем</b>                          | 人物 |
| опровергать тот факт, что    | Ли Бо | их <b>великий поэт</b>                                             | 评价 |
| китайское сравнение, что     | Ли Бо | <b>- поэт неба</b>                                                 |    |
| Наследие                     | Ли Бо | включает <b>классические стихотворения</b>                         | 文学 |
| “Три чуда” обозначают        | Ли Бо | поэзия, Пэй Минь фехтование, каллиграфия Чжан                      | 造诣 |
| Стиль поэзии                 | Ли Бо | <b>прост, лишен украшательства</b>                                 |    |
| /                            | Ли Бо | отличался <b>романтическим стилем</b>                              | 诗歌 |
| Луна у                       | Ли Бо | высокая и яркая, что соответствует его <b>романтическому стилю</b> | 风格 |
| Ссылка опальных писателей на |       |                                                                    |    |
| край ойкумены — того же      | Ли Бо | была для них трагедией удаления от центра власти                   | 政治 |
| Овидия                       |       |                                                                    | 生涯 |
| желание стать чиновником,    | Ли Бо | <b>не стал готовиться к экзаменам</b>                              |    |
| /                            | Ли Бо | <b>был приглашен ко двору</b>                                      |    |
| Гениальный безумец           | Ли Бо | черпал <b>свое вдохновение</b>                                     |    |
|                              |       | <b>в вине</b>                                                      | 饮酒 |
| у                            | Ли Бо | оказалось достаточно времени, чтобы <b>предаться</b>               | 作诗 |
|                              |       | <b>винопитию</b>                                                   |    |

观察俄罗斯媒体李白话语语料库中“Ли Бо”检索词索引行可以发现，俄罗斯媒体从以下多维图景建构李白形象。

其一，俄罗斯媒体对李白进行评价时，通过 **великий поэт**（伟大的诗人）、**поэт неба**（诗仙）、**выдающийся представитель «золотого века китайской поэзии»**（中国诗歌“黄金时代”的杰出代表）、**политический и общественный деятель**（政治和公众人物）给俄罗斯读者建构了一个在文学、政治、社会等多个领域均留下深远影响的李白形象，如下例。

（1）Ли Бо является выдающимся представителем «золотого века китайской поэзии», политическим и общественным деятелем.

李白不仅是中国诗歌“黄金时代”的杰出代表，同时还是一位政治家和公众人物。

(2) В разговоре с китайскими историками я не стал опровергать тот факт, что Ли Бо их великий поэт.

在与中国历史学家的对话中，我从未反驳李白是他们伟大的诗人这一事实。

(3) Например, приводится китайское сравнение, что Ли Бо - поэт неба.

例如，中国人把李白称为“诗仙”。

例(1)通过指称李白为中国诗歌“黄金时代”的杰出代表，彰显出李白在中国文学史中“高山仰止，景行行止”的杰出地位。политический и общественный деятель (政治和公众人物) 则表明李白不仅专注文学创作，同时在政治和社会领域均有所建树，从而多维度地构建了一个对政治活动和社会事件有着深刻洞察力，并取得卓越贡献的人物形象。从例(2) не стал опровергать (没有反驳) 这一动词词组可以看出，即使身处不同的语言系统和文化体系，李白的诗歌才华也在跨文化交际中，得到了海外学者“他者”视角中的认可与尊重。其对李白“伟大诗人”这一形象不容置疑的认可，印证了李白诗歌作品跨越了地域和文化的界限，成为全人类共同的文化瑰宝。例(3)中 поэт неба (诗仙) 一词可使俄罗斯读者感知李白在世人眼中是一个神仙化的存在。在中国古代文化中，“仙”常用于描述超脱凡人的特质(彭兆荣，2015: 60)，“诗仙”这一称号是对李白杰出的文学成就和卓越的诗歌艺术的最高赞誉。正如杜甫《春日忆李白》诗中“白也诗无敌，飘然思不群”所言，“诗仙”这一指称，建构了一个飘然若仙、卓尔不群的诗人形象。

其二，俄罗斯媒体在探究李白的文学造诣时，Три чуда (三绝)、стихи в жанре народных песен «юэфу» (乐府)、четверостишия и ритмическую прозу «фу» (辞赋) 等词组的使用，介绍了李白在创作不同体裁诗词方面的卓越建树，如下两例。

(4) Наследие Ли Бо включает классические стихотворения, стихи в жанре народных песен «юэфу», четверостишия и ритмическую прозу «фу».

李白给后世留下了古风、乐府、绝句和辞赋等诗歌形式的文学作品。

(5) Выражение “Три чуда” обозначают Ли Бо поэзия, Пэй Минь фехтование, каллиграфия Чжан Сюя.

李白的诗歌、裴旻的剑舞、张旭的草书被誉为大唐“三绝”。

例(4)通过列举李白多种形式的诗歌作品，包括古风、乐府、绝句和辞赋

等，为俄罗斯读者勾勒出了一个在不同诗歌体裁中均有所建树的诗人形象。这一多样化的诗歌体裁驾驭能力不仅体现了李白对传统诗歌形式的继承与创新，更突显其深厚的文学造诣和卓越的文化素养。例（5）中 Три чуда（三绝）出自《新唐书·文艺传》：“文宗时，诏以白歌诗、裴旻剑舞、张旭草书为‘三绝’”。唐文宗李昂曾向天下颁布诏书，李白、裴旻和张旭三位名士在文化领域取得了登峰造极的成就，御封其为“唐代三绝”。该例句不仅展示出李白“笔落惊风雨，诗成泣鬼神”的高超文学天赋，同时也体现出尽管唐朝文化发展兴盛、不断涌现出诸多优秀诗人，但无人可媲美李白的杰出成就，进一步突显李白在唐朝文坛上的崇高地位，塑造了一个才华横溢、享有盛誉的文士形象。

其三，俄罗斯媒体在表征李白的诗歌风格时，通过 прост（纯粹的）、лишен украшения（没有过多修饰）、романтический стиль（浪漫主义风格）等词汇构建了一个诗歌作品不依赖于华丽辞藻的修饰，但又不乏诗意与浪漫的诗人形象。

（6）Стиль поэзии Ли Бо прост, лишен украшения.

李白诗风简朴，无须过多修饰。

（7）Ли Бо отличался романтическим стилем письма, и теперь получил звание бессмертный поэт.

李白的诗歌有着浪漫主义的特点，他被誉为“不朽的诗人”。

（8）Луна у Ли Бо высокая и яркая, что соответствует его романти-ческому стилю.

李白笔下高耸明亮的月亮，符合他的浪漫主义风格。

例（6）中的 прост（简朴）和 лишен украшения（没有过多修饰）向俄罗斯读者介绍了李白诗歌创作时所用语言质朴，不追求过度的雕琢修饰。正如其诗句“清水出芙蓉，天然去雕饰”所传达的清新隽永、文雅脱俗的诗歌风格。例（7）和例（8）通过 романтический стиль（浪漫主义风格）强调了李白诗作具有强烈的浪漫主义色彩。李白是中国文学史上继屈原之后又一位伟大的浪漫主义诗人。李白在继承前辈浪漫主义创作风格的同时，极大地丰富了浪漫主义创作的艺术技巧，并使浪漫主义精神和表现手法在诗句中达到高度统一，将浪漫主义艺术创作推向一个新的高峰（李连发，2003：179）。从《梦游天姥吟留别》中“天姥连天向天横，势拔五岳掩赤城”可见李白常以热情奔放的语言、瑰丽的想象和奇特的比喻塑造形象。这一浪漫主义的诗歌创作手法也间接塑造了一个注重个人情感、追求思想自由、不受传统束缚的浪漫主义诗人形象。

其四，俄罗斯媒体在阐述李白的政治抱负和抉择时，通过 опальный писатель

(被贬黜的作家)、**сильное желание стать чиновником** (强烈的入仕想法)、**не стал готовиться к экзаменам** (不为科举考试做准备)、**ушел из дворца** (离开皇宫)等短句,构建了一个具有政治理想,希望通过从政实现自己“如逢渭川猎,犹可帝王师”的人生抱负,但始终坚持自己的原则,不为权贵折腰的诗人形象,如下例。

(9) Ссылка опальных писателей на край ойкумены — того же Овидия и Ли Бо — была для них трагедией удаления от центра власти и культуры.

和奥维德一样,李白也是一位遭受政治迫害被发配边疆的诗人,这对他们来说是一场远离权力和文化中心的悲剧。

(10) Несмотря на сильное желание стать чиновником, Ли Бо не стал готовиться к экзаменам на государственную службу.

尽管李白非常希望步入仕途,但他并没有参加科举考试。

(11) Ли Бо был приглашен ко двору в качестве придворного поэта, однако император не стал слушать про политику, требуя лишь стихов о красоте своих наложниц и вскоре просто ушел из дворца.

李白曾应召进宫随侍作诗,但皇帝无心听李白关于朝政的看法,只要求他做一些赞扬妃嫔美貌的诗句,没过多久李白便离开了皇宫。

从例(9)可以看出,李白和古罗马诗人奥维德同样经历了被流放的悲惨命运。当一个文人,尤其是李白和奥维德这样的文学巨匠被迫离开政治和文化中心时,这种边缘化和孤立化的遭遇对其无疑是沉重的打击。此句为读者展现了一个即使在文学领域有着非凡成就,也无法免受政治波折的影响,在政治漩涡中满怀无奈与悲情的文人形象。例(10)中的**сильное желание стать чиновником** (强烈的入仕想法)说明李白曾有十分明确的政治抱负,却放弃实现这一理想最为便捷的途径。“学而优则仕”一直是中国古代文人的信念,对于绝大多数文人来说,苦读四书五经、百阅圣贤之书,为的是有朝一日能够金榜题名、入朝为官,施展自己的才华。该例句表明李白不屑于一般读书人通常选择的科举道路,相信凭借自己的才华依然能够登台拜相。这在一定程度上建构了一个不随波逐流、孤高傲然的狂士形象。例(11)中的**был приглашен ко двору в качестве придворного поэта** (被邀请进宫随侍作诗)说明李白曾受玄宗赏识,被授予翰林待诏一职,一时风光无两。但在面对皇帝无心朝政,只要求他写诗赞美妃嫔的美貌时,他选择坚持自己的立场,忠于自己的内心,最终辞官离开了宫廷。正如杜甫诗中所写“天子呼来不上船,自称臣是酒中仙”,尽管李白身处权力中心,也曾渴望步入政坛,但其并没有在权力和地位面前妥协自己的信仰。体现出李白在“安能摧眉折腰事权贵,使我不得开心颜”一

句中表达的对自由、洒脱和独立精神的追求，向俄罗斯读者展现了一个潇洒随性的逸士形象。

其五，俄罗斯媒体在解读李白与酒之缘时，通过 черпал свое вдохновение в вине（从酒中获取灵感）、предаваться винопитию（沉醉于饮酒）、восемь бессмертных винной чаши（醉八仙）等表达，构建了一个以酒入诗、结识兴趣相投的酒友诗客，一同“狂醉于花月之间”的“酒仙”形象，如下例。

(12) Гениальный безумец Ли Бо черпал свое вдохновение в вине.

酒是狂人李白创作的灵感源泉。

(13) У Ли Бо оказалось достаточно времени, чтобы предаваться вино-питию и наслаждаться обществом друзей - единомышленников. Они называли себя “восемь бессмертных винной чаши”.

在李白的生活中，时刻有酒和朋友的相伴，他们自称为“醉八仙”。

例(12)表明，酒对李白来说不仅是简单的生活调剂品，更是其文学创作时的灵感源泉。李白将酒作为生活与艺术间的桥梁，通过饮酒激发创作灵感。正如其酩酊大醉后所作的千古绝唱“人生得意须尽欢，莫使金樽空对月”，展现诗人在酒中寻找灵感和诗意，进而道出其及时行乐的生活态度以及豁达洒脱的人生哲思。例(13)中，восемь бессмертных винной чаши（醉八仙）出自杜甫的《饮中八仙歌》，指唐开元年间长安市上八位嗜酒名士，即李白、贺知章、李适之、李璡、崔宗之、苏晋、张旭和焦遂，诗中描绘了“醉八仙”的酒后之态。盛唐时期，饮酒之风盛行，文人常呼朋唤友尽兴痛饮，以彰显其名士风流（詹福瑞，2020：51）。该例句中“醉八仙”的典故为俄罗斯读者展示出李白等人嗜酒如命、纵饮不羁的性格特点。他们以酒结友，不拘泥于世俗礼仪、不追求权力地位，以免陷入官场的纷争和世俗的枷锁，从而保持心灵的自由和创作的独立。

### 3.3 话语策略分析

话语策略分析是话语历史分析法的主要环节之一，本研究将从提名策略、述谓策略和视角化策略三个角度出发，探究俄罗斯媒体对李白形象建构的话语策略。

#### 3.3.1 提名策略

提名策略是指运用语言手段对语篇中的社会活动者、事物现象、行为过程等进行指称，以建构某一个体或群体的身份形象（Wodak et al., 2009：94）。本研究借助语料库检索工具对俄罗斯媒体指称“李白”的相关词汇进行检索，并将检索结果通过词云图进行可视化呈现。词云图中单词的大小取决于其在文本语料中出现的频率，频率越高的单词字号越大，同时由中心向四周扩散。

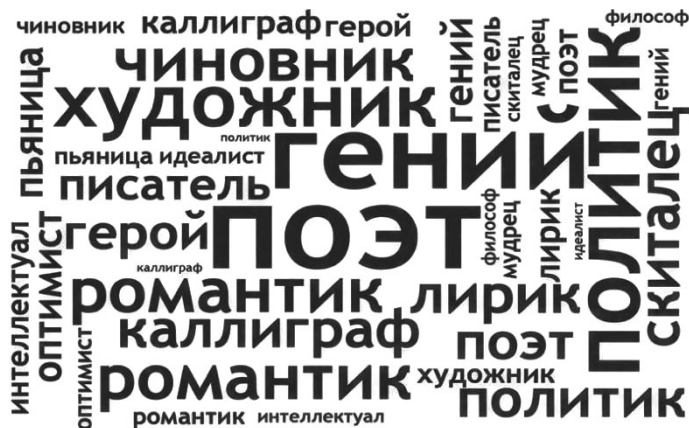


图2 李白形象高频词云

观察图2,除标准译文 Ли Бо 外,俄媒还广泛使用 поэт (诗人)、гений (天才)、политик (政治家)、скиталец (漂泊者)、романтик (浪漫主义诗人)、каллиграф (书法家)、художник (艺术家)、чиновник (官员)、лирик (抒情诗人)、пьяница (酒鬼)等词汇对李白进行指称。可见,俄罗斯媒体频繁运用提名策略对李白形象进行多维度解读和呈现,其指称内容跨越政治、文学、艺术、生平等多个方面。这一多视角的指称符合当代读者对历史人物立体化和生动化诠释的需求,有助于俄罗斯民众更为客观且全面地欣赏李白的诗歌风格、了解李白的人生经历、洞察李白的多元身份,进而能够近距离感知李白形象。

### 3.3.2 述谓策略

述谓策略是指使用形容词、副词、介词短语、从句等评价性话语,对表征的对象进行话语修饰和限定,以赋予其积极或消极的情感态度(田海龙,2009:85)。结合研究语料特点,本研究依托语料库检索工具 AntConc 的检索结果,制作俄罗斯媒体对李白形象的评述性词表(见表3)。

表3 李白形象评述性词

| 类型   | 示例                                                                                                                                                                   |
|------|----------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| 形容词  | талантливый, великий, лирический, романтический, философский, блестящий, независимый, китайский, выдающийся, популярный, гениальный, вдохновенный                    |
| 副词   | искренне, безудержно, умно, творчески, эксцентрично, глубоко, эмоционально, смело, энергично, свободно, легко, страстно                                              |
| 名词结构 | с великими предшественниками Тао Юаньмином и Цюй Юанем, представитель золотого века китайской поэзии, один из самых почитаемых китайских поэтов                      |
| 动词结构 | ставят в один ряд с Данте и Петраркой, стоит в ряду крупнейших мировых лириков, был сильно пьян, чиновником мечтает быть, побывал в знаменитых горах и у истоков рек |
| 从句   | как для русских – Пушкин, как для англичан – Шекспир                                                                                                                 |

从表3可以看出，俄罗斯媒体广泛使用表意积极、肯定的词汇表征李白的诗歌创作成就及其个人形象。此类评价性话语汇聚成一个更为具象的正向语义场，如 талантливый（才华横溢的）、свободно（自由地）、чиновником мечтает быть（成为一名官员）、один из самых почитаемых китайских поэтов（中国最伟大的诗人之一）、побывал в знаменитых горах и у истоков рек（遍访中国的名山大川）等，与上述提名策略协同塑造出一个兼具文采斐然的诗人、逍遥自在的游侠、满腔抱负的官员等多元身份的李白形象。

(14) Цюй Юань, Тао Юаньмин, Се Линьюн, Ли Бо - имена крупнейших китайских поэтов, живших в древности и в средние века, чьи произведения переводились на русский язык.

屈原、陶渊明、谢灵运、李白是中国古代和中世纪最为杰出的诗人，他们的作品被翻译成俄语。

(15) Ли Бо - один из самых почитаемых китайских поэтов. Его имя ставят в один ряд с Данте и Петраркой, Гете и Шиллером, Пушкином и Шекспиром.

李白是中国最受尊敬的诗人之一，他与但丁、彼特拉克、歌德、席勒、普希金和莎士比亚齐名。

(16) Ли Бо - всё для китайцев, как для русских — Пушкин, а для англичан — Шекспир.

李白对于中国人来说，就像普希金之于俄罗斯人、莎士比亚之于英国人一样重要。

此外，俄罗斯媒体通过述谓策略将李白与屈原、莎士比亚、普希金等著名文人相提并论的话语值得进一步探究。屈原、但丁、歌德和莎士比亚不仅是各自国家的文学巨擘，他们的作品和思想也跨越了国界，影响了数百年来世界文学的发展进程，塑造了无数文人墨客的文学理想。俄罗斯媒体通过例（14）和（15）将李白与这些文学巨匠并列，不仅强调了李白在中国文学史上举足轻重的地位，同时彰显出其作为中国文学的象征符号，在世界文学史上也具有一定的影响力。此外，例（16）将李白与普希金齐名，进一步表明俄罗斯媒体对李白文学造诣的认可，同时也突显其对中国文化的尊重。文化交流是建立在对其他文化认同的基础之上，而文化尊重又无疑是文化交流的前提（高佳红等，2022：22）。中俄山水相连，两国文化交流源远流长，例句这一表述展示出俄罗斯媒体对两国文化相互尊重的积极态度，对进一步强化中俄两国文化交流互鉴有积极的推动作用。

### 3.3.3 视角化策略

视角化策略通过直接引述、间接引述和自由引述等方式对话语生产者的观点进行立场化标记，可以使话语更为真实准确（赵林静，2009：89）。结合语料发现，俄罗斯媒体频繁使用视角化策略对李白形象进行建构，人工筛查语料库共得到39条引述结果。在这些语料中，多数倾向于通过直接引述李白的原诗，刻画其内心情感与性格特征，进而将抽象的诗人形象具象化到具体的历史与情感语境，现选取其中两例进行分析。

(17) Ли Бо почувствовал себя совсем одиноким и написал: Взмахну бокалом — приходи, луна! Ведь с тенью нас и вовсе будет трое.

李白感到孤独并写下这样的诗句：举杯邀明月，对影成三人。

(18) Боль по уходящему, переданная через образ текущей воды, просвечивает и в стихах Ли Бо, посвященных прощанию с друзьями: В лазури сирий парус тает белым клином, И лишь Река стремится за кромку облаков.

李白在其送别诗中，通过流水的意象传达离别的悲伤：孤帆远影碧空尽，唯见长江天际流。

例（17）中 Взмахну бокалом — приходи, луна! Ведь с тенью нас и вовсе будет трое（举杯邀明月，对影成三人）出自李白的《月下独酌》。公元744年，李白不羁于繁杂礼仪束缚，不愿违心奉承，遂解印辞官，游历名山大川。同年某夜，李白独自在月下徘徊，因政治失意而倍感孤寂忧愁，奈何寂寥无人可说，故作此诗。诗中李白与天边的明月和身旁的影子做伴共饮，使得原本冷清的场面瞬间热闹起来。这一诗句的引述体现了李白独自排解寂寞时旷达不羁的性格，构建出一个超然独立的诗人形象。例（18）则体现出李白对离别悲伤之情的刻画，通过流水的意象表达与友人离别时的不舍之情。诗中 В лазури сирий парус тает белым клином, И лишь Река стремится за кромку облаков（孤帆远影碧空尽，惟见长江天际流）描绘出一幅诗人伫立江岸，望着友人渐行渐远的帆影，最终只剩一江春水，浩浩荡荡流向远方水天交接之处的悲情画卷。诗句所述情景充分展示了李白对友人深情厚谊的重视，这一情感投入体现了李白人际交往和情感交流的深度，为后世对李白诗歌文学意象的解读提供了丰富的情感观照。综上，话语生产者以“他者”视角，直接引述李白所作诗句，可使话语建构的形象有据可循、有源可溯、有权威可依（乌楠等，2019：226），更符合客观历史事实，为俄罗斯读者展现了一个更加贴近历史的李白形象。

## 4 社会历史语境分析

话语是具有一定交际目的和功能的语言单位，具有形象建构的功能（曹静娴，2022：42）。话语的产生无法脱离特定的社会历史背景，故话语分析需观照话语所处的社会政治以及历史语境（陈晗霖，2023：861）。从前文分析可见，俄罗斯社会普遍采用积极正向的话语对李白形象进行建构。究其原因，从政治关系来看，2001年7月中俄两国签署《中俄睦邻友好合作条约》，条约第十六条指出：“缔约双方将大力促进发展文化、教育、卫生、信息、旅游、体育和法制领域的交流与合作”（中华人民共和国外交部，2001）。这一条约的签订，掀起了中俄人文交流合作的新浪潮。2005年7月，中俄两国元首共同确定2006年在中国举办“俄罗斯年”、2007年在俄罗斯举办“中国年”活动。“国家年”期间，两国共同举办500余项活动，使两国民众对彼此悠久的历史 and 灿烂的文化有了更深层次的了解（文记东，2016：141）。随着中俄两国战略协作伙伴关系的全面发展和不断深化，两国元首于2019年决定将两国关系提升为“新时代中俄全面战略协作伙伴关系”。李白作为中国古代诗人的杰出代表，其形象的传播和塑造无疑是一种较为有效的文化外交手段。在中俄两国政治关系稳步发展的背景下，俄罗斯媒体跨越语言和文化的边界，以积极的话语构建和诠释李白形象，一定程度上表达了对中国传统文学的尊重和认可，同时也深化了中俄文化间的交流，为构建两国信得过、靠得住、压不垮的友好合作关系奠定了坚实的文化基础（李哲，2022：78）。

从历史背景来看，中俄两国文化交流源远流长。中俄文化交流最早可追溯至13世纪，彼时成吉思汗西征，将先进的中国文化带到了罗斯诸公国（张玉侠，2009：59）。1741年俄国汉学家罗索欣（И.К. Россохин）从俄国东正教驻北京传教士团回国，随后进入圣彼得堡皇家科学院担任满汉语翻译一职，自此开启了俄国汉学的新篇章（阎国栋，2004：152）。新中国成立之初，苏联向中国提供了大量援助，两国之间的文化交流得到了空前的发展。21世纪初，俄罗斯总统普京多次访华，签署多项声明及条约，中俄两国的关系不断跃上新的台阶，文化交流也出现了前所未有的盛况。这一历史背景下，俄罗斯媒体以正向的话语向其民众介绍中国诗人李白，不仅呼应了中俄两国历史上的深厚友好基础，同时巩固了两国间的文化纽带，让中俄世代友好的理念深入两国民众之心。

## 5 结论

本研究以道琼斯语料库中2000年1月1日至2025年6月30日有关“李白”（Ли Бо）的俄文报道为研究语料，运用语料库辅助话语分析的研究范式，从高频实词、索引行和话语策略三个方面切入，剖析俄罗斯主流媒体话语建构何种“李白”形象

以及运用何种话语策略对其形象进行建构。研究表明,在语料库高频实词使用方面,俄媒主要从人物评价、文学造诣、诗歌风格、政治生涯和饮酒作诗五大核心语义场建构李白形象。在索引行分析方面,发现俄媒建构了一个集卓尔不群的诗仙、纵饮不羁的酒仙、才华横溢的文士、潇洒随性的逸士、孤高傲然的狂士于一身的“李白”形象。在话语策略方面,俄媒主要使用提名策略、述谓策略和视角化策略对李白形象进行建构。其中,提名策略通过对李白进行命名、提及和指涉,从多维度对李白形象进行解读和呈现,使俄罗斯读者能够近距离感知李白兼具浪漫主义诗人、满腔抱负的官员、潇洒随性的逸士等多元身份。述谓策略通过广泛使用积极肯定等正向评价性话语对李白的诗歌创作成就及其个人形象进行话语修饰和限定,使其建构的形象更加具象化。视角化策略通过引述李白所作的诗句,使建构的形象有据可循、有源可溯、有权威可依,更符合客观历史事实,为读者展现了一个更加贴近历史的李白形象。

在中国文学于海外广泛传播的当下,李白这一中国古代的文学巨匠,日益成为国际文化交流的焦点。通过话语历史分析,可以从“他者”视角理性地观察李白在异质文化国家中被话语建构的多元形象。此外,在关注国际文化场域对中国文学代表的形象表征与建构的同时,国内相关机构也应制定更为合理的方针与政策,积极开展与域外世界“他者”之间的沟通与对话,以期不断推动中国文学的海外传播与发展,致力于构建人类命运共同体的文化伟业(姚建彬,2020:9)。

## 参考文献

- Wodak R, De Cillia R, Reisigl M, et al., 2009. The discursive construction of national identity[M]. Edinburgh: Edinburgh University Press.
- Лобанова Т, 2019. Идеологические результаты дискурсивных практик китайских медиа: стратегия и методология исследования[J]. Вестник Московского государственного областного университета, 86(4):161-172.
- Пороль П, 2019. Эстетика перевода А. Ахматовой в стихотворении Ли Бо «На западной башне в городе Цзиньлин читаю стихи под луной»[C]. Россия в мире: проблемы и перспективы развития международного сотрудничества в гуманитарной и социальной сфере. М.: Пенза.
- Ракова А, 2017. Переводы поэзии эпохи Тан (на примере произведений Ли Бо)[J]. Тенденции развития науки и образования, 24(3):5-7.
- Шилихина К, 2014. Использование корпусов в исследованиях дискурса[J]. Вестник Воронежского государственного университета, 61(3):21-26.
- Эгамназаров Х, 2018. О понятии лексико-семантического поля в лингвистике[J]. Ученые записки Худжандского государственного университета им. академика Б. Гафурова, 54(1):185-189.
- 曹静娴, 2022. 俄罗斯媒体话语建构的孔子形象[J]. 中国俄语教学 (3): 42-51.
- 陈晗霖, 2023. 话语—历史分析: 趋势与展望[J]. 现代外语 (6): 1-11.

- 高佳红, 贺东航, 2022. 新时代人类文明交流互鉴的话语体系构建[J]. 学术探索 (7): 22-28.
- 黄昌宁, 李涓子, 2007. 语料库语言学[M]. 北京: 商务印书馆.
- 景若冰, 2021. 英美世界中的李白神话[D]. 上海: 华东师范大学.
- 李连发, 2003. 李白浪漫主义诗风探源[J]. 社会科学辑刊 (2): 178-179.
- 李玉箫, 2022. 从《克林格梭尔的最后夏天》看黑塞对中国文化的转向[J]. 黄冈师范学院学报 (1): 102-107.
- 李哲, 2022. 俄罗斯汉学在中俄文化交流中的历史价值和现实意义——俄罗斯档案馆馆藏中国文献典籍译本文献管窥[J]. 中国档案 (1): 76-78.
- 刘立华, 韦怡, 2023. 语料库视角下的话语研究[J]. 外国语言文学 (4): 15-24+133.
- 彭兆荣, 2015. 文化遗产关键词: 仙[J]. 民族艺术 (1): 57-62+70.
- 钱毓芳, 2010. 语料库与批判话语分析[J]. 外语教学与研究 (3): 198-202+241.
- 田海龙, 2009. 语篇研究: 范畴、视角、方法[M]. 上海: 上海外语教育出版社.
- 文记东, 2016. 21世纪中俄人文交流与展望[J]. 中国社会科学院研究生院学报 (6): 140-144.
- 乌楠, 张敬源, 2019. 中美企业机构身份的话语建构策略[J]. 现代外语 (2): 220-230.
- 熊晓霜, 2013. 从阿瑟·韦利和斯蒂芬·欧文看李白的西方形象[J]. 合肥工业大学学报(社会科学版) (3): 95-99.
- 徐臻, 2015. 日本汉诗对“谪仙”李白的接受[J]. 广东外语外贸大学学报 (3): 64-68+77.
- 许家金, 2019. 语料库与话语研究[M]. 北京: 外语教学与研究出版社.
- 阎国栋, 2004. 十八世纪俄国汉学之创立[J]. 中国文化研究 (2): 152-162.
- 姚建彬, 2020. 对中国文学海外传播的反思与建议[J]. 外国语文 (4): 1-10.
- 詹福瑞, 2020. 中国古代文学“生命意识”的传统与现代——以李白诗文研究为中心的讨论[J]. 文学遗产 (5): 172-180.
- 张玉侠, 2009. 早期中俄文化交流中的东正教因素[J]. 西伯利亚研究 (1): 59-61.
- 赵林静, 2009. 话语历史分析: 视角、方法与原则[J]. 广东外语外贸大学学报 (3): 87-91.
- 中华人民共和国外交部, 2001. 中华人民共和国和俄罗斯联邦睦邻友好合作条约[EB/OL]. (2001-07-16) [2026-04-30]. <https://www.fmprc.gov.cn>.

## 通信地址

300071 天津市 南开大学外国语学院

(责任编辑: 李添豪)

# 东南亚高级汉语学习者介词框式结构习得实证研究<sup>\*</sup>

吉林外国语大学 刘博洋 徐俊丽

**提 要：**本文基于HSK动态作文语料库，以东南亚地区泰国、越南和印度尼西亚高级汉语学习者为研究对象，运用Python编程实现语料分词、介词框式结构提取和标注介词偏误类型，并利用SPSS27.0对数据进行量化统计，探讨了学习者语言水平、国别与介词框式结构使用正确率和偏误类型之间的关联性。结果显示，高级学习者在介词使用准确性和语境适应性方面表现突出，其遗漏和语义偏误率显著低于初、中级学习者；不同母语背景学习者呈现不同偏误模式，泰国学习者冗余偏误显著，越南学习者错序偏误突出，印度尼西亚学习者则以遗漏偏误为主；不同语言水平学习者习得过程呈现阶段性特征。基于上述发现，研究总结了针对国别可供推广的学习策略，即注重学习者结构化输入、梯度学习和加强元认知监控；并提出针对语言水平和国别的教学建议，为东南亚汉语学习者的介词框式结构教学提供了实证参考。

**关键词：**东南亚；高级汉语学习者；介词框式结构；习得效果

## An Empirical Study on the Acquisition of Prepositional Frame Structures by Advanced Chinese Learners From Southeast Asia

Boyang Liu and Junli Xu

**Abstract:** This article focuses on advanced Chinese language learners in Southeast Asia,

<sup>\*</sup> 本文系2024年度吉林外国语大学校级重大科研项目“汉语二语高级学习者语言表现研究”（项目编号：JW2024ZDA004）、2025年度吉林外国语大学研究生科研项目“大语言模型视域下东南亚汉语学习者介词框式结构习得研究”（项目编号：JWXSKY2025A030）的阶段性研究成果。

刘博洋为本文通信作者。

作者贡献：

刘博洋：选题构思、研究方法、讨论结论、论文撰写、字数占比（40%）；

徐俊丽：数据收集、数据分析、初稿撰写、修改润色、字数占比（60%）。

specifically those in Thailand, Vietnam, and Indonesia. Based on the HSK Dynamic Composition Corpus, Python programming was used to segment the corpus, extract prepositional frame structures, and annotate types of prepositional errors. SPSS 27.0 was employed for quantitative data analysis to explore the relationships among learners' language proficiency, nationality, and the accuracy and error types in their use of prepositional frame structures. The results show that advanced learners perform well in preposition accuracy and contextual appropriateness, with omission and semantic error rates significantly lower than those of beginner and intermediate learners. Learners from different L1 backgrounds exhibit distinct error patterns: Thai learners show prominent redundancy errors, Vietnamese learners frequently make sequence errors, and Indonesian learners primarily make omission errors. The acquisition process also displays staged characteristics across different proficiency levels. Based on these findings, the study summarizes nationally applicable learning strategies, emphasizing structured input, gradual learning, and enhanced meta-cognitive monitoring, and proposes teaching suggestions tailored to learners' proficiency and nationality, providing empirical references for teaching prepositional frame structures to Southeast Asian Chinese learners.

**Keywords:** Southeast Asia; advanced Chinese learners; prepositional frame structures; acquisition effects

## 1 引言

介词框式结构作为汉语特有的语法现象,是由前置词和后置词共同构成的一个完整的句法-语义单元(Greenberg, 2005)。国内学者从不同角度对其进行了系统研究。陈昌来(2002)从句法功能角度指出其具有介引名词性成分和限定语义关系的双重功能。刘丹青(2002)从类型学视角提出其“前置词+X+后置词”的典型结构模式,强调其句法标记功能。贾君芳等(2020)则进一步提出“双向依存性”理论,认为介词框式结构在语义上具有不可分割的整体性。这种结构的完整性对二语学习者构成特殊挑战,也是二语教学中的一个难点。在二语习得领域,黄理秋等(2010)研究了英美、日韩学习者常用介词框式结构的使用频率与偏误率,结合构式语法理论分析习得难度序列。周文华等(2011)主要考察各类介词框式结构的普遍习得规律,结合学习者介词框式结构正确率划分习得阶段,发现空间介词习得早于抽象介词,且不同结构习得难度存在显著差异。周文华(2013)又深入揭示了母语迁移的具体表现,如韩语母语者受SOV语序影响易产生语序偏误。崔希亮(2005)揭示了英语母语者常见的近义介词混淆,以及东南亚学习者典型的成分遗漏等现象。值得注意的是,学习者的母语类型与其介词偏误类型呈现系统性对应关

系,即泰语、越南语等介词前置类型母语者倾向于遗漏后置词,而缅甸语等介词后置类型母语者则更多出现前置词误用。

尽管已有研究探讨了汉语作为第二语言的介词框式结构习得,但研究多集中于母语为英美、日韩国家的汉语学习者,而东南亚地区的汉语学习者习得介词框式结构的情况少有研究。此外,我们发现,现有的习得研究多集中于初级和中级阶段,以偏误分析为主,高级阶段的研究较少,且正确分析较少,多集中于错误分析。本研究以泰国、越南和印度尼西亚等东南亚国家高级汉语学习者为研究对象,以HSK动态作文语料库为语料来源,通过分析不同语言水平、不同国别学习者介词框式结构正确率和偏误类型比率数据,揭示东南亚地区高级汉语学习者介词框式结构的习得优势和习得困难,提出有效的学习策略,为介词框式结构的语法教学提供参考。

## 2 研究方法

### 2.1 研究问题

本文基于HSK汉语水平考试作文语料库,探究东南亚高级汉语学习者介词框式结构的使用情况。主要围绕以下五个问题进行。

(1) 高级汉语学习者是否仍然在介词框式结构使用中偏误较多?

(2) 不同语言水平学习者偏误类型分布有何差异?

(3) 不同母语背景学习者在介词框式结构偏误类型上是否表现出特定的优势或困难?

(4) 是否可以通过分析习得成功案例总结出具有推广意义的学习策略?

(5) 如何结合学习者的策略设计更有效的教学方法,以促进介词框式结构的习得?

### 2.2 语料选择

本文选取北京语言大学HSK动态作文语料库作为研究的语料来源,数据真实性与可信度较高。为聚焦东南亚地区汉语学习者,语料筛选严格限定为母语为东南亚国家官方语言(越南语、泰语、印尼语)的留学生,排除其他区域(欧美、日韩)学习者的作文,以确保研究对象的同质性。

本研究依据HSK动态作文语料库中学习者的语言能力等级字段,结合旧版HSK考试体系与《国际中文教育中文水平等级标准》(2021)中的词汇等级标注体系,对学习者的汉语水平进行分级。沿用旧版HSK高等考试中获得A、B、C等级证书的标准(相当于《国际中文教育中文水平等级标准》的9—11级),代表学习者具备较高的汉语综合运用能力,对应高级组。未获得HSK高级证书的学习者,其语言能力总体表现为低级至中级水平。为进一步细分两组水平,本研究对选取的

40个未获得高级证书的学习者根据得分进行分组。40个未获得高等证书的学习者，其平均得分为270分。以此为界，将270分及以上的学习者划分为中级组，他们具备较为稳定的语法和表达能力，将270分以下的学习者划分为初级组，其语言输出整体表现为词汇与语法结构较为基础、句式简单、语篇长度较短。

HSK 高等作文的主题覆盖个人生活、社会议题、文化对比三类，这避免了单一主题对语言表达的局限性，确保了语料在话题广度与语言复杂度上的代表性。

2.3 数据处理

鲁健骥（1994）将介词框式结构偏误类型分为遗漏、误加、误代、错序四大类。黄理秋等（2010）将介词框式结构偏误分为句法结构偏误和语义偏误两大类。本研究在鲁健骥经典偏误分类基础上，结合黄、施二人的研究及东南亚汉语学习者的偏误特征，采用遗漏、冗余、错序、语义偏误四分法。确定分类方法后，我们利用 Python 对语料进行分词和词性标注，通过系统初步识别介词框式结构，生成带有介词框式结构标记的文本文件。借助人工校验 Python 标注的结果，补充了未识别出的介词框式结构，在此步骤我们共得到 229 例有效语料。接着我们按照遗漏、冗余、错序、语义偏误四分法利用 Python 再次对 229 例语料进行偏误类型分类，系统自动分类后，手动计数每种偏误类型出现的次数和介词框式结构总使用次数，按照语言水平，整理进 Excel 表格，利用 Excel 计算学习者正确率及偏误类型比率，以备后续数据分析。

关于利用 Python 如何对作文语料进行介词框式结构特征提取和偏误标注，步骤如图 1 所示。

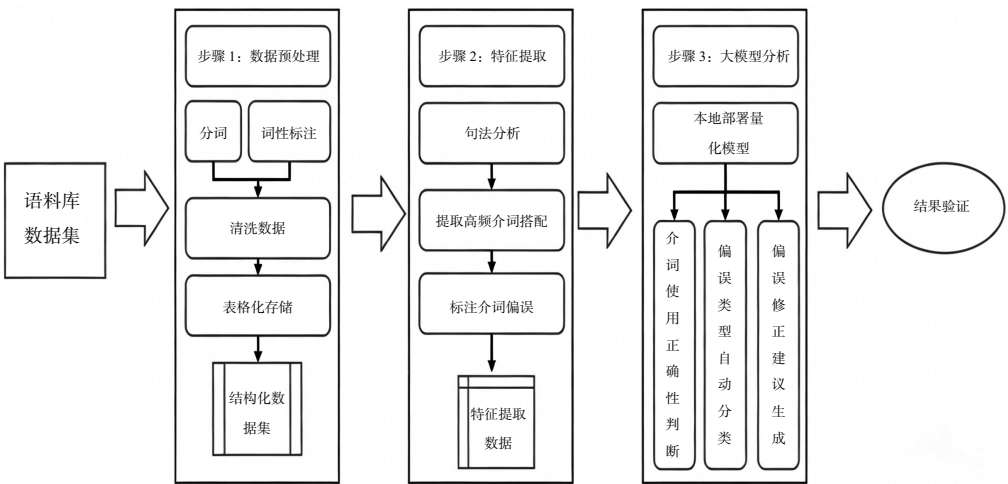


图 1 Python 处理数据流程

利用 SPSS 来处理变量之间的关系，如用方差分析来探究语言水平对准确率的影响，用卡方检验来探讨国别与偏误类型的关联。

通过以上分工，利用 Python、SPSS 和 Excel 最大化工具效率，人工校验保证数据质量，以此保证数据分析的有效性和准确性。

3 数据分析

本研究的数据分析部分系统考察了东南亚不同水平汉语学习者介词框式结构习得的正确率、偏误类型、使用频率、多样性。基于以上四个维度，主要考察了三组关键数据关系：语言水平与正确率和偏误类型的关联性、国别与偏误类型的关联性、语言水平和国别对介词框式结构使用频率及多样性的影响。具体分析如下。

3.1 不同水平学习者介词框式结构正确率和偏误类型比率分析

为探究高级汉语学习者是否仍然在介词框式结构使用中偏误较多，本研究将其与初级、中级汉语学习者进行了偏误类型横向对比，如表 1 所示。为探究不同语言水平学习者偏误类型分布有何差异，我们利用 Excel 对四种偏误类型进行了百分比计算，如表 2 所示。

表 1 各等级学习者介词框式结构正确率

| 学习者等级 | 介词框式结构总数 | 正确使用 | 正确率 (%) |
|-------|----------|------|---------|
| 初级    | 73       | 40   | 54.79   |
| 中级    | 103      | 79   | 76.70   |
| 高级    | 53       | 49   | 92.45   |

表 2 各等级学习者偏误类型比率对比

| 偏误类型 | 初级 (%) | 中级 (%) | 高级 (%) |
|------|--------|--------|--------|
| 遗漏   | 15.07  | 12.62  | 0      |
| 冗余   | 5.48   | 4.85   | 3.77   |
| 错序   | 5.48   | 1.94   | 1.89   |
| 语义偏误 | 19.18  | 3.88   | 1.89   |

由表 1 可知，介词框式结构使用正确率随着学习者等级的提升逐步提高，初级（54.79%）、中级（76.70%）汉语留学生在介词框式结构的准确使用上明显不如高级（92.45%），差异明显。

由表 2 可知，遗漏和语义偏误在初级和中级学习者中占比较高，尤其在初级学习者中，遗漏占比达到 15.07%，语义偏误占比为 19.18%。高级学习者几乎没有遗

漏，语义偏误和错序占比较低仅为1.89%。冗余偏误在高级阶段占比为3.77%，横向对比初级、中级比例，进步不显著，呈现顽固特征。

3.2 不同国家偏误类型分布差异分析

为探究不同母语背景学习者在介词框式结构偏误类型上是否表现出特定的优势或困难，我们通过卡方检验不同国家偏误类型分布差异及多重比较，探讨国别对学习 者偏误类型的影响，以及不同国家学习者在偏误类型上呈现的特征，如表3所示。

表3 不同国家偏误类型占比及卡方检验

| 偏误类型 | 泰国 (%) | 越南 (%) | 印度尼西亚 (%) | $\chi^2$ | $p$    |
|------|--------|--------|-----------|----------|--------|
| 遗漏   | 33.33  | 33.33  | 52.94     | 28.00    | <0.001 |
| 冗余   | 23.80  | 6.66   | 5.88      |          |        |
| 错序   | 9.52   | 26.66  | 5.88      |          |        |
| 语义偏误 | 33.33  | 33.33  | 35.29     |          |        |

由表3可知，对不同国家学习者偏误类型分布进行卡方检验得到 $p<0.001$ ，表明母语背景与偏误类型分布存在统计学关联，我们需进一步进行事后多重比较找出哪两国之间差异显著，如表4。

表4 不同国家差异率卡方检验多重比较

| 对比组         | $\chi^2$ | $p$    |
|-------------|----------|--------|
| 泰国<br>越南    | 68.00    | <0.001 |
| 泰国<br>印度尼西亚 | 201.00   | <0.001 |
| 越南<br>印度尼西亚 | 201.00   | <0.001 |

如表4所示，两两国家之间学习者呈现的偏误类型都具有显著差异 $p<0.0125$ ，泰国学习者遗漏、冗余和语义偏误占比高，越南遗漏、错序和语义偏误比例突出，而印度尼西亚集中于遗漏和语义偏误，其中遗漏偏误比率最高（52.94%）。

3.3 语言水平和国别对介词框式结构的频率、准确率和多样性分析

为探究语言水平和国别对准确性影响的具体差异，我们统计了不同国别和语言水平学习者在介词框式结构使用中的频率、准确性、多样性三项指标，如表5所示。将表5的数据导入SPSS27.0，通过双因素方差分析检验语言水平（初级/中级/高级）和国别（泰/越/印度尼西亚）对介词框式结构使用频率、准确率和多样性的影响，以验证这两个因素和习得效果之间的关联性，得出结果如表6所示。

表5 不同国别不同语言水平的学生介词框式结构的频率、准确率和多样性指标

| 学生ID | 语言水平 | 国别    | 介词框式结构频率 | 准确率(%) | 多样性指标 |
|------|------|-------|----------|--------|-------|
| 1    | 低    | 泰国    | 3        | 33.33  | 1     |
| 2    | 低    | 泰国    | 7        | 14.28  | 4     |
| 3    | 低    | 泰国    | 2        | 100    | 1     |
| 4    | 低    | 泰国    | 3        | 33.33  | 2     |
| 5    | 低    | 泰国    | 7        | 28.57  | 4     |
| 6    | 低    | 泰国    | 7        | 85.71  | 2     |
| 7    | 低    | 越南    | 5        | 80     | 4     |
| 8    | 低    | 越南    | 2        | 50     | 2     |
| 9    | 低    | 越南    | 3        | 66.66  | 2     |
| 10   | 低    | 越南    | 2        | 50     | 1     |
| 11   | 低    | 越南    | 1        | 0      | 1     |
| 12   | 低    | 越南    | 6        | 66.66  | 3     |
| 13   | 低    | 印度尼西亚 | 5        | 60     | 1     |
| 14   | 低    | 印度尼西亚 | 4        | 50     | 4     |
| 15   | 低    | 印度尼西亚 | 3        | 66.66  | 2     |
| 16   | 低    | 印度尼西亚 | 1        | 0      | 1     |
| 17   | 低    | 印度尼西亚 | 1        | 0      | 1     |
| 18   | 低    | 印度尼西亚 | 4        | 75     | 2     |
| 19   | 低    | 印度尼西亚 | 4        | 0      | 1     |
| 20   | 中    | 泰国    | 3        | 100    | 2     |
| 21   | 中    | 泰国    | 6        | 100    | 3     |
| 22   | 中    | 泰国    | 6        | 100    | 2     |
| 23   | 中    | 泰国    | 3        | 100    | 1     |
| 24   | 中    | 泰国    | 2        | 50     | 1     |
| 25   | 中    | 泰国    | 3        | 100    | 1     |
| 26   | 中    | 越南    | 3        | 100    | 1     |
| 27   | 中    | 越南    | 4        | 75     | 2     |
| 28   | 中    | 越南    | 6        | 100    | 3     |
| 29   | 中    | 越南    | 1        | 100    | 1     |
| 30   | 中    | 越南    | 4        | 75     | 2     |
| 31   | 中    | 越南    | 4        | 50     | 3     |
| 32   | 中    | 越南    | 5        | 100    | 2     |
| 33   | 中    | 越南    | 5        | 80     | 3     |
| 34   | 中    | 印度尼西亚 | 5        | 100    | 2     |
| 35   | 中    | 印度尼西亚 | 5        | 40     | 2     |
| 36   | 中    | 印度尼西亚 | 2        | 100    | 1     |
| 37   | 中    | 印度尼西亚 | 5        | 20     | 3     |
| 38   | 中    | 印度尼西亚 | 4        | 100    | 2     |
| 39   | 高    | 泰国    | 6        | 100    | 3     |
| 40   | 高    | 泰国    | 3        | 66.66  | 1     |

续表

| 学生ID | 语言水平 | 国别    | 介词框式结构频率 | 准确率 (%) | 多样性指标 |
|------|------|-------|----------|---------|-------|
| 41   | 高    | 泰国    | 1        | 100     | 1     |
| 42   | 高    | 泰国    | 4        | 100     | 2     |
| 43   | 高    | 泰国    | 3        | 100     | 1     |
| 44   | 高    | 泰国    | 4        | 75      | 3     |
| 45   | 高    | 泰国    | 1        | 100     | 1     |
| 46   | 高    | 越南    | 1        | 100     | 1     |
| 47   | 高    | 越南    | 2        | 50      | 1     |
| 48   | 高    | 印度尼西亚 | 3        | 100     | 2     |
| 49   | 高    | 印度尼西亚 | 1        | 0       | 1     |
| 50   | 高    | 印度尼西亚 | 9        | 100     | 2     |
| 51   | 高    | 印度尼西亚 | 4        | 100     | 4     |
| 52   | 高    | 印度尼西亚 | 3        | 100     | 1     |
| 53   | 高    | 印度尼西亚 | 2        | 100     | 1     |
| 54   | 高    | 印度尼西亚 | 2        | 100     | 2     |

由表5可知，学习者语言水平准确率呈现显著层级差异：低水平组平均准确率45.27%，中级组提升至83.68%，高级组达86.98%。泰国学习者表现突出，中级组准确率（91.7%）甚至超过越南高级组（75%）。越南低水平组准确率（52.2%）高于泰国同级（49.2%）和印度尼西亚同级（36.0%）。印度尼西亚学习者的准确率波动最大，低水平组36.0%到高级组85.7%，增幅明显。

表6 语言水平和国别对学习者的介词框式习得各因素影响F值检验

| 自变量     | 因变量      | 均方       | F值    | 显著性   |
|---------|----------|----------|-------|-------|
| 语言水平    | 介词框式结构频率 | 6.403    | 1.787 | 0.179 |
|         | 准确率 (%)  | 8164.813 | 9.495 | 0.000 |
|         | 多样性指标    | 1.094    | 1.068 | 0.352 |
| 国别      | 介词框式结构频率 | 3.908    | 1.091 | 0.345 |
|         | 准确率 (%)  | 785.915  | 0.914 | 0.408 |
|         | 多样性指标    | 0.071    | 0.069 | 0.933 |
| 语言水平*国别 | 介词框式结构频率 | 3.239    | 0.904 | 0.470 |
|         | 准确率 (%)  | 240.421  | 0.280 | 0.890 |
|         | 多样性指标    | 0.790    | 0.771 | 0.550 |

分析表6数据可知，语言水平对学习者的介词框式结构准确率的影响非常显著（ $p=0.000$ ， $p<0.05$ ），而国别对学习者的介词框式结构的准确率的影响上，主效应不显著（ $p=0.408$ ， $p>0.05$ ）。在对介词框式结构频率进行分析时，语言水平（ $p=0.179$ ， $p>0.05$ ）、国别（ $p=0.345$ ， $p>0.05$ ）以及语言水平与国别的交互作用（ $p=0.470$ ， $p>0.05$ ）均未达到显著水平，这意味着语言水平、国别以及它们之间的交互作用对介词框式结构的频率没有显著影响。

语言水平栏 $p=0.352$  ( $p>0.05$ ), 说明语言水平对多样性指标没有显著影响。国别栏 $p=0.933$  ( $p>0.05$ ), 说明国别对多样性指标也没有显著影响。

综上, 只有语言水平对学习者的介词框式结构准确率有显著影响, 而国别对准确率影响不显著, 语言水平与国别的交互作用对频率和多样性两个指标影响不显著。

## 4 研究结果

本研究从习得准确率、偏误类型分布、结构使用频率和多样性四个角度对泰国、越南、印度尼西亚不同水平汉语学习者介词框式结构习得特征进行系统考察, 并运用方差分析和卡方检验对不同国别、不同语言水平学习者在各维度上的差异性进行分析, 初步得出以下结论。

### 4.1 语言水平显著影响介词框式结构习得准确性

通过对初级、中级和高级汉语学习者介词框式结构使用情况的对比分析, 我们发现三个语言水平阶段呈现出明显的偏误类型差异。初级学习者(正确率54.79%)存在较多遗漏和语义混淆问题, 这表明他们虽然能掌握基本结构, 但对复杂语境下的介词功能理解不足。中级学习者(正确率76.70%)的准确性显著提升, 但仍会出现结构错序和特定语义偏误, 显示出他们在结构运用上还不够熟练。高级学习者(正确率92.45%)的表现接近母语者水平, 仅存在少量冗余, 如过度使用正式表达“就……而言”, 以及来自学习者母语的形态迁移, 如印度尼西亚学生的“从早直到晚”, 体现了他们对介词框式结构的精准把握和语境适应能力。整体来看, 介词框式结构的习得呈现明显的层级性特征, 初级阶段以形式掌握为主, 中级阶段发展语义区分能力, 高级阶段达到了语用层面。

### 4.2 母语背景对偏误类型的差异化影响

通过卡方检验我们得出学习者母语背景显著影响学习者介词框式结构的偏误类型分布。泰国学习者相比越南(6.66%)和印度尼西亚(5.88%)表现出明显的冗余偏误特征(23.80%), 越南学习者错序偏误比例最高(26.66%), 而印度尼西亚学习者遗漏偏误最为突出(52.94%)。两两比较显示三国学习者的偏误模式均存在显著差异( $p<0.0125$ ), 其中泰国和越南在冗余偏误、越南和印度尼西亚在错序偏误、泰国和印度尼西亚在遗漏偏误上的差异尤为显著。这一结果印证了母语迁移效应的国别化特征, 即泰语介词系统具有冗余特征、越南语SOV语序干扰学习者对汉语SVO语言的习得导致错序比例(26.66%)高于泰(9.52%)、印(5.88%)两国学习者, 而印尼语介词简化缺少后置词导致成分省略倾向。这些母语特点塑造了三国学习者独特的偏误模式。研究结果为针对不同母语背景学习者的差异化教学提供了实证依据。

### 4.3 习得进步体现于质量而非使用频率或多样性

本研究还通过双因素方差分析考察了语言水平和国别对介词框式结构使用正确率、频率及多样性的影响，语言水平对介词框式结构习得准确率具有决定性影响( $F=9.495, p<0.001$ )，学习者呈现显著的进步趋势，低、中、高级组准确率分别为45.27%、83.68%、86.98%。值得注意的是，语言水平和国别对介词使用频率( $p=0.470$ )及多样性均无显著影响，说明习得进步主要体现在质量(准确性)而非数量(使用频次和类型)上。

## 5 介词框式结构的习得策略建议

东南亚汉语学习者在介词框式结构的习得上呈现出显著的语言水平差异和国别化偏误模式，这些结果不仅揭示了学习者习得过程中的优势和难点，也为针对性教学策略的制定提供了实证依据。下面我们将结合研究结果，从习得策略和教学启示两个层面提出具体建议，以优化介词框式结构的教学实践。

### 5.1 介词框式结构的习得策略建议

东南亚高级汉语学习者在介词框式结构习得准确率方面达到了86.98%，表现出明显的层级优势，不同国别的学习者也在习得过程中表现出特定的优势。通过深入探讨学习者在介词框式结构习得过程中展现出的典型特征和有效策略，我们总结了以下三点习得策略。

#### 5.1.1 通过语境理解掌握介词用法

高级学习者在习得介词框式结构时，倾向于通过语境来理解和掌握介词的用法。许多学习者在实际交流中，通过对话、阅读材料以及写作中的实例，逐步积累了对介词的感性认识。根据HSK作文语料库的分析，高级学习者在“在……上”“对……来说”等框式结构的使用中，能够根据上下文准确选择介词，并灵活调整其用法。例如，在表达空间关系时，他们能够正确使用“在……里”或“在……上”，而在表达时间关系时，则能够准确使用“在……时候”。这种通过语境反复接触的学习策略，使学习者能够更深入地理解介词的功能，并将其应用于实际表达中。

#### 5.1.2 通过母语对比促进介词框式结构习得

东南亚地区的高级汉语学习者，在学习介词框式结构时，会通过对比母语(泰语、越南语、印尼语)与汉语之间的介词差异来促进学习。例如，泰语和越南语属于介词前置类型语言，与汉语的介词类型较为接近，但具体用法存在差异；而印尼语的部分介词语序则与汉语有较大不同。根据HSK作文语料库的数据，高级学习

者能够逐渐避免母语干扰,并在对比中掌握汉语介词的准确用法。例如,泰语学习者在习得“对……来说”时,能够意识到泰语中类似结构的差异,并调整自己的语言输出。这种对比学习策略不仅帮助学习者克服语义范畴的概念迁移,还促进了他们对汉语介词系统的深入理解。

### 5.1.3 针对国别差异推广习得策略

研究结果显示,不同国别学习者在介词框式结构的使用中表现出特定的优势。泰国学习者凭借泰语同为介词前置类型语言,在初中级阶段表现突出(中级准确率91.7%),证明其形式规则内化能力突出,泰国学习者在学习初期即注重介词结构的完整输出,注重句式的概念意识。这种学习方法使得泰国学习者遗漏偏误随着语言水平提高大大降低,证明了这种方法能有效避免缺词少词问题的出现,也给越南学习者一定的学习启示。越南学习者虽然错序偏误较多,但其低水平组平均准确率(52.2%)高于同级泰、印两国,这表明越南学习者是从语义理解入手掌握介词框式结构,而非机械形式记忆,随着语言水平的提高,学习者主动比较汉越语序差异,建立显性认知对照,能有效降低由母语带来的句法负迁移,注重语义也能更深入地理解第二语言,不易遗忘。印度尼西亚学习者在习得中最明显的优势是准确率增幅最大,初期准确率最低(36.0%),但高级阶段进步显著(85.7%),展现出强元语言意识。这表明他们更擅长系统学习汉语介词规则并积极应用于输出,由此可以推广的策略是学习者可通过强化输出练习来弥补母语简化特征,从而达到语言水平的提高。

总结以上三个不同国家学生的学习策略,我们归纳出初级阶段要注重结构化输入优先,随着语言水平提高由易到难渐进式梯度学习,以及加强学习元认知监控这些策略,这些学习方法均源自学习者在实际习得过程中表现出的有效行为模式,具有推广价值。

## 5.2 介词框式结构的教学启示

从学习者习得结果得到启发,教师在实际教学中可以根据学习者的国别背景和语言水平设计有针对性的教学策略,帮助学生克服特定的语言偏误。本文从以下三方面给出了介词框式结构的教学启示。

### 5.2.1 针对不同国别背景设计教学内容

泰国学习者在初中级阶段表现突出,但进展缓慢,冗余偏误顽固。教师可以对泰国学生开展“近义介词删减游戏”,要求其在写作中标注所有介词并删除冗余项,强化学生记忆。越南学习者受越南语介词前置类型语言影响,易出现框架残缺,遗漏前置词或后置词。教师可采取针对性策略如“显性框架”教学,使用双色

标记法，例如前置词用红色标记，后置词用蓝色标记，配合“结构补全”练习强化结构意识，总结汉越语序对比表，纠正越南留学生的错序问题。对印度尼西亚学生可采用“方位词强制完形法”，通过填空练习和视觉提示，例如用红色标记缺失成分，解决其遗漏偏误。教师在实施策略的同时配合即时反馈机制，以强化学生正确形式的输出。

### 5.2.2 针对语言水平分阶段实施教学法和教学策略

Talmy (2003) 的认知语义学理论认为，语言中的空间表达反映人类对物理空间的认知图式，方位词本质上是空间关系的心理表征，不同语言通过特定形式编码这些图式，汉语强制使用“在X上”，而印尼语可省略方位标记。教师可以根据此认知特点，对不同语言水平的学生采取分阶段教学法。初级阶段采用“语义图示法”，通过实物具象化讲解方位介词；中级阶段设计“母语对比任务”，如要求泰国学习者将母语双重介词结构转换为汉语单介词表达；高级阶段宜采用“语篇完形填空”，在文本中嵌入近义介词选项，帮助越南学习者在复杂语境中降低语义偏误率。

### 5.2.3 借助技术数据实现智慧学习

教师可基于 Ollama 模型开发智能纠错插件，实时标记学习者的介词冗余和语序偏误语句，该技术结合人工智能对偏误的识别有较高准确率，可用于辅助教学。教师构建语料库学习模块，指导学生用 AntConc 检索 HSK 语料库中的结构搭配，如学术语篇用法和日常对话用语的不同，强化学生记忆。如今随着时代发展，技术在教学中的应用越来越普遍，为教学带来了个性化学习路径、即时精准反馈、大数据驱动优化、认知可视化强化等优势。

## 6 结论

本研究基于 HSK 动态作文语料库，深入考察了东南亚汉语学习者介词框式结构的习得特征，揭示了几个关键发现。高级汉语学习者在介词框式结构使用中的偏误显著减少，尤其是在遗漏方面几乎不再出现错误，与初级和中级学习者相比，高级学习者表现出更高的准确性和稳定性。研究还发现，随着语言水平提升，高级汉语学习者随着水平提高语义偏误显著减少，但冗余偏误仍具顽固性。学习者的母语背景显著影响其偏误类型分布，泰国学习者冗余偏误突出，越南学习者错序偏误明显，印度尼西亚学习者则以遗漏偏误为主。根据学习者在习得过程中表现出来的特定优势和困难，我们总结了三条系统性学习策略：通过真实语境中的反复接触来内化介词用法、借助母语与汉语的对比分析克服迁移干扰、针对不同母语背景采取差异化学习路径。基于这些发现，我们提出了一套针对国别的教学方案，对泰国学生开展冗余修正训练，为越南学习者设计语序对比练习，帮助印度尼西亚学生强化结

构完整性意识。这些发现不仅深化了对汉语二语习得规律的认识，也为实际教学提供了具体可行的改进方向。

## 参考文献

- Greenberg J H, 2005. Language universals: with special reference to feature hierarchies[M]. Berlin: Mouton de Gruyter.
- Talmy L, 2003. Toward a cognitive semantics[M]. Cambridge: MIT Press.
- 陈昌来, 2002. 介词与介引功能[M]. 合肥: 安徽教育出版社.
- 崔希亮, 2005. 欧美学生汉语介词习得的特点及偏误分析[J]. 世界汉语教学 (3): 83-95+115-116.
- 黄理秋, 施春宏, 2010. 汉语中介语介词性框式结构的偏误分析[J]. 华文教学与研究 (3): 33-41.
- 贾君芳, 何洪峰, 2020. 后置介词结构语序演变的类型学观照[J]. 中南大学学报 (社会科学版) (2): 200-208.
- 刘丹青, 2002. 汉语中的框式介词[J]. 当代语言学 (4): 241-253+316.
- 鲁健骥, 1994. 外国人学汉语的语法偏误分析[J]. 语言教学与研究 (1): 49-64.
- 周文华, 2013. 韩国学生不同句法位“在+处所”短语习得考察[J]. 华文教学与研究 (4): 35-43.
- 周文华, 肖奚强, 2011. 现代汉语介词习得研究[J]. 语言文字应用 (2): 144.

## 通信地址

130117 长春市 吉林外国语大学国际传媒学院

(责任编辑: 刘若冰)

# 基于文本挖掘的国家生态发展观探究： 以“十四五”规划英译文本为例\*

广东海洋大学 黎曜玮 张 朦

**提 要：**本研究以“多元和谐，交互共生”生态哲学观作为判断标准，通过社会语义网络模型、词云与词文本位置分布、多维空间语义计算和情感计算进行生态性文本挖掘与话语分析，探究国家政策文件“十四五”规划英译文本中体现的国家生态文明建设的发展观念。研究发现：(1) 生态主题是“十四五”规划中的重要一环，生态发展并非独立环节，与经济、社会发展等主题相互关联、有机融合；(2) 国家在生态文明建设中作为参与主体，表现出积极的态度和努力，重视全面的环境保护和治理，致力于体系化和机制化建设，并积极推动国际合作；(3) 文本体现的生态情感以中性与积极情感为主，呈现出“理性客观为主、积极引导为辅”的特点，既突出国家生态文明建设的决心，也体现国家政策落实的稳健严谨与可操作性。本研究旨在对未来探究国家政策文件生态发展观及其国际话语传播影响的研究起到一定的方法论参考。

**关键词：**生态文明建设；国家发展观；文本挖掘与话语分析

## Research on National Ecological Development Outlook Through Text Mining: A Case Study of the English Version of the 14th Five-Year Plan

Yaowei Li and Meng Zhang

**Abstract:** Guided by the ecological philosophy of “diversity in harmony and interactive

\* 本文系2023年广东海洋大学人文社会研究项目“深度神经网络机器文学翻译系统研发及平行语料库构建”（项目编号：030301162306）、教育部人文社会科学研究基金项目“中国海洋治理对外话语体系的汉英译介与传播研究”（项目编号：24YJC740088）、江门市哲学社会科学规划课题“江门地方志在英语国家的译介与传播研究”（项目编号：JM2025C120）的阶段性研究成果。

张朦为本文通信作者。

作者贡献：

黎曜玮：选题构思、研究方法、数据分析、讨论结论、论文撰写、修改润色、字数占比（70%）；

张朦：选题构思、研究方法、修改润色、字数占比（30%）。

coexistence” as the evaluative standard, this study employs social semantic network modeling, word cloud visualization with text-location distribution, multidimensional spatial semantic computation, and sentiment analysis to conduct ecological text mining and discourse analysis on the English translation of China’s “14th Five-Year Plan” policy document, examining its articulation of national ecological civilization development concepts. Findings indicate that: (1) ecological themes form a crucial, non-isolated component of the Plan, organically integrated with economic and social development domains; (2) as the primary actor in ecological advancement, the state demonstrates proactive engagement through comprehensive environmental governance, systematic institutional development, and international cooperation promotion; and (3) the text exhibits predominantly neutral-to-positive ecological sentiment, balancing rational objectivity with constructive guidance—reflecting both resolution in ecological progress and pragmatic rigor in policy implementation. This research provides methodological reference for future studies analyzing ecological perspectives in policy documents and their internationally oriented discourse dissemination.

**Keywords:** ecological civilization construction; national development philosophy; text mining and discourse analysis

## 1 引言

生态发展观（又称生态文明观）是国家在经济发展与生态环境冲突过程中形成的一种全新的发展观念，涉及经济社会发展全局以及人类社会发展的总体性方向，体现“人与自然和谐共生”的生态自然观（姚修杰，2020），“绿水青山就是金山银山”的论断展现以习近平同志为核心的党中央引领国家推进绿色发展的坚强意志和坚定信心（杨卫军等，2020）。2007年党的十七大首次提出建设生态文明，2012年党的十八大明确了生态文明建设的总体要求，并将生态文明建设纳入中国特色社会主义“五位一体”总体布局。党的十九大对我国社会主义现代化建设作出新的战略部署，明确以“五位一体”总体布局推进中国特色社会主义事业，并从经济、政治、文化、社会、生态文明五个方面制定了统筹推进“五位一体”总体布局的战略目标。

在此时代背景下，黄国文（2017）针对中国语境下的政治、经济、文化等话语提出“和谐”之生态哲学观，旨在促进中国语境下人与人、人与自然之间的和谐。何伟等（2018）针对表征国际社会生态系统提出“多元和谐、交互共生”生态哲学观，强调生态系统内多样要素通过平等互动实现动态平衡与共同繁荣，旨在引导人

类摒弃自我中心主义，构建自然与社会生态良性循环的可持续发展模式。生态语言学的出现与兴起，既回应了生态问题的现实要求，又体现了语言学经世致用的价值和语言学者的责任与担当（孙炬等，2022）。

基于“多元和谐、交互共生”生态哲学观及生态话语分析模式（何伟、魏榕，2017；何伟等，2020），国内学者开展环保公益广告、网络评论、新闻报道、企业话语、景观标示语与古典典籍等的生态话语分析（如何伟、耿芳，2018；邵珊珊等，2021；董典，2021；赵蕊华，2022；何冰艳，2023），解读普遍性话语传播中体现的群体生态意蕴，探究其中的生态主题（如环境保护、可持续发展、生物多样性等），以及与这些主题相关的语言表达和符号（如生态词汇、隐喻、象征等），从而揭示背后的生态意义和观念对现实生态问题的影响与启示。何伟等（2023；2024）从国家层面的角度切入，分别基于网络语料分析中国对外话语传播的国家生态形象塑造，以及探究美国能源新政的国家生态观。Ma等（2022）指出中国生态语言学家应把“生态文明”政策作为生态哲学观来分析中国语境中的话语，国家政策话语对国内生态话语构建具有重大意义。目前，鲜有研究对国家政策文件进行生态性话语分析来挖掘国家生态层面发展的意识形态、情感及落实的途径措施。本研究拟以“多元和谐、交互共生”生态哲学思想为标尺，对《中华人民共和国国民经济和社会发展第十四个五年规划和2035年远景目标纲要》（简称“十四五”规划）英译整体文本及其生态文明建设部分（第11章），进行文本挖掘与话语分析，探究中国在新时期的英译政策文件国际话语传播中所蕴含并传达给世界的生态发展观。

## 2 国家环境政策文本挖掘文献回顾

生态话语分析应充分利用语料库工具与计算机技术，尝试更复杂的考察视角与统计方法，最大程度地挖掘话语的词汇语法、语义及语篇信息（张琳等，2024）。相关研究对国家环境政策文件进行了文本深度计量分析，Peng等（2016）使用T-LAB 9.1进行语言学和统计分析，采用量化内容分析法（文本挖掘、词频统计、共现分析、主题聚类），系统研究1997—2013年中国政府发布的清洁生产政策文本，以揭示政策关键词关联、核心主题、制定与实施特征（如协作性、约束性、引导性）。斯丽娟等（2018）利用政策文本量化分析（内容分析、工具分类统计）和耦合协调模型实证方法，研究1982—2018年20份中央一号文件中农村绿色发展政策的演进规律、工具特征及其实践绩效（以经济发展与生态协调为评价核心），检验绿色发展政策在协调区域经济发展与生态环境保护方面的实际效果。Liao（2018）采用内容分析法，从政策工具类型（命令控制型、市场型、信息型）和环境创新阶段（研发阶段、采用阶段）两个维度，对中国1987—2016年发布的231项促进企业环境创新的政策文本进行系统性挖掘。冉连（2020）聚焦“绿色治理政策”相关内

容,使用Nvivo软件对1949—2020年政府发布的52份《政府工作报告》进行文本挖掘,对政策关键词进行计量统计、分类(如词频分析、节点编码),梳理政策在价值层面、时间层面、内容层面的变迁逻辑。贺丹等(2021)采用文本内容分析法,包含政策力度量化、主体协同度计算、Nvivo关键词分析及政策工具编码,针对1986—2021年中国绿色服务政策体系(以《绿色产业指导目录》为依据的859份中央政策文本),研究政府绿色服务政策的演进轨迹、部门协同特征及政策工具结构。这些研究多从政策的基本信息(发文时间、文件类型)、政策主体、政策主题、政策工具、政策目标等维度展开,再辅以协同度模型或社会网络分析等方法进行深入挖掘。

然而,上述文献仅对以往的环境政策文件进行历时研究,鲜有涉及对国家未来长期生态文明建设规划的发展观探究。本研究拟基于“多元和谐,交互共生”生态哲学观作为判断标准,聚焦形容词表达及其所在语境模式,通过生态视角下的文本挖掘与话语分析模式,探究“十四五”规划英译整体文本及其生态文明建设部分(第11章)中体现的国家生态文明建设发展观。“十四五”规划是我国全面建设社会主义现代化国家、向第二个百年奋斗目标进军的第一个五年规划,体现了当代中国的历史使命和基本定位。以其语料为分析对象,使本研究得以在语料的权威性、规范性与准确性上科学地开展对国家生态发展观的分析。同时,本研究选择“十四五”规划中华人民共和国外交部英译本,旨在关注国家长期发展规划文件在生态方面对外的国际传播效果,探索中国生态话语在国际上发挥的影响作用。生态翻译学认为译者与翻译生态环境的关系问题,是译者与语言、交际、文化、社会、作者、读者与委托者等之间的互动关系(胡庚申,2011),从翻译文本中窥见译者对源语言社会环境的认识,其翻译质量及对外传播的效果,对国家政治文化形象构建具有重要的作用。因此,选择“十四五”规划的英译本,能够以国际化的视角,客观全面地分析国家的生态情感倾向及在生态问题上发挥的国际角色作用。

### 3 研究方法

本研究使用Python第三方库SpaCy等自然语言处理技术对目标语料(“十四五”规划英译整体文本)进行数据预处理,去除文本中噪声数据,并提取正文的干净文本。文本总词数为45,201(生态部分文本总词数为2,722),标识符为3,767,文本词汇丰富度为11.99(每个词平均被使用12次)。本研究的文本挖掘步骤如下。

(1) 采用VOSviewer对整体英译文本进行可视化社会语义网络分析与生态主题挖掘。

(2) 使用SpaCy的分词与词性赋码检索生态部分文本中的形容词,统计词频筛选出高频关键词。

(3) 使用BertTokenizer与BertEmbedding对“十四五”规划文本的生态部分进

行高维空间向量建模，并将模型数据导入Embedding Projector，对关键词三维空间语义分布可视化，获取空间距离相近的语域信息。

(4) 利用 Python 第三方库 SpaCy、VaderSentiment 对“十四五”规划英译文本生态部分的句子和形容词进行情感量化分析,统计总体句子与形容词不同情感的表现,挖掘其体现的生态情感倾向。

(5) 运用正则表达式 ([A-Z][^!~?]\*\b target word \b[^!~?]\*[!~?]) 检索关键高频形容词所在语境, 结合质性研究范式挖掘文本体现的生态发展观。

## 4 文本挖掘与数据分析

#### 4.1 “十四五”规划英译文本社会语义网络

本研究使用 VOSviewer 对“十四五”规划整体英译文本进行社会语义网络构建与挖掘分析。VOSviewer 是基于网络科学的信息可视化工具，主要用于构建文本语义结构及知识图谱，探究语义关系类型与关联强度。在文本预处理上，本研究使用 Python 语言对目标文本进行词语去重、停用词去除与词干提取，将预处理后的文本导入软件中，使用聚类算法（如 K-Means、DBSCAN、层次聚类等）、共现分析（如词语共现分析等）、引用分析（如引用网络分析、引用关系图谱等）与主题建模（如 Latent Dirichlet Allocation、Topic Modeling 等）构建模型并观测建模效果。通过 VOSviewer 的图谱模型（如力导向布局、网络映射、热度图等）、矩阵与散点图等进行数据可视化，根据研究问题、数据特点与可视化结果的可解释性，本研究选取图 1 和图 2 为“十四五”规划英译文本的社会语义网络图谱。

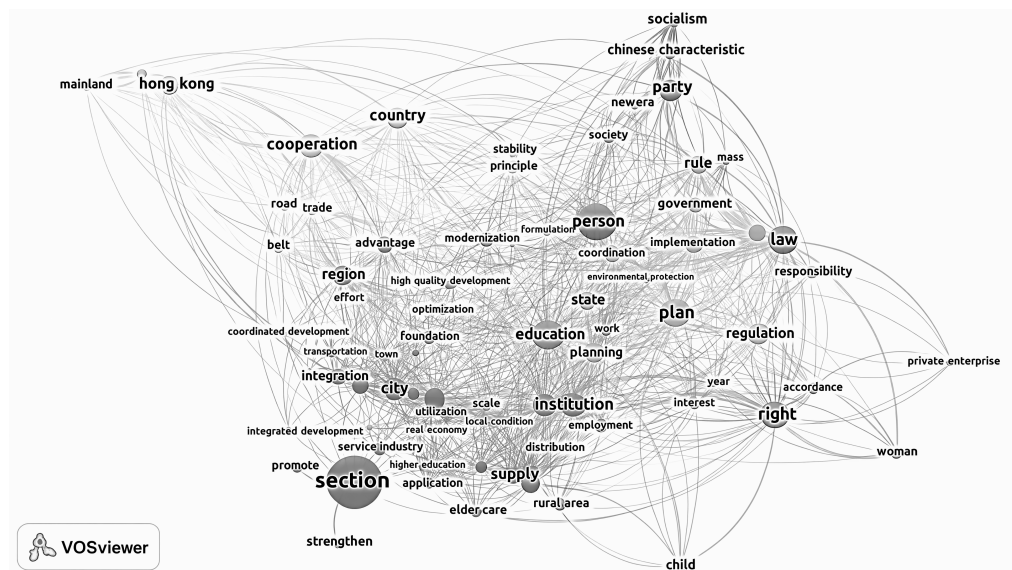


图1 “十四五”规划整体英译文本社会网络

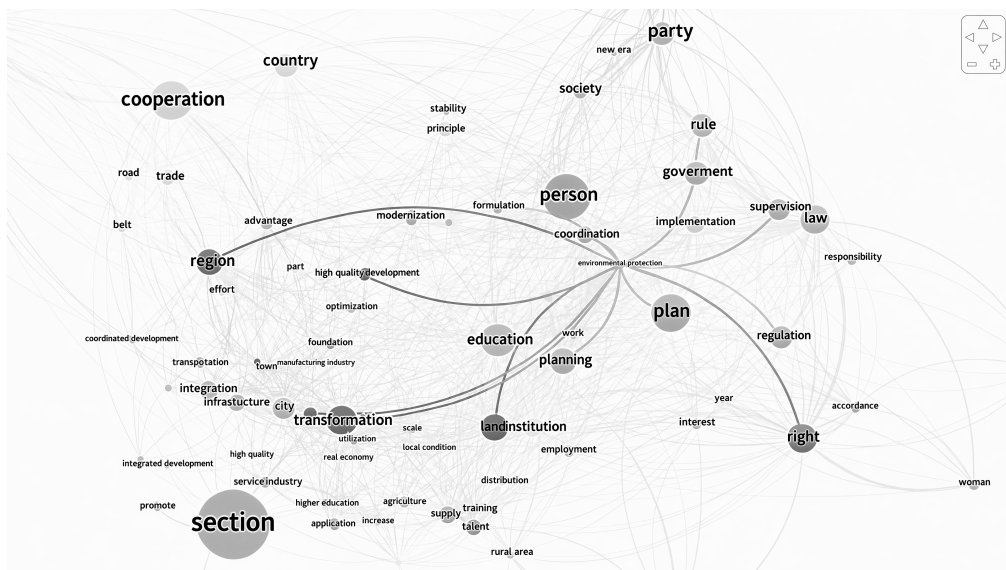


图2 “十四五”规划整体英译文本“environmental protection”社会网络

图1为VOSviewer以共现频次阈值大于10的85个高频关键词绘制的共现聚类图谱,共分为五个结构主题,85个关键词产生2042个网络联接,整体关联强度为11,334。图谱中圆形节点表示关键词,其大小表示关键词的出现频次,节点连线粗细表示共现强弱程度。基于表1呈现的关键词主题聚类,依据第一个主题结构中的transformation、institution、land、region、infrastructure、supply、agriculture、service industry、application、foundation、energy等主题词,主题1可概括为现代化建设。依据第二个主题结构中的plan、planning、regulation、supervision、implementation、rule、government、state、coordination、responsibility、optimization、formulation、environmental protection等主题词,主题2可概括为生态监管。依据第三个主题结构中的education、law、right、employment、interest、accordance、rural area、woman、child、private enterprise、year、scope等主题词,主题3可概括为社会管理。依据第四个主题结构中的cooperation、country、Hong Kong、belt、road、trade、Macao、part、effort、principle、Mainland等主题词,主题4可概括为外交合作。依据第五个主题结构中的person、party、modernization、work、society、Chinese characteristic、socialism、new era、direction、mass、stability等主题词,主题5可概括为政治建设,见表1。

表1 “十四五”规划整体英译文本主题聚类及关键词共现

| 结构主题<br>(85)   | 关键词及共现频次                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                           |
|----------------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| 现代化建<br>设 (38) | section (194)、city (55)、transformation (54)、institution (54)、land (53)、region (44)、<br>infrastructure (42)、supply (42)、integration (40)、talent (36)、advantage (35)、agriculture<br>(26)、service industry (26)、application (24)、foundation (24)、energy (23)、traning (23)、<br>utilization (22)、high quality development (21)、transportation (19)、town (18)、scale (18)、<br>promote (17)、urban agglomeration (16)、upgrading (15)、strengthen (15)、elder care (14)、<br>increase (13)、manufacturing industry (13)、integrated development (12)、distribution (12)、<br>coordinated development (12)、location condition (11)、high quality (11)、higher education<br>(11)、proposition (11)、real economy (11)、western region (10) |
| 生态监管<br>(13)   | plan (73)、planning (53)、regulation (52)、supervision (44)、implementation (43)、rule<br>(38)、government (34)、state (31)、coordination (26)、responsibility (24)、optimization<br>(15)、formulation (13)、environmental protection (12)                                                                                                                                                                                                                                                                                                                                                                                                                                                                                     |
| 社会管理<br>(12)   | education (80)、law (70)、right (62)、employment (26)、interest (24)、accordance (20)、<br>rural area (19)、woman (17)、child (15)、private enterprise (14)、year (13)、scope (11)                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                            |
| 外交合作<br>(11)   | cooperation (68)、country (47)、Hong Kong (29)、belt (27)、road (26)、trade (25)、<br>Macao (22)、part (21)、effort (19)、principle (18)、Mainland (10)                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                      |
| 政治建设<br>(11)   | person (106)、party (37)、modernization (32)、work (28)、society (27)、Chinese<br>characteristic (17)、socialism (16)、new era (15)、direction (14)、mass (12)、stability (17)                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                               |

图1右部中间区域为生态发展规划的社会语义网络，本研究以“生态监管”主题下的environmental protection为中心词，探究其产生的共现网络知识图谱（见图2），发现government、planning、supervision、regulation、coordination、transformation、region与high quality development和中心词高度共现，其中环境保护（environmental protection）与规划（plan）关联强度较大，在共现网络中较为突出，表明生态主题是“十四五”规划中的重要一环，并与其他主题的关键词相互关联，如转型（transformation）与高质发展（high-quality development），说明生态发展并非独立环节，而是与其他主题发展有机融合，突出政府在生态发展规划中充当行为施事者，在环境保护、生态保护和生态修复方面，政府不断加强监管、立法与治理的力度，制定和完善生态文明与区域协调发展政策，明确发展规划、转型方向和推动机制，坚持经济发展与生态文明建设、区域发展与协调发展相统一的发展观，以实现高质量的可持续发展。

4.2 “十四五”规划英译文本生态部分词语挖掘

本小节从词语层面挖掘语料的生态性。图3为“十四五”规划英译文本生态部分的高频词云图，其中出现频率较高的关键词如promote、strengthen、improve、control、implement等动词表现出国家作为环境管理与治理的主体参与者的行为动

作，突出其在生态治理中的积极态度和努力，强调对生态环境的保护和治理政策的推动，管控污染等生态性破坏行为，进一步优化、加强措施的实行。



图3 “十四五”规划英译文本生态部分词云

近年来，政府采取了一系列措施来推动生态环境保护和治理，如加强环境监管、推广绿色技术、强化环境执法等，并积极加强国际合作，共同应对气候变化和环境问题，推动全球生态文明建设。water、energy、resource、river basin、national park、farmland、climate change 等实义名词突出了国家对环境保护的全面性和整体性的重视，如 water、energy、resource 突出对资源的保护和管理全面性，资源是国家发展的重要基础，对于国家经济和社会的发展至关重要。国家致力于保护和管理这些资源，确保其可持续利用，从而实现经济、社会和生态的协同发展。river basin、national park 突出对生态环境的保护和管理全面性，我国拥有广阔的地域和多样的生态系统，国家致力于确保这些生态系统的可持续利用和保护，特别是通过实施流域和自然保护区的管理，来保护和修复生态系统。farmland 突出对农业和土地资源的保护和管理全面性，农业是中国经济的重要组成部分，也是土地资源利用的重要方式，国家致力于保护和管理农业和土地资源，确保其可持续利用，从而保障国家粮食安全和农民的生计。climate change 突出中国政府对气候变化问题的处理与在国际环境保护中的角色担当，气候变化已成为全球性的挑战，国家致力于通过积极的政策和行动，来减缓气候变化和应对其影响，推广低碳技术和能源转型，减少碳排放和促进绿色发展。system、mechanism 等概念名词强调国家在建设生态文明方面致力于做到体系化与机制化，采取一系列措施来保护环境、促进可持续发展。近年来，国家制定了一系列法律法规来保护环境，如《环境保护法》《大气污染防治法》《水污染防治法》等，并建立起一系列机构来协调环境保护工作，如国家环境保护总局、国家海洋局、国家林业和草原局等。此外，中国政府还积极推动生态文明建设的国际合作，与其他国家共同应对气候变化和环境问题，例如参与《巴黎协定》的制定，并主办世界环境日等活动。总体而言，国家在生态文明建设

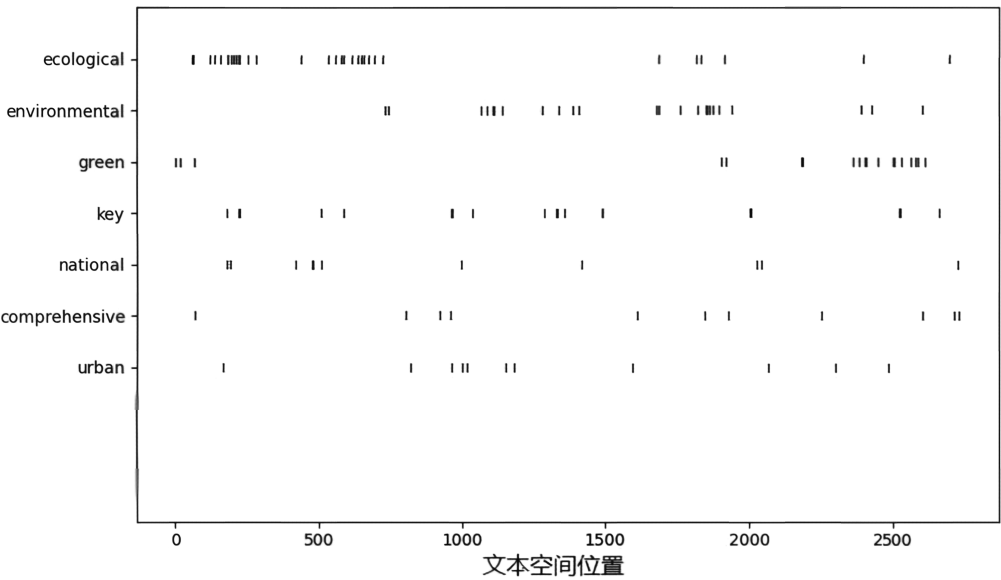
设中作为参与主体，表现出积极的态度和努力，重视全面的环境保护和治理，致力于体系化和机制化建设，并积极推动国际合作。

4.3 生态评价性形容词的词频与文本分布

形容词是评价意义的典型体现方式（Martin et al., 2005），早期研究也多选择形容词进行网络社交媒体评论的民意调查，被学者认为是较好地体现出情感分析的重要特征（Hunston et al., 2019；Su et al., 2019）。本研究以“十四五”规划英译文本生态部分的形容词及其所在的生态语境作为研究对象，按词频大小筛选7个高频形容词（词频≥10，见表2），并使用NLTK工具包生成形容词文本位置分布图（见图4）。

表2 生态性形容词词频

| 序号 | 形容词           | 词频 |
|----|---------------|----|
| 1  | ecological    | 32 |
| 2  | environmental | 24 |
| 3  | green         | 19 |
| 4  | key           | 19 |
| 5  | national      | 12 |
| 6  | comprehensive | 11 |
| 7  | urban         | 11 |



由表2可见, ecological与 environmental在文本内出现的频率较高,但其文本位置分布不同,前者集中在文本前部,后者则在文本中后部。基于文本内容分析, ecological多描述自然景观和生态环境(如森林、河流、湖泊、草原等自然生态系统),对维持生命和人类福祉非常重要, environmental则描述人类居住的生活环境,例如城市、农村、家庭、工作场所等,对人们的生活质量、身心健康产生着深远的影响。自然环境和人类生活环境存在紧密的相互作用,人类活动对自然环境的破坏和污染会导致生态系统的崩溃和生物多样性的丧失,生态系统的破坏和衰退也会对人类生活环境产生负面影响。因此,国家把生态保护放在首要位置,表明国家对可持续发展的重视,优先保护生态环境,实现经济、社会和环境的协调发展。green一词集中分布于后半部分,主要源于政策文本的绿色发展理念定位,前半部分侧重宏观战略与基础性目标(如经济发展、科技创新),而后半部分聚焦具体实施路径与保障措施。绿色发展作为实现高质量发展的关键抓手,被系统性地嵌入生态保护、能源转型、产业升级等实践层面,该词高频代指“绿色发展”“绿色经济”等复合概念,在后半部分涉及生态环保、国际合作等内容时呈现密集分布,凸显绿色理念从顶层设计到落地执行的政策闭环逻辑。key、national、comprehensive、urban四个形容词在文本中呈现出同步且均匀地分布特征,反映了中国生态治理的全局性、系统性和多层次性。战略定位的“关键性”(key)表明生态议题被明确列为国家发展的核心任务之一,例如,规划提出“加强生态环境保护法治建设”需聚焦“关键领域”(key areas),并将“碳中和”目标作为“关键路径”(key pathway)。这些表述凸显了生态治理在整体战略中的优先性,体现了key对重点任务的强调。治理层级的“国家性”(national)既体现在“国家发展规划”(national development plan)等宏观框架中,也用于描述“国家公园体制试点”(national park system)等具体制度,这种用法强化了生态治理的国家主导性,呼应了“统筹中华民族伟大复兴战略全局”的全局观。政策措施的“综合性”(comprehensive)强调多维度协同,构建全过程、全方位生态环保体系,突显了跨领域、全链条的系统治理逻辑。空间维度的“城市性”(urban)体现在城市化与生态保护的平衡上,反映了城市化进程中生态修复的紧迫性。

#### 4.4 生态性语义三维空间向量分析

本研究通过自然语言处理的词嵌入(Word Embedding)技术对“十四五”规划英译文本生态部分进行词向量(Word Vector)模型的构建。词嵌入将文本在不同颗粒度上(词、N-gram、句或段落等)映射为高维空间实数向量,借此表示其语义和语法信息。Word2Vec、GloVe和FastText是词嵌入常用的算法模型,本研究使用基于Transformer神经网络架构的BERT模型对文本进行向量化,BERT使用滑动式上

下文窗口动态捕捉词语的位置信息，可以准确区分不同语境中的多义词汇。在 Python 环境中，实验导入 transformers 第三方库的 BertTokenizer 与 BertModel 模块，加载英文分词与词嵌入预训练模型，对目标文本进行高维向量化，并将生成的 tsv 与 tsv-metadata 数据上传至 Embedding Projector，实现三维空间向量语义可视化。

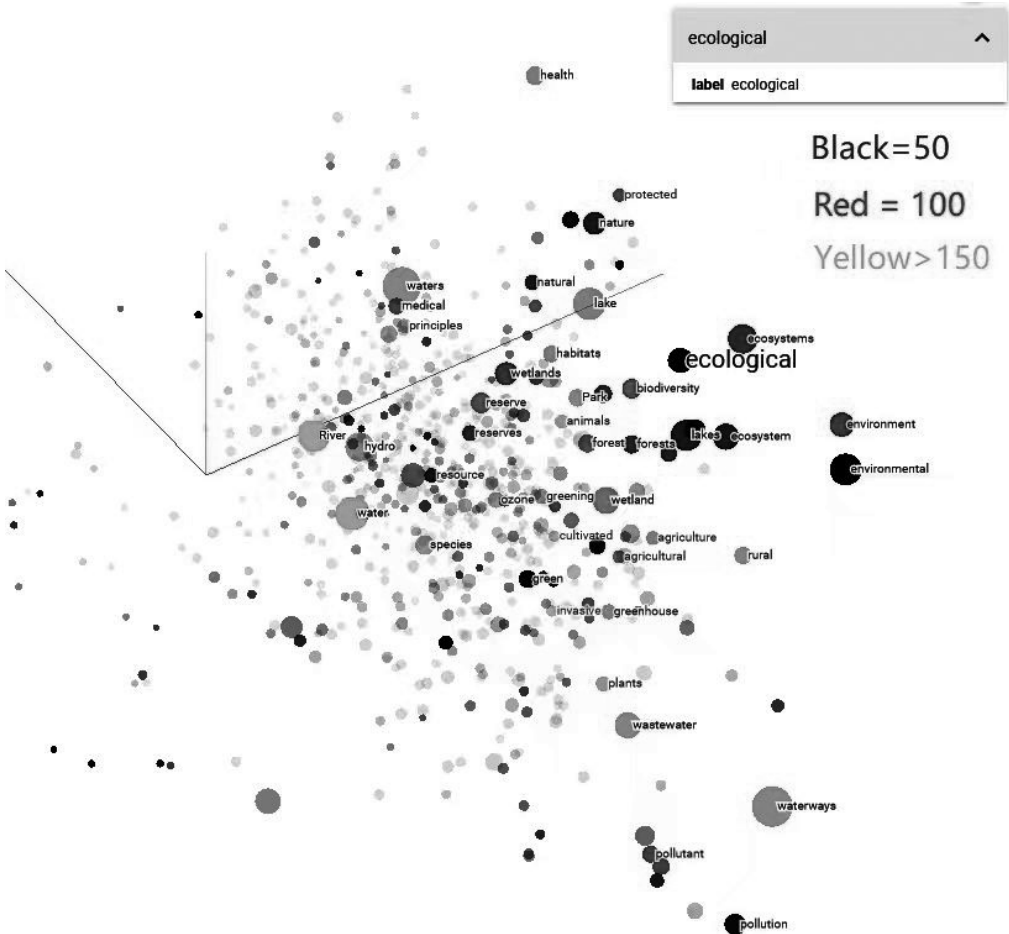


图5 “十四五”规划英译文本生态部分三维空间语义网络

图5显示，以 ecological 为核心节点的语义网络呈现出三层辐射结构：核心理念层（黑色节点，语义距离  $\leq 50$ ）包含 ecological、environmental、green，ecological-environmental-green 构成铁三角，印证了国家生态文明建设的整体性，从单一的环保向“五位一体”的范式跃迁，体现“保护与发展”的动态平衡，对应“绿水青山就是金山银山”的转化机制；生态基础层（红色节点，语义距离  $\geq 100$ ， $< 150$ ）的词汇 ecosystems、environment、biodiversity、nature、resource 体现出生态

文明建设的宏观战略性，生态系统作为物质能量循环的载（ecosystems），承载着环境质量的整体性平衡（environment）和生物多样性的遗传基因库功能（biodiversity），是自然资源永续利用（resource）的根基，也是人类与自然和谐共生（nature）的价值纽带，其通过“山水林田湖草沙生命共同体”的系统治理逻辑，共同构成国家生态安全屏障的核心要素。在生态红线管控、自然资源资产产权制度、国土空间规划等政策实践中，这些要素的协同保护正是实现生态产品价值转化、推动绿色低碳循环发展的底层支撑；黄色层级为具体的生态分支（黄色节点，语义距离>150）包含 lake、river、wetlands、waterways、wastewater、habitats、animals、plants、park、reserve、agriculture、ozone、greenhouse 等，体现了国家生态文明建设中“山水林田湖草沙生命共同体”的整体治理理念。其中，水域生态系统（湖泊、河流、湿地、水道）的节点集群凸显了水资源管理与水环境治理在国家生态安全格局中的核心地位，既涉及城市建设中的雨洪调蓄，也包含对水污染（废水、污染物）防治的刚性约束。生物多样性保护维度（栖息地、动植物）与自然保护地体系（公园、保护区）相互交织，印证了生态保护红线制度对关键物种及其生境的严格守护，同时农业生态节点（耕地、农村）与绿色转型（绿化、温室）的关联，反映了粮食安全与生态农业协同发展的战略导向。这些要素通过语义网络的密集连接，生动诠释了中国在统筹生产、生活、生态空间过程中，坚持系统治理、源头治理的生态现代化路径。

## 4.5 文本生态情感分析

### 4.5.1 文本句子情感分析

本研究采用 VaderSentiment 情感分析工具，基于“多元和谐，交互共生”生态哲学观，对“十四五”规划英译文本生态部分语境进行情感量化，通过 Python 正则表达式（英语分句函数：'(?<=[!?!]) +'）对文本进行分句并计算每句的复合情感得分（范围在[-1,1]），并使用 python 第三方制图库 matplotlib 对数据分析结果进行可视化，依据语义取向将句子划分为三类：（1）积极情感（复合得分 $\geq 0.5$ ）：体现生态保护与和谐共生意愿；（2）消极情感（复合得分 $< 0$ ）：指向生态破坏或威胁；（3）中性情感（ $0 \leq \text{得分} < 0.5$ ）：描述客观事实或政策机制。

如图 6 所示，“十四五”规划英译文本生态部分共有句子 75 句，其中积极情感占比 81.3%（61 句），表征政策文本的生态建设主导取向（如：We will persist in the management of mountains, rivers, forests, fields, lakes, and grasslands, focus on improving the self-repair capacity and stability of the ecosystem...）；中性情感占比 9.3%（7 句），集中于生态相关的政策机制描述（如：We will formulate and implement ecological protection compensation regulations.）；消极情感占比 9.3%（7

句)，主要围绕生态环境问题治理（如：We will build an environmental infrastructure system that integrates sewage, garbage, solid waste, hazardous waste, and medical waste treatment...）。总的来说，我国生态文明政策呈现“目标引领—制度支撑—问题治理”的立体性框架：积极情感主导（81.3%）彰显以“绿色发展”“碳中和”为核心的生态优先战略方向，通过高比例正向表述强化政策执行的信心与社会动员力；中性情感（9.3%）集中于生态补偿、环保督察等机制设计，以规范化语言保障政策落地操作性；消极情感（9.3%）精准聚焦污染治理、生态脆弱性等现实挑战，凸显问题导向的治理紧迫性。该情感分布结构既契合“统筹保护与发展”的治理逻辑——以愿景激励行动、以制度固化路径、以反思优化实践，也在传播策略上平衡了公众信心塑造与风险警示需求，反映了中国生态文明建设中顶层设计与问题靶向相结合的系统性思维。

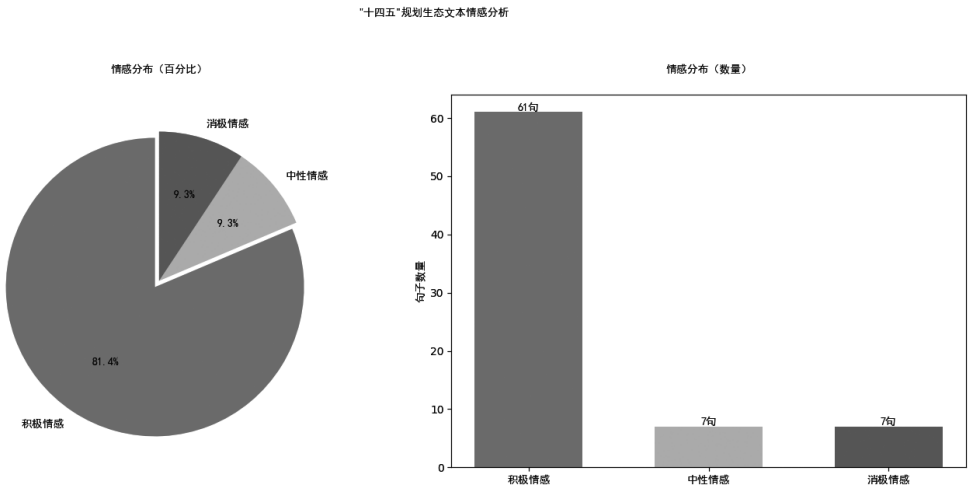


图6 “十四五”规划英译文本生态部分句子情感分析

4.5.2 文本形容词词汇情感分析

形容词能较好地体现出文本情感的重要特征（Hunston et al., 2019；Su et al., 2019；Wen et al., 2022），本研究采用Python第三方库Spacy、VaderSentiment与matplotlib等库工具，对“十四五”规划英译文本生态部分的形容词进行识别与统计，计算并可视化形容词积极情感、消极情感与中性情感的情感得分（积极为[0.5, 1]，消极[-1, 0.0]，中性为[0.0, 0.49]）与其在文本中出现的频率。

由图7可知，文本中共338个形容词，积极情感共46个（占比13.6%），主要有comprehensive（11次）、natural（8次）、clean（5次）等；消极情感共10个（占比3%），主要有low（4次）、vulnerable（1次）、blind（1次）等；中性情感共282个（占比83.4%），高频词主要有ecological（32次）、environmental（24次）、key（19

次)等。关键词检索与具体话语分析表明,政策表述呈现出“理性客观为主、积极引导为辅”的特点。中性情感形容词占比高达83.4%,高频词如ecological(32次)、environmental(24次)和key(19次)的密集使用,反映出文本聚焦于系统性、专业化的生态战略描述,强调环境要素的客观属性与重点领域的规划部署,符合政策文件注重事实陈述与路径设计的特征。积极情感词占比13.6%,以comprehensive(全面)、natural(自然)、clean(清洁)为核心,既凸显了生态治理目标的整体性、自然本底保护的重要性,也传递出污染防治的积极导向,体现政策对绿色发展愿景的引领作用。消极情感词仅占2.95%,且low(低碳)、vulnerable(脆弱)、hazardous(有害的)等词多用于客观陈述现存挑战或技术指标(如低碳发展),而非渲染负面情绪,既暗示了消极的生态问题转化为具体治理措施,又保持了政策文本的严谨性。总体而言,文本通过情感词汇的梯度配置,构建了“立足现实问题—明确重点任务—展望积极目标”的叙事逻辑,既彰显生态文明建设的决心,又维持了政策表述的稳健性与可操作性。

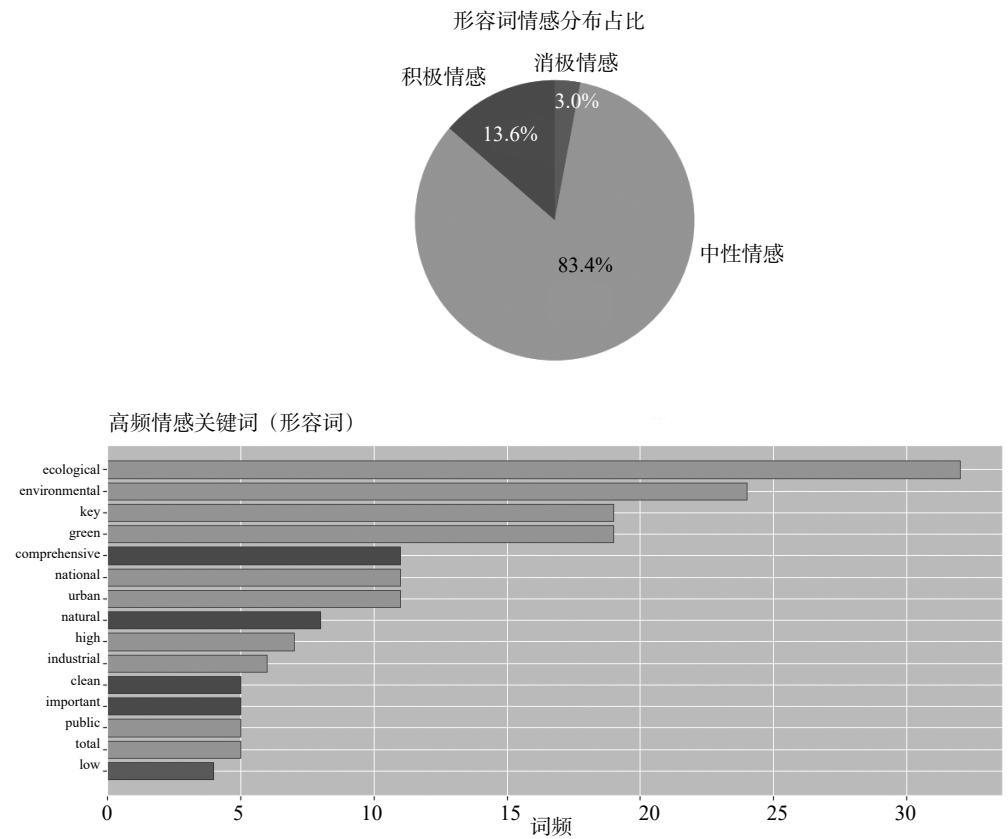


图7 “十四五”规划英译文本生态部分形容词情感分析

## 5 结语

本研究深度剖析文本所体现出的国家生态发展观念、意识情感倾向、国家在生态环境问题上的参与角色及蕴含的语言生态学意义，阐释国家如何通过“十四五”规划政策文件引导人们形成与自然和谐共生的生态型发展观念。生态环境部副部长翟青在党的二十大新闻中心举办的记者招待会上介绍：“过去十年，在习近平生态文明思想科学指引下，我们坚持绿水青山就是金山银山的理念，坚持山水林田湖草沙一体化保护和系统治理，生态文明建设和生态环境保护发生历史性、转折性、全局性变化。”本研究在“十四五”规划英译政策文件话语中，也挖掘出同样的生态性情感倾向，发现国家对待生态问题以生态积极性态度为主，扮演着积极推进生态措施实施的角色；在面对负面的生态问题上扮演着严格的纠正者、严肃的监管者的角色，表现出生态监管、生态保护的强烈决心。国家正以前所未有的决心、前所未有的力度、前所未有的成效积极推进生态文明建设与生态环境保护。人类通过语言来反映现实和对世界的认知，同时用语言来构建世界（Halliday, 1990），作为语言的使用者与研究者，关心生态问题、关心话语中的生态倾向、促进人与自然和谐共生应是自觉承担的责任，但也要警惕“以自然为中心”的非人道的思想倾向（何冰艳, 2023），国家的社会经济发展，应坚持在以人为本的前提下，倡导人与自然和谐共生的理念。

## 参考文献

- Halliday M A K, 1990. New ways of meaning: a challenge to applied linguistics[J]. *Journal of applied linguistics*, 6(13): 7-36.
- Hunston S, Su H, 2019. Patterns, constructions, and local grammar: a case study of ‘evaluation’ [J]. *Applied Linguistics*, 40(4): 567-593.
- Liao Z J, 2018. Content analysis of China’s environmental policy instruments on promoting firms’ environmental innovation[J]. *Environmental science and policy*, 88(1): 46-51 .
- Ma C, Stibbe A, 2022. The search for new stories to live by: a summary of ten ecolinguistics lectures delivered by Arran Stibbe[J]. *Journal of world languages*, 8(1): 164-187.
- Martin J, White P, 2005. *The language of evaluation: appraisal in English*[M]. New York: Palgrave Macmillan.
- Peng H T, Liu Y, 2016. A comprehensive analysis of cleaner production policies in China[J]. *Journal of cleaner production*, 135(1): 1138-1149 .
- Su H, Hunston S, 2019. Language patterns and attitude revisited: adjective patterns, attitude and appraisal[J]. *Functions of language*, 26(3): 343-371.
- Wen J, Lei L, 2022. Adjectives and adverbs in life sciences across 50 years: implications for emotions and readability in academic texts[J]. *Scientometrics*, 127(1): 4731-4749.

- 董典, 2021. 新时代新闻话语的多维度生态话语分析[J]. 外语电化教学 (1): 92-97.
- 何冰艳, 2023. 基于《道德经》哲学思想的新生态故事研究[J]. 解放军外国语学院学报 (2): 68-76+161.
- 贺丹, 唐娅华, 2021. 中国绿色服务政策演进、协同及文本内容分析[J]. 中国环境管理 (3): 92-101.
- 何伟, 程铭, 2023. 新时代生态文明建设对外传播话语与国家生态形象塑造研究[J]. 外语电化教学 (4): 89-91+125.
- 何伟, 程铭, 2024. 基于文本挖掘技术的美国清洁能源推文生态观研究[J]. 中国外语 (6): 26-34.
- 何伟, 耿芳, 2018. 英汉环境保护公益广告话语之生态性对比分析[J]. 外语电化教学 (4): 57-63.
- 何伟, 马子杰, 2020. 生态语言学视角下的评价系统[J]. 外国语 (1): 48-58.
- 何伟, 魏榕, 2017. 国际生态话语之及物性分析模式构建[J]. 现代外语 (5): 597-607+729.
- 何伟, 魏榕, 2018. 多元和谐,交互共生——国际生态话语分析之生态哲学观建构[J]. 外语学刊 (6): 28-35.
- 何伟, 张瑞杰, 2017. 生态话语分析模式构建[J]. 中国外语 (5): 56-64.
- 胡庚申, 2011. 生态翻译学的研究焦点与理论视角[J]. 中国翻译 (2): 5-9+95.
- 黄国文, 2017. 从系统功能语言学到生态语言学[J]. 外语教学 (5): 130-141.
- 冉连, 2020. 1949—2020我国政府绿色治理政策文本分析:变迁逻辑与基本经验[J]. 深圳大学学报 (人文社会科学版) (4): 46-55.
- 邵珊珊, 彭俊, 2021. 新冠疫情下基于生态评价视角的中外电商消费者话语分析[J]. 当代外语研究 (06): 132-143.
- 斯丽娟, 王佳璐, 2018. 农村绿色发展的政策文本分析与政策绩效实证[J]. 兰州大学学报 (社会科学版) (6): 148-157.
- 孙炬, 国金晖, 2022. 及物性系统视角下亚马孙雨林大火报道的生态话语分析[J]. 中国外语 (06): 70-75.
- 杨卫军, 孙慧明, 2020. 中国共产党生态发展观的思想逻辑[J]. 学习论坛 (7): 5-11.
- 姚修杰, 2020. 习近平生态文明思想的理论内涵与时代价值[J]. 理论探讨 (2): 33-39.
- 张琳, 胡开宝, 2024. 基于语料库的环境和生态话语分析研究: 现状、问题与未来[J]. 外语教学理论与实践 (3): 28-34+98.
- 赵蕊华, 2022. 生态语言学视域下的环境优先与环境外溢——以美国三大报北极话语为例[J]. 中国外语 (5): 42-50.

## 通信地址

524088 广东省湛江市 广东海洋大学外国语学院

(责任编辑: 李添豪)

# 西夏文数字化的进程<sup>\*</sup>

中国社会科学院语言研究所 / 中国社会科学院（中国社会科学院大学）

语言学重点实验室 孙 森

**提 要：**20世纪初，黑水城出土了大量西夏文献，远超此前零星发现，奠定了现代西夏学基础。西夏文研究在过去一世纪已经有了相当的文献解读与研究基础。当前人工智能浪潮下，高精度西夏文自动识别已经实现，并正向自动分词、机器翻译及大规模语料库构建迈进，从而提升了研究效率与广度，推动西夏学从传统释读迈向计算语言学与语料库语言学驱动的数字西夏学新纪元。中国学者的努力正改变“西夏研究在国外”的格局。西夏文语料库及数字工具的完善，将加强中国在国际西夏学中的话语权，推动西夏学成为高度自动化、智能化的现代人文学科，为构建跨学科知识体系提供核心支撑。

**关键词：**西夏文；数字化；自动分词；数据库

## The Digitalization Process of the Tangut Script

Sen Sun

**Abstract:** Since the early twentieth century, the substantial corpus of Tangut manuscripts unearthed at Khara-Khoto laid the foundation for modern Xixia Studies, which was far exceeding earlier sporadic discoveries. Over the past century, research on the Tangut script has established a solid foundation in textual interpretation and research itself. In the current wave of artificial intelligence, high accuracy of optical character recognition (OCR) for the Tangut script has been achieved, and the field is now advancing toward automatic word segmentation, machine translation and the construction of large-scale corpus. These developments have enhanced both the efficiency and scope of research, propelling Xixia Studies from traditional textual decipherment into a new era of digital Xixia Studies driven by computational and corpus linguistics. The efforts of Chinese

<sup>\*</sup> 本文系中国社会科学院“绝学”、冷门学科资助计划（项目编号：25JXLM08）和国家社科基金青年项目“西夏文解经文献字词研究”（项目编号：25CYY115）的阶段性研究成果。

scholars are reshaping the conventional paradigm whereby “Xixia studies were conducted overseas.” The continued refinement of the Tangut corpus and related digital tools will strengthen China’s voice in international Xixia Studies, transform the discipline into a highly automated and intelligent modern humanities field, and provide core support for the construction of an interdisciplinary knowledge system.

**Keywords:** Tangut Script; digitalization; word segmentation; corpus

## 1 引言

数字化(digitalization)是通过转换、呈现和增强等技术过程,创建数字制品(Gradillas et al., 2023)。西夏文数字化是以字符编码标准化、自然语言处理工具及古籍文献数字重建为核心,以期实现对西夏文信息进行形式化建模与算法化再造,既包括从纸质载体到数字文本的技术转换,也包括由此引发的计算语言学研究 with 数字人文方法。西夏文数字化既可以实现文献保护,也可推动西夏学从传统研究范式向数字人文转型,其核心难题集中于西夏文编码、自动识别、自动分词、机器翻译、知识图谱构建和数据库构建等多个方面。

## 2 西夏文的编码数字化

西夏文为西夏主体民族党项族创制,西夏时期得到广泛推行,为方块表意文字。出土文献材料证明西夏文至晚使用至明朝中期。20世纪初,西夏文献零星发现,如武威发现的《凉州护国寺感通塔碑铭》、北京白塔发现的《妙法莲华经》、居庸关发现的六体石刻“六字真言”(其中一种为西夏文)等。1909年,俄国探险家科兹洛夫在内蒙古额济纳黑水城发现了大量西夏文献,今藏于俄罗斯科学院东方文献研究所。在这批黑水城文献中,有一本西夏文汉文双解词典《番汉合时掌中珠》,为学界提供了打开西夏未解之谜的钥匙。

早在20世纪初,学者如罗福苕、本哈第和查赫,主要借用汉字研究范式,对西夏文的“部首”进行了归类。1914年,罗福苕对西夏文的字形等问题进行了分析,认为西夏文是一种表意文字,构成原则与汉字相同,并划分出了西夏文字字根或部首(罗福苕,1914)。罗福成则在《西夏国书类编》里详细列出了西夏字部首169个(罗福成,1915)。本哈第和查赫则通过对西夏字的分析分出了几个部首,与罗福苕的观点不谋而合,但新发现了一些部首(Bernhardi et al., 1919)。伊凤阁认为西夏字总体上是在汉文字典《说文解字》的基础上创制而成的(伊凤阁,1923)。西田龙雄在其以更为系统和穷尽性的方法深化了结构分析,总结出350个“西夏文字要素”,并确定了44种文字组合样式(如“左右结构”“左中右结构”

等)。至于文字构造所反映的西夏人思维特点,西田龙雄指出西夏人领会客观事物的思路与汉族人、藏族人等存在显著差异,但这需要历史化学或民族学的充分理据来阐明,而非简单的机械归纳(西田龙雄,1989)。出土文献印证西夏不止《番汉合时掌中珠》这一本词典,同时存有类似《说文解字》的《文海》,同样按照“从某从某”形式对字形进行解说,此外还有《同音》《同义》《五音切韵》等不同字典或者词典,为后世研究提供了宝贵的内部材料。西夏学的研究也不仅仅局限于对西夏文字的研究,诸多学者如毛利瑟、劳费尔、伊凤阁、聂历山、西田龙雄、克平、索夫罗诺夫等学者已经开始进行语言学层面上的研究,对西夏语音面貌、语系归属等问题进行了研究。而如克恰诺夫、戈尔巴乔娃、龙果夫、西田龙雄、周叔迦等学者则对西夏文献的基本面貌进行了研究,编纂了西夏文献目录,为西夏学研究提供了便利。诸多中外学者对西夏文献进行了解读,使得西夏文献的基本面貌得以清晰地展现在世人面前。

西夏文编码最早的方案是格林斯蒂德于1972年发表的编码方案,他的工作基于苏敏整理的5819个西夏单字表,并采用了鄂山荫《汉俄字典》的部首检索方式(从右下角一笔算起,然后是右边,最后是左边部首)。然而,他的思考始终集中于如何避免“重码”,导致其设计出的是一套缺乏理据的“电报码”,未能作为字符集内码使用(Grinstead, 1972)。使西夏文字进入电脑的尝试在20世纪80年代末由莱顿大学的范德利姆和俄罗斯的克平共同主持。但由于当时电脑字体制作技术尚不成熟,难以处理“上中下结构”等复杂的复合西夏字,最终仅约三分之二的文字得以录入,项目被迫中断。1996年,中岛干起主持研制的第一个完整的西夏字库在日本问世,该字库在技术层面上达到了极高水平,并被用于公布西夏文字的各种统计分析结果。后中岛干起与中国学者李范文合作出版有关相关成果(李范文等,1997)。随后,宁夏大学计算中心马希荣带领开发的“夏汉字处理系统”于1999年出版发行。该系统在实用性上有所侧重,采用了李范文《夏汉字典》中的四角号码作为输入方法,但由于课题组成员缺乏西夏语文专业知识,未能进行严密校核,导致字形错误、错码、漏码等瑕疵丛生。相比之下,有西夏学家自始至终深度参与校勘的字库如台湾“中央研究院”的字库,其质量则得到了国际学术界的普遍认可。从字库设计实践中不难看到,字形的准确率已成为衡量字库成功的首要指标。西夏文字笔画繁复,区别只在毫厘之间,学者们在缺乏西夏文字规范文献的情况下,只能凭个人经验和《文海》的字形解说进行校对。经过历代学者的不断努力,西夏文编码工作的努力在国际标准层面取得了突破。2007年,美国加州大学伯克利分校的Richard Cook提交了西夏文编码字符集提案。中国专家随即提出了专业修订意见,强调应采用景永时教授开发的字体,并收入其所列出的所有“正字”。在各国专家的共同努力下,西夏文最终于2016年成功被纳入Unicode 9.0版本(编码空间为U+17000 - U+187FF),共收录6,125个字符。这一里程碑式的成就,标志着困扰

学界数十年的西夏文信息基础设施建设的圆满完成，为后续西夏学数字化工作奠定了坚实基础。

### 3 西夏文的自动识别技术

自动识别技术，即通常所说的光学字符识别技术（Optical Character Recognition，简称OCR）。汉文自动识别技术较为成熟，现代排版的自动识别十分准确，历史文献如竖排古汉语的自动识别也较为成熟，处理平台有古联OCR系统、中华书局古籍智能整理平台、如是古籍数字化工具平台、识典古籍等。云聪古籍数字化平台支持多种少数民族文字的OCR识别，比如藏文、蒙古文、哈萨克文、维吾尔文、彝文和苗文等，但关于西夏文OCR的处理平台目前较少，目前有中国社会科学院民族学与人类学语言学实验室开发的“龙水西夏文识别软件”可供识别西夏文，中国社会科学院语言研究所语言学重点实验室开发的平台可识别印刷体西夏文、原始图版与汉文、西夏文混合文种识别。

目前，西夏文识别主要集中在西夏原始图版的采集、标注、预处理与识别上，即先对原始图版进行采集，形成一个数据集。而在方法上，目前学界基本已经进展到基于深度学习的文字识别阶段，这种识别一般采用的是监督式学习，即训练和测试的数据是经过专家识别并进行标注。深度学习与传统的机器学习相比其优势在于，只需要将尽可能多的、准确的标注数据提供给神经网络，而无须教给神经网络如何进行分类（即无须提供人工特征进行分类），实际分类能力是深度神经网络从提供给它的训练集中字形学到的。基于深度学习的文字识别需要比较大规模的训练与测试数据集，每个字符往往需要几百上千个不同的实例。训练数据集包含了神经网络的学习目标，即数据的标签，我们用训练数据集“教会”神经网络“认识”文字。而测试数据集的数据与训练集没有任何交集，在训练神经网络的过程中用来检验其学习效果。训练深度神经网络模型时，标注数据集的规模和质量对识别的准确性和适应性至关重要。

西夏文识别的研究始于将通用图像识别领域的深度学习模型进行迁移和定制。文字识别本质上是图像识别的一个特化任务。面对实际的古籍文献，文本的准确检测与分割是实现高精度OCR的前置关键环节。近年来，研究焦点已从单字识别转向复杂版式处理。文本行与区域检测方面，刘佟采用基于迁移学习的CRAFT文本检测模型，在中文数据集预训练的基础上，应用于自制的西夏文数据集，有效解决了文本区域的定位问题（刘佟，2022）。景诗云（2024）针对西夏古籍中文本行间距密接、长短分布不均等问题，基于PSENet实例分割方法构建了SAC-PSENet检测模型。岳霄（2024）提出了SA-CE-DBNet西夏文古籍文本检测方法，在与主流算法的对比中展现出优势。这些研究极大地提升了系统处理复杂古籍图像的能力。

西夏文字由于笔画繁多导致很多字形十分相似，这种事实对识别算法提出了极高的要求，不同的学者使用不同模型对西夏文 OCR 提出了自己的方案。孟一飞基于四角号码编码的识别路线，通过改进短点引导的霍夫变换应用于文字笔画检测，提高了系统容错率，同时建立了西夏文字的单字和文本数据集，为后续研究提供了基础语料。此外，该研究利用生成对抗网络扩展西夏文字样本集，有效缓解了样本不均衡分布的问题，并证明了扩展样本集在提高深度学习模型识别率方面的有效性（孟一飞，2019）。张光伟（2020）提出可以借鉴成熟的神经网络架构，并针对西夏文的特点进行优化调整，在特定标注数据集上获得了超过 98% 的识别准确率。王维淇（2023）基于改进 YOLOX 的西夏文检测方法，利用深度神经网络结构学习西夏古籍图像中整体和部分的分界关系，通过特征提取网络获取有用的特征信息，检测定位后将图像的特征信息输送到分类器进行分类识别，并使用边界框来标记西夏文的位置。二阶段目标检测算法总体将目标的检测和分类分为两个阶段，而一阶段目标检测算法无须形成候选框，研究显示一阶段目标检测算法在不影响检测速度的前提下，大大提高了检测精确度，甚至超越了二阶段目标检测方法。刘佟（2022）利用一种基于 DS-CBAM-DenseNet 的识别模型，通过通道注意力模块（CBAM）抑制背景干扰，使模型更专注于西夏文字的笔画特征。景诗云（2024）基于 Res2Net 的识别方法，可以利用其多尺度特征提取能力扩大感受野，应对复杂字形。在低资源条件下，使用 500 张真实标注图片的前提下实现了 83.10% 的准确率（郑宇熹 等，2025）。这些技术标志着西夏文 OCR 研究正从依赖大量标注的监督范式转向探索高效的自学习范式，以应对数字人文领域中文献资源的复杂性和稀缺性。

总之，学界通过各种技术在西夏文 OCR 领域取得了显著进展，直接促进了西夏文献研究。

## 4 西夏文的自动分词研究

词是能独立运用的最小语言单位，是承载意义最基本的单位，而分词是指针对词与词之间不存在间隔标记的文本，基于一定规范，将文本按分词单位进行划分的过程。分词能够应用于词性标注、命名实体识别、智能检索等下游任务中，是自然语言处理中很重要的环节，具有重要的应用价值。进行自动分词的研究，分词规范的制定是基础且不可或缺的一环。目前我国，汉语和藏语都有了分词的国家标准，这些规范大大促进了各文种自动分词研究的发展。

目前西夏文的自动分词领域基本为空白，因此有必要借鉴已有现代汉语、古代汉语及民族文字如藏文的自动分词经验进行研究。

## 4.1 汉语自动分词研究

汉语自动分词是中文自然语言处理的核心基础内容。由于汉语以连续字序列为基本书写形式，空格的缺失使词语边界识别成为首要难题。自动分词旨在将字序列合理切分为语义和语法独立的词单元，而词性标注在词序列上赋予语法范畴标签。

汉语分词研究经历了统计学习模型、深度学习模型和预训练语言模型（pre-trained model，简称PLM）等不同阶段。统计学习模型中，隐马尔可夫模型（hidden Markov model，简称HMM）和条件随机场（conditional random field，简称CRF）成为主流，HMM利用马尔可夫性质简化序列建模，而CRF作为判别式模型，则能更灵活地利用上下文特征进行联合建模。然而，统计方法严重依赖人工设计的特征模板，难以实现跨领域泛化。进入深度学习时代后，研究焦点转向字符嵌入（embedding）与循环神经网络（recurrent neural network，简称RNN）。在此基础上，双向长短时记忆网络（bi-directional long short-term memory，简称BiLSTM）作为RNN的改进模型，能够自动捕捉长距离上下文依赖，再结合CRF解码层来约束标签间的合法转移，从而显著提升了分词与词性标注的性能。这一阶段的研究通过优化词向量预训练和混合输入策略，推动模型性能接近甚至超越复杂结构。

自BERT等PLM问世以来，中文自动分词和词性标注迅速进入预训练时代。预训练语言模型通过在大规模无标注语料上进行预训练，将丰富的词汇和语义知识内化到模型参数中，极大缓解了对外部词典和复杂特征工程的依赖。例如，全词掩码（whole word masking，简称WWM）策略显著改善了BERT对中文词级语义的表示（Cui et al., 2021）。学界目前大多数方法都是直接在预训练模型中加入序列层，局限在于不能在BERT模型底部的层加入词汇信息，导致BERT能力得不到充分利用。有学者提出Lexicon Enhanced BERT（LEBERT）模型，该模型在BERT底部的层中加入一个Lexicon Adapter层来融合文本的词汇特征，实验结果表明该模型在序列标注类任务的10个数据集上的表现超越了学界现有的其他模型（Zong et al., 2021）。此外，研究者还提出一种用于中文分词的对抗性多准则学习方法，该方法整合了来自多个异构切分标准中的共享知识。在八个采用异构切分标准的语料库上的实验表明，与单准则学习相比，每个语料库的分词性能均有显著提升（Chen et al., 2017）。

然而，当前研究仍面临古籍文本与低资源场景的适应性的问题。现代汉语分词的研究相对充分，但面对古汉语、民族历史文献等低资源文本，性能大幅下降。最新研究表明，大语言模型（large language model，简称LLM）在经过指令微调后，在跨语言典籍分词任务中展现出显著的性能优势，验证了大语言模型在跨语言典籍自动分词任务中的有效性和潜力（王希羽等，2024）。

综上，汉语自动分词与词性标注的研究已进入高度成熟的预训练阶段，为数字人文领域的进一步发展提供了坚实的技术基础。

## 4.2 藏语自动分词研究

当前, 诸多民族语言学界也已或着手准备自动分词研究, 如藏语、彝语等民族语言研究已经卓有成效, 因藏语自动分词已经发展较长时间, 取得了较大成就, 因此也成为我们的借鉴对象。

藏文自动分词 (Tibetan word segmentation, 简称 TWS) 是藏文自然语言处理的基础性工作。与汉语自动分词不同的是, 藏文以 “ ” 分隔音节, 而词通常由若干音节构成, 词边界模糊但极富变化。这种 “音节分写, 词语连写” 的混合结构, 加上不同语素组合、方言差异及借词现象, 使得藏文自动分词成为低资源语言自然语言处理中的一项独特挑战。

藏文信息处理经历了几十年的发展。早期, 基于字符串匹配的分词方法运用较为广泛。藏文分词基本框架的搭建基于 12 万词条和 500 万字藏语真实文本语料实践, 归纳为藏文计算机分词的 36 条基本规则 (罗秉芬 等, 1999)。后有学者根据藏语丰富的句法标记, 构建不同的藏语句法组块类型, 建立相应的分块规则, 然后进行分词 (陈玉忠 等, 2003)。但这种做法对于未登录词的识别不够精准。在汉语分词理论和方法的基础上建立了为切分、格框架和标注一体化的藏语三级切分体系 (祁坤钰, 2006)。“班智达藏文分词系统” 可分三步实现分词功能, 即先将文本分词句、字, 再通过格助词将句子分成组块, 块内再通过词典匹配切分成词 (才智杰, 2010)。后期, 基于统计的分词方法运用在藏文分词中。央金藏文系统借鉴汉语分词系统 Segtag, 该系统主要采用 HMM 模型。SegT 则运用双向切分检测交集型歧义字进行处理, 使得系统分词的正确率得到提升 (刘汇丹 等, 2012)。TIP-LAS 藏文分词系统则是利用 CRF 实现基于音节标注的藏文分词 (李亚超 等, 2015)。近年来, 由于藏文语料规模小、分词准确率较低等问题, 基于统计和规则相结合的分词方法被应用在藏文分词上。条件随机场和规则融合方法用来解决藏文分词问题, 正确率达到了 96.11% (洛桑嘎登, 2016)。最大熵模型和 CRF 被同样应用于藏文分词当中, 基于转换的学习方法用于对条件随机场模型的分词结果进行修正 (龙从军 等, 2016)。进入深度学习时代后, RNN 及其变体, 特别是 BiLSTM 与 CRF 层的结合 (BiLSTM-CRF), 已经被应用在藏文分词领域并且效果较好 (李博涵 等, 2018)。有学者提出一种藏文分词的神经网络架构, 只需监督式训练的标注数据 and 无监督学习迁入表示的未标注语料, 无须介入人工特征工程 (桑杰端珠 等, 2018)。

近年来, 预训练语言模型如 BERT 等为藏文自动分词研究提供了巨大帮助, 这些模型通过在大量语料上进行预训练, 提供了强大的上下文双向编码能力。有学者在基于多语言 BERT 的藏文预训练基础上, 利用小规模藏文分词语料进行微调, 得到一个使用藏文分词的特定模型或者本地模型, 验证藏文预训练语言模型的性能 (索朗次仁, 2023)。

目前藏文分词取得了重大成绩,但由于藏文本身的文字形式、语法样式及规则等问题,在具体研究过程当中还存在虚词识别、歧义消除、OOV识别等问题。尽管如此,藏文作为汉藏语系中典型的低资源语言,其处理经验对西夏文的数字化研究具有直接的借鉴意义。藏文信息处理的成功得益于国家标准的制定,如《信息处理用藏文分词规范(GB/T 36452-2018)》,这为TWS提供了统一的理论框架和评测基准,是实现学科规范化的关键一步。

### 4.3 西夏语自动分词研究

当前学界还未有人涉及西夏语自动分词研究,我们根据学界解读基础,在“四行对译”的基础上添加第五行,进行人工分词标注。当前已经对世俗文献和宗教文献两种不同性质的文献标注约25,000字。未来将借鉴汉语自动分词和藏文分词的研究方法,在人工标准训练的基础上,采用统计模型、深度学习模型尤其是预训练模型等三种不同路径,获得可用的分词系统。

## 5 西夏文数字化研究的未来

西夏文数字化研究的未来,需从当前的字符图像识别迈向计算认知的更高维度。由于西夏文数字语料的稀缺性与非标准化现状,首要任务是基础设施建设。目前有关单位都在加紧相关研究工作,基本逻辑与方法是相似的,即都意识到要建立一个具有层级的西夏数字语料体系,该体系至少包含基础数据、标注材料等两个方面。基础数据用于长期保存拓片与文献等高清影像,并要求严格遵循元数据标准。而标注材料的来源各家则各不相同,一般采用“四行对译”或者“隔行对照”的格式,这一部分工作能够具备为计算机所用的可能,进行字级、句级标注,以确保数据可共享性和互操作性。大部分单位基本在这一部分停滞不前,原因是无法短时间内建构起十分庞大的已经解读的文献数据群,而这需要西夏学界的共同努力。

我们目前在从事自动识别、自动分词及词性标注等研究,等研究成果上线后,即可利用人工智能构建西夏文献辅助释读框架,实现以自动化四行对译为显示结果的机器翻译,这将极大提升文献释读效率。未来应将西夏文自动分词、词性标准、命名实体识别作为优先任务,为西夏语构建能够标注分词、词性的计算语料库,从而支撑西夏语的词汇学、语法学以及语言谱系研究。同时,通过结合地理信息系统(Geographic Information System,简称GIS)与知识图谱,将西夏人名、地名与文本内容进行时空绑定,构建西夏文化数字地图和时空西夏学知识本体,为历史地理学和宗教考古研究提供动态分析与计算工具。关于知识本体将音韵、语义、部件、语法、历史地理等信息,以知识图谱(knowledge graph)形式进行结构化表示。这一体系将超越简单的数据存储,旨在构建一个可计算的西夏文明知识数据库,为后续的语义关联和知识检索提供统一的机器可理解框架。

未来的技术发展趋势将是由单纯的“视觉识别”向“语义理解”进行智能跃迁。当前的 OCR 研究主要集中于字符识别精度，而对复杂的文本结构、语法和深层语义解析仍处于起步阶段。为此，必须发展结构化视觉建模方法，结合视觉 Transformer 与图卷积网络（graph convolutional network，简称 GCN），在单字符内部建立笔画构件的拓扑关系，并扩展至句法层面，实现文字形体与语言结构的交互式建模。这不仅能提高识别的鲁棒性，还将使系统具备语义自校正能力。更进一步，多模态语义融合技术将推动语义识别进入新阶段，通过融合图像、声音（拟音）、跨语言对译文本（如汉文、藏文）等多维数据，构建跨模态的语言表示模型，能够实现文字、语言与文化信息的深层互解。

西夏学数字化任重道远，期望学界同仁齐心协力，共同致力于推动以西夏文为代表的“绝学”、冷门学科的复兴与繁荣。一方面要依托扎实严谨的传统文献学研究，深挖文本内涵、厘清历史脉络；另一方面要积极融合数字人文技术，借助计算机在数据挖掘、图像识别、知识组织等方面的强大能力，为西夏文献的整理、释读、分析与传播注入新的活力，真正实现传统与数字的双轮驱动，赋能“绝学”、冷门学科在新时代焕发出持久的生机。

## 参考文献

- Bernhardi A, von Zach E, 1919. Einige Bermerkungen über Si-hia-Schrift und -Sprache[J]. *Ostasiatische Zeitschrift*, 232-238.
- Chen X, Shi Z, Qiu X, Huang X, 2017. Adversarial multi-criteria learning for Chinese word segmentation[C]//*Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*. Vancouver: ACL: 1193-1203.
- Cui Y, Che W, Liu T, et al., 2021. Pre-training with whole word masking for Chinese BERT[J]. *IEEE/ACM transactions on audio, speech, and language processing*: 3504-3514.
- Grinstead E, 1972. *Analysis of the Tangut script*[M]. Lund: Studentlitteratur, 1972.
- Gradillas M, Thomas L, 2023. Distinguishing digitization and digitalization: a systematic review and conceptual framework[J]. *Journal of product innovation management*, 42(1):112-143.
- Zong C, Xia F, Li W, et al., 2021. Lexicon-enhanced Chinese sequence labeling using BERT adapter[C]//*Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*. Stroudsburg: ACL: 5847-5858.
- 才智杰, 2010. 班智达藏文自动分词系统的设计与实现[J]. *青海师范大学民族师范学院学报*, 21(2): 75-77.
- 陈玉忠, 李保利, 俞士汶等, 2003. 基于格助词和接续特征的藏文自动分词方案[J]. *语言文字应用* (1): 75-82.
- 景诗云, 2024. 基于深度学习的西夏古籍文本检测与识别算法研究[D]. 银川: 宁夏大学.
- 李博涵, 刘汇丹, 龙从军, 等, 2018. 基于深度学习的藏文分词方法[J]. *计算机工程与设计*, 39(1): 194-198.

- 李范文, 中嶋幹起, 1997. 电脑处理西夏文〈雜字〉研究[M]. 东京: 東京外国語大学アジア・アフリカ言語文化研究所.
- 李亚超, 江静, 加羊吉, 等, 2015. TIP-LAS: 一个开源的藏文分词词性标注系统[J]. 中文信息学报, 29 (6): 203-207.
- 刘汇丹, 诺明花, 赵维纳, 等, 2012. SegT: 一个实用的藏文分词系统[J]. 中文信息学报, 26 (1): 97-103.
- 刘佟, 2022. 基于深度学习的西夏文古籍文献识别研究与实现[D]. 银川: 宁夏大学.
- 龙从军, 刘汇丹, 2016. 藏文自动分词的理论与方法研究[M]. 北京: 知识产权出版社.
- 罗秉芬, 江荻, 1999. 藏文计算机自动分词的基本规则[C]//中国少数民族语言文字现代化文集. 北京: 民族出版社.
- 罗福苾, 1914. 西夏国书略说[M]. 京都: 东山学社.
- 罗福成, 1915. 西夏国书类编[M]. 京都: 东山学社.
- 洛桑嘎登, 2016. 藏文自动分词与词性标注研究[D]. 北京: 中央民族大学.
- 孟一飞, 2019. 改进的霍夫变换在文字笔画检测识别中的应用研究[D]. 银川: 宁夏大学.
- 祁坤钰, 2006. 信息处理用藏文自动分词研究[J]. 西北民族大学学报(哲学社会科学版) (4): 92-97.
- 桑杰端珠, 才让加, 2018. 神经网络藏文分词方法研究[J]. 青海科技, 25 (6): 15-21.
- 索朗次仁, 2023. 基于预训练语言模型的藏文分词与词性标注研究[D]. 拉萨: 西藏大学.
- 王维洪, 2023. 基于深度学习的西夏古籍文本检测与识别算法研究[D]. 银川: 宁夏大学.
- 王希羽, 王东波, 2024. 基于大语言模型的跨语言典籍自动分词研究[J]. 图书馆杂志, 2024, 44 (9): 104-115.
- 西田龙雄, 1966. 西夏语の研究——西夏语の构拟と西夏文字の解读 2[M]. 东京: 座右宝刊行会.
- 西田龙雄, 1989. 西夏王国的话[M]. 东京: 大修馆书店.
- 伊凤阁, 1923. 西夏国书说[J]. 北京大学国学季刊, 1 (4): 122-133.
- 岳霄, 2024. 基于深度学习的西夏文古籍文献文本检测与识别算法研究[D]. 银川: 宁夏大学.
- 张光伟, 2020. 基于深度学习的西夏文献数字化[C]//第二届西夏学国际论坛论文集. 银川: 宁夏大学.
- 郑宇熹, 周子茗, 张永伟, 等, 2025. 低资源条件下的西夏文OCR与机器翻译研究[J]. 数字人文 (3): 113-135.

## 通信地址

100732 北京市 东城区建国门内大街5号

(责任编辑: 江姝婷)

# 《功能性译者风格研究：语料库辅助研究法》述评

北京外国语大学 郭嘉雯

**提 要：**吴侃和李德凤的专著《功能性译者风格研究：语料库辅助研究法》聚焦语料库翻译学中的译者风格问题，以“功能性译者风格”为核心，系统梳理相关理论脉络、语料库辅助研究路径及其在文学翻译中的应用。针对既有研究偏重形式特征、忽视文本功能与文学效果的不足，该书将译者风格界定为译者在目标文本中重构功能时呈现的语言特异性，并以老舍《二马》及其多个英译本为例展开实证分析。该书推动了译者风格研究由“语言指纹”识别向“功能重构”阐释的拓展，强化了语言描写与文学阐释的结合，对文学翻译研究与语料库翻译学发展具有启发意义。本文拟从内容框架、理论方法、研究价值与不足等方面加以评述。

**关键词：**功能性译者风格；语料库辅助研究法；文学功能；文本功能

## ***A Review of Researching Translator's Functional Style: A Corpus-Assisted Approach***

Jiawen Guo

**Abstract:** The monograph *Researching Translator's Functional Style: A Corpus-Assisted Approach* by Kan Wu and Defeng Li focuses on the issue of translator style within corpus-based translation studies. Focusing on the notion of “translator's functional style,” it offers a systematic review of the relevant theoretical traditions, corpus-assisted research methods, and their application to literary translation. Addressing the shortcomings of existing studies, which have tended to emphasize formal features whilst neglecting textual function and literary effect, the book conceptualizes translator style as the distinctive linguistic idiosyncrasies shown by translators as they (re)create functional aspects in target texts. This perspective is substantiated through an empirical case study of Lao She's *Er Ma* and its various English translations. Advancing the study of translator style from the identification of linguistic fingerprints to the interpretation of functional reconstruction, the

book reinforces the integration of linguistic description with literary interpretation. It offers valuable insights for literary translation research and the development of corpus-based translation studies. This paper reviews the monograph in terms of its content, theoretical and methodological approaches, scholarly significance, and limitations.

**Keywords:** translator's functional style; corpus-assisted approach; literary function; textual function

## 1 引言

译者风格研究是语料库翻译学的重要议题，但其核心概念界定与研究路径迄今仍存在分歧。一方面，既有研究大多沿袭Baker（2000）以来的“语言指纹”思路，侧重从词汇、句法等形式特征识别译者个性；另一方面，强调源语与译语关系的研究则表明，若脱离文本功能与文学效果，仅凭形式层面的统计，难以充分说明译者风格的生成机制。尤其在文学翻译研究中，如何打通语言描写与文学阐释，将功能维度纳入可操作的语料库分析，仍是当前研究中的明显缺环。吴侃（Kan Wu）与李德凤（Defeng Li）2025年由Springer出版的《功能性译者风格研究：语料库辅助研究法》一书正是围绕这一问题展开。该书在理论上提出“功能性译者风格”概念，将译者风格界定为译者在目标文本中重构文学功能与文本功能时所体现的语言特异性，回应了相关研究长期偏重形式、忽视功能的局限。在方法层面，该书融合语料库文体学与风格计量归因路径，综合运用主题词（keywords）、词汇链（lexical chain）、语义场（semantic field）、多维统计等多重手段，体现出较强的方法整合意识。在研究对象上，该书将考察范围由零散的语言特征扩展至文学翻译中的功能重构，既拓宽了译者风格研究的对象边界，也为文学翻译研究提供了具有示范意义的实证路径。

## 2 内容简介

全书共12章，分为前后两部分。第一部分涵盖第1—4章，重点讨论译者风格研究的理论谱系、语料库辅助研究的发展脉络及“功能性译者风格”概念的提出。这一部分旨在回应这样一个根本问题：译者风格何以不仅表现为形式层面的差异，而且能够作为一种功能重构现象被加以识别和阐释。第二部分包括第5—12章，以老舍小说《二马》及其四个英译本为个案，对译者功能风格进行操作化检验。

第1章主要讨论什么是译者风格，以及为何要研究译者风格两个问题。作者梳理了该领域多位学者的代表性界定，指出译者风格既可被理解为跨文本持续显现的语言选择模式，也可被理解为译者在对源文风格的感知、转化与重构过程中表现出

的特异性。基于此，本章进一步论述研究译者风格在凸显译者主体性、揭示源语文本与译语接受之间的关系以及推动翻译理论发展等方面的重要意义，并由此引出语料库文体学在弥合“语言描写”与“文学阐释”之间断裂方面所发挥的关键作用。

第2章对过去二十余年的语料库辅助译者风格研究做系统回顾，重点考察其概念界定、语料设计、分析类型及解释路径。作者区分目标文本导向（TT-oriented）与源文本导向（ST-oriented）两类研究，讨论可比语料与平行语料两种常见设计，并强调研究不应止于统计呈现，而应进入意义建构（sense-making）的解释层面。本章指出，既有研究虽已通过词汇、句法等形式指标识别译者“指纹”，但对主题建构、人物塑造等功能性特征的探讨仍相对薄弱。即便个别研究涉及功能层面，也多限于视角、话语呈现等局部维度，译者风格研究的关键缺环在于功能维度尚未获得充分的理论建构与操作化展开。

第3章转向语料库文体学本身，探讨其由语料库语言学发展而来并逐步与文体学融合的学术演进过程。作者尝试厘清语料库文体学的定义与学科定位，并系统梳理其主要研究路径。该领域持续围绕文本型式（textual pattern）与功能阐释之间的关系，展开理论与实证层面的探索。作者指出，语料库文体学并不止于对语言形式的量化统计，而在于揭示文本型式与文学功能、审美效果之间的内在关联及其阐释意义。将译者作为研究对象纳入语料库文体学视野，不仅能够拓展语料库文体学的适用范围，也能为译者风格研究提供一种兼顾语言描写与文学解释的方法论支持。

第4章是全书理论建构的核心。作者提出“功能性译者风格”概念，将其界定为译者在目标文本中重构或建构功能时所呈现的语言特异性。作者从范围和类型两个维度对功能特征进行分类，区分整体性功能与局部性功能、文学功能与文本功能。整体性功能特征指作用于文本宏观层面的功能，服务于全局功能，局部性功能特征则体现在类符/形符比、叙述视角、话语表达等特定风格要素中。文学功能主要关涉作品意义、审美效果和虚构世界的生成，文本功能则主要关涉语篇组织、连贯衔接和整体可读性。据此，作者细化出三种具有代表性的功能风格：主题风格（thematic style）、人物形象重塑风格（recharacterization style）和文本全景风格（textual-panoramic style）。本章同时讨论分析功能性风格的方法基础，包括语料库文体圈（corpus stylistic circle）与风格计量归因（stylometric attribution）两条路径，并给出具体分析框架、核心模型与操作步骤，强调通过多源数据和多种指标的三角互证，才能较为全面地把握译者风格的功能面向。

第5章开始进入个案研究。本章简要梳理老舍的旅英经历及小说《二马》的相关情况，并进一步论述该文本作为功能性译者风格研究对象的学术价值与研究意义。作者指出，《二马》以20世纪初旅英华人的生存状态为叙事中心，围绕中国人与英国人的文化认知、身份关系和社会互动展开叙述，是老舍早年经历的真实写照。作品中对“中国人”和“英国人”族群形象的反复书写，构成了小说中两条重

要的主题线索。基于对作品内容及既有文学批评研究的讨论，本章将“离散华人”与“本土英国人”界定为后续分析的两个功能焦点，从而使个案研究在研究起点上即与文学功能的识别、提炼及重构形成紧密衔接。

第6章系统交代案例研究的方法论设计。作者提出五个相互衔接的研究问题：一是识别与小说中“离散华人”和“本土英国人”直接相关的主题词；二是分析这些主题词如何与语境互动并在原作中生成主题建构与人物塑造功能；三是比较不同译者在重构这些文学功能时表现出的语言特征；四是考察译者在目标文本中如何建构“文本全景”这一文本功能；五是解释这些差异背后的可能因素。在语料方面，本章说明以《二马》中文原文、四个英译本以及参考语料库为基础，并对文本校勘、中文分词、软件使用及各分析阶段作出详细说明。由此，本章完成了从理论框架到可操作研究流程的转换。

第7章的任务是生成并分类《二马》的主题词，以建立从语言形式进入文学功能分析的起点。作者借助参考语料和中文分词技术，提取作品主题词，并依据主题词分类模型对其进行归类，最终筛选出与“中国人”和“英国人”相关、具有较强功能负载能力的主题词，如“马则仁”“黄脸”“中国人”“英国人”“洋鬼子”等。本章指出，主题词并非孤立词汇，而是主题和人物意义网络的入口。通过对其进行筛选和分类，后续章节才能进一步追踪这些词在原文与译文中所形成的扩展意义、词汇链、话语表达等词汇—语义型式（lexico-semantic pattern）。

第8章以主题重构为中心，考察四个英译本如何再现《二马》中“离散华人”与“本土英国人”两条主题。作者首先分析原文中相关主题词的搭配关系、语义场构成分布和词汇链延展方式，揭示两条主题在原作中的语言实现机制，继而比较四个译本对功能负载主题词的译法、对原有词汇链的维系程度以及对语义场构成比例的调整方式。研究发现，不同译者虽大体保留了主题框架，但在具体语义成分上表现出不同程度的偏移，从而形成各自不同的主题风格。由此说明，译者风格不仅体现为某些语言习惯，更体现为其对文本“所写为何”的重构方式。

第9章将分析重点转向人物形象重塑，讨论译者如何在目标文本中重构中国人与英国人的人物形象。作者先梳理原文中主要人物标签和相关语义场，再比较四个译本在人物标签翻译、语义场重构和话语表达方式上的差异。作者发现，译者在中国人物标签的英译形式、自由间接思想、直接思想、直接言语与自由直接言语使用倾向以及语义场构成分布的调整等方面，均存在可识别差异，这些差异进一步影响了目标读者对人物身份、性格及中英文化关系的理解，并揭示出不同译者在塑造人物形象、读者—人物心理距离与叙事沉浸性方面的不同取向。

第10章从文本功能角度讨论文本全景风格。有别于前两章主要采用语境化语料库文体学分析路径，本章结合最常用词（most frequent words）、最常用词串（most frequent word sequences）及Biber（1988）六维模型，对四个英译本实施主成

分分析（principal component analysis）和多维分析（multidimensional analysis）。作者基于 MFW 和 MFWS 列表通过 PCA 比较不同译本的整体距离关系，同时计算不同译本在六个维度上的相应得分，以描绘各译本的文本全景。研究显示，不同译本在文本全景上的聚合与离散关系并不完全等同于其在主题与人物层面的表现，例如，Dolby 的译本与 Huang 和 Finkelstein 的合译本在某些 PCA 结果中较为接近，而 James 译本与其他译本差异较大，多维分析揭示部分译本在可读性、叙事性与信息密度方面存在明显分化。

第 11 章在总结第 8—10 章分析结果的基础上，转入对风格成因的讨论。作者将相关因素区分为译者因素与超译者因素两大类，前者包括译者的语言文化背景、个人立场和个人偏好，后者则涉及社会文化语境、赞助人背景及其政策以及既有译本的后续影响。作者结合四个译本的出版背景、译者身份及相关副文本材料，尝试解释为何某些译者在主题重构上更趋近原文，某些译者在人物塑造和文本全景上显示更强干预。该章的意义在于将前述统计差异从可观察现象推进为可解释现象，使全书的实证结果真正回到翻译学意义上的主体性、意识形态与社会语境问题。

第 12 章为结语，概括全书主要发现并提出未来研究方向。作者指出，从主题重构、人物形象重塑和文本全景建构三个维度看，四个英译本既有相通之处，也存在明显差异，而且不同分析视角所呈现的聚散格局并不完全一致。这表明译者风格具有多维性和层次性，难以由单一指标加以概括。本章进一步认为，语料库文体学路径能够同时兼顾形式与功能，适合对译者风格进行多角度考察。未来研究应在更广泛的文学功能、文本功能和跨学科工具之间建立更系统的分析框架。

### 3 简评

#### 3.1 从“语言指纹”到“功能重构”

与其说本书是对既有译者风格研究的补充，不如说它回应了该领域一个持续存在的核心争议：译者风格究竟应主要被理解为可量化的语言“指纹”，还是应被放回源语和译语关系、文学功能与意义建构过程中加以把握。Baker（2000）以来的语料库研究为译者风格识别提供了坚实的经验基础，使词汇、句法、搭配等形式特征成为考察译者个性的重要入口。但是，Malmkjær（2003）、Boase-Beier（2006）等研究者亦不断提示，若脱离文本功能、审美效果和文学阐释，仅凭形式统计尚不足以说明译者风格何以在文学翻译中发挥作用。本书正是在这一争议背景下展开，其基本思路并非否定前者，而是试图在前者基础上补入“功能”维度。就材料组织而言，本书先在理论层面指出既有研究偏重形式、相对忽略功能的问题，再通过语料库语言学的理论模型，逐步把“功能性译者风格”转化为可操作、可比较的分析框架，最后再将文本差异放回译者背景、立场与社会文化语境之中加以说明。

本书最值得重视的理论推进，在于将译者风格由形式统计意义上的稳定特征进一步推进为与文本功能生成相关的语言选择。作者将译者风格界定为译者在目标文本中重构文学功能与建构文本功能时呈现的语言特异性，并据此提出主题风格、人物形象重塑风格和文本全景风格三种代表性类型。这一处理一方面延续了语料库翻译学强调经验材料和可验证性的传统，另一方面也吸收了语料库文体学关于“文本型式—文学功能”关联的研究成果，从而在“语言描写”与“文学阐释”之间建立了更紧密的联系。就学术史意义而言，这种尝试并不是另起炉灶，而是在既有“指纹”研究之外，为译者风格研究增添了一个更能进入文学翻译内部机制的维度。本书接受译者风格作为译者特异性语言选择的基本前提，并将研究重心置于同一作品的多译本比较。因此，它首先揭示的是不同译者在同一文本中的风格差异。这一研究设计具有很强的问题针对性，也使“功能性译者风格”这一概念获得了较为清晰的实证支撑。同时，这一概念也提示后续研究仍可继续拓展。如果将考察对象由单一作品的多译本比较进一步延伸至同一译者跨文本、跨体裁、跨时期的翻译实践，便有可能更充分地回应 Saldanha (2011) 所强调的、具有跨文本稳定性的译者风格问题。换言之，本书在概念建构上迈出了重要一步，其更广泛的适用层级和解释范围，尚有待今后研究继续充实。

### 3.2 文学意义的再建构

文学翻译的实质是风格的翻译。就文学翻译而言，译者风格的功能性意义，并不仅体现为某些语言特征的可识别差异，更体现为这些差异如何参与原文风格的重构，并影响人物形象的生成、主题意涵的凸显以及叙事氛围的重组。换言之，风格差异若要真正落实为意义差异，尚需进一步说明局部语言选择如何在篇章推进和组织中累积为整体性的文本效果。从这一角度看，意义的再建构之所以构成一个值得重视的问题，并不只是因为文学文本比一般文本更复杂，而在于文学翻译中的语言选择本身就具有高度的功能负载性。若仅停留于语言特征的识别，固然可以说明不同译者“写得不一樣”，但只有进一步追问这些差异如何进入功能的生成，译者风格研究才真正触及文学翻译内部的意义生产机制。

本书的启发意义，正在于它较为自觉地推动这一研究重心的转移。作者并未将主题词、语义场、词汇链、人物标签和话语表达等指标仅作为统计结果加以陈列，而是努力将其纳入文学和文本功能的分析框架之中，从而使语料库方法不再只是服务于差异的识别，而开始进入意义阐释的层面。尤其是在多译本比较中，不同译者的语言选择并非彼此孤立的局部现象，而可能在持续累积中塑造出不同的人物面貌、主题重心和整体文体图景。就此而言，本书的重要推进或许正在于说明翻译如何通过一系列可观察的语言选择，参与文学意义的重新建构，为译者风格研究由语言描写走向意义阐释打开了更具建设性的讨论空间。

### 3.3 研究方法的创新

本书的另一个贡献，在于较有意识地整合了语料库文体学与风格计量方法。前者使主题词、语义场、词汇链和话语表达等语言特征能够与主题建构、人物塑造等文学功能发生关联，后者则通过多维统计手段等技术描绘译文的整体文体轮廓，从而将“文学功能”与“文本功能”纳入同一研究框架。与以往较多依赖单一形式指标的研究相比，这种多层次、多路径的设计无疑增强了译者风格研究的解释力和立体性，也为语料库翻译学中“多方法互证”的研究思路提供了具有示范意义的范例。

在此基础上，本书具体操作的部分分析路径仍提示出方法进一步细化的空间。例如，在主题风格的讨论中，作者主要从少数“功能负载主题词”出发，进入语义场和词汇链考察，这一设计具有较强的操作性，也有效实现了从局部语言现象进入主题功能分析的过渡。然而，此类文学功能的生成往往还涉及叙事结构、修辞组织以及更大篇章单位之间的互动关系，因此，若在主题词网络之外进一步吸纳篇章层面和叙事层面的证据，功能建构的解释图景或许会更加完整。因此，对于更复杂、更高层级的文学功能生成过程，尚有继续扩展的可能。

总体来看，本书是近年来语料库译者风格研究中一项有价值的探索。最为重要的是，本书并未将“功能性译者风格”处理为一个封闭性的结论，而是将其呈现为一个仍可继续检验、拓展和深化的研究方向。就此而言，它不仅拓宽了译者风格研究的问题意识，也为今后进一步讨论翻译中的描写方式、文学功能与意义建构之间的关系提供了富有启发性的参照。

### 参考文献

- Baker M, 2000. Towards a methodology for investigating the style of a literary translator[J]. *Target*, 12 (2): 241-266.
- Biber D, 1988. *Variation across speech and writing*[M]. Cambridge: Cambridge University Press.
- Boase-Beier J, 2006. *Stylistic approaches to translation*[M]. Manchester: St Jerome Publishing.
- Malmkjær K, 2003. What happened to God and the angels: an exercise in translational stylistics[J]. *Target*, 15(1): 37-58.
- Saldanha G, 2011. Translator style: methodological considerations[J]. *The translator*, 17(1): 25-50.

### 通信地址

100089 北京市 北京外国语大学中国外语与教育研究中心

(责任编辑：李添豪)

# 《文体学：文本、认知与语料库》述评

上海电机学院外国语学院 王艳伟

**提 要：**《文体学：文本、认知与语料库》的两位编者基于对认知科学和语料库方法的深刻理解，围绕文体学热点话题及其理论框架的讨论，为读者提供了一整套理论性和实践性兼具的文体分析工具包。它不仅夯实了当代文体研究的理论基础，还为文学文本和非文学文本的文体分析提供了成熟的分析模型和可操作性强的语料库描写工具。在理论框架的整合、研究方法的综合和内容编写的体例上，本书都有所创新，在我国国内既可作为学术专著，供从事文体学、文学、语料库语言学、认知语言学等相关领域研究的学者研读，也可作为教材，供文体学相关课程的师生参考和选用。

**关键词：**文体研究；文学文本；认知科学；语料库方法

## *A Review of Stylistics: Text, Cognition and Corpora*

Yanwei Wang

**Abstract:** Based on a deep understanding of cognitive science and corpus methods, the co-authors of *Stylistics: Text, Cognition and Corpora* provide readers with a set of theoretical and practical stylistic analysis kits while discussing hot topics in stylistics and their theoretical frameworks. The coursebook not only consolidates the theoretical basis of contemporary stylistic research, but also provides a mature analytical model and a viable corpus description tool for both literary and non-literary texts. The innovation of the book lies in the integration of theoretical frameworks, the synthesis of research methods and the style of content compilation. In China it can be used as an academic reference for scholars engaged in stylistics, literature, corpus linguistics, cognitive linguistics and other linguistic fields as well as a textbook for teachers and students of stylistics.

**Keywords:** stylistic research; literary texts; cognitive science; corpus methods

### 1 引言

文本一直是文体学研究中的核心要素，经过近三四十年的发展，文体学的研究

方法已经从传统上严谨、详细的文本分析逐渐转向对语言产生和接收背后的认知结构和过程的理论性考量 (Semino et al., 2002), 以及使用语料库技术为文体研究提供经验数据、增强客观性的方法 (Semino et al., 2004), 其中前者被称为“认知转向”, 后者被称为“语料库转向” (Leech et al., 2007: 286)。值得注意的是, 作为这两种转向中的重要学者, Elena Semino 既是认知学者, 又是语料库学者, 这似乎暗示了认知科学与语料库研究的某种兼容性。认知转向、语料库转向与文本之间分别存在怎样的相互关系? 在文体分析中三者能否融合? 怎样融合? 这些问题在 Jane Lugea 和 Brian Walker 合著的教材《文体学：文本、认知与语料库》中逐一得到了解答。

本书的目标是帮助读者识别和解释对文本风格做出贡献的语言特征。在本书中, 两位编者提出文本、认知与语料库是当代文体研究的三大支柱, 三者之间是互补关系, 共同指向文本意义的建构与阐释, 将三者有机整合在一起的综合方法可以概括当代文体学的研究目标——以证据为基础分析文本如何在读者中产生影响。

## 2 内容简介

本教材正文由体量均衡的 10 章构成。第 1 章首先引入风格 (或文体) 的概念, 语言风格是语言行为, 是人们运用语言做事的一种特定方式, 是由文本生产者有意识或潜意识地在语言系统中所做的语言选择的总和, 这些选择显现在文本中因而可为读者识别 (Lugea & Walker, 2023: 2)。文体学研究文本的风格, 考察文本生产者在诸多可选形式中所做的语言选择以及这些选择如何与文本的艺术功能 (Leech et al., 2007: 12)、诗学功能 (Jakobson, 1960: 356)、文体功能 (Carter et al., 1989: 5) 联系在一起。更重要的是, 文体学关注文本的语言描写, 并在描写基础上对取得的效果进行解释。故编者提出文本、认知和语料库可以结合在一起, 以促进文本的语言描写和所取得效果的解释, 帮助读者了解文本风格的创造过程。

编者提出, 文体学的一个良好开端是基于这样一种理念: 文本创造了一个“世界”, 即话语的心理表征。该隐喻有助于厘清文本与话语语境之间的关系, 也强调了读者的认知在创造文本理解中的作用。因此, 第 2 章详细总结了基于“世界”的文体学研究方法, 提供了话语建构过程的一个概览。根据文本世界理论 (Text World Theory, 简称 TWT), 话语参与者结合文本线索和先前知识创造了一个话语的心理表征, 即文本“世界”。接着编者运用 TWT 理论, 以《德伯家的苔丝》为例, 通过其起始两段对文本世界的建构和阐述, 揭示了这部小说从一开始就关注社会的不平等和僵化问题。鉴于 TWT 基于文本细读, 对于分析较长的文本而言可行性较低, 编者建议使用语料库来找到文本的“切入口” (way in), 以揭示小说的语言特征, 进而聚焦部分特征进行定性分析。通过基于词性赋码的语料库词表文内比

较，编者选择了小说中情态助动词使用频率最高的第36章进行了更详细的文本世界分析。语料库方法与TWT创新性的结合阐释了小说中新婚人物对情态世界的创造如何影响他们的婚姻讨论和决策。

第3章旨在阐明散文小说不仅是一个故事（story），而且是对该故事的讲述（narrative），一种蕴含视角的叙事（narration）。该章综述了大量关于叙述视角的研究，同时提供了一种基于话语结构思考叙事、叙述与故事之间关系的新方式。编者描述了在话语世界中，应当如何通过叙事文本与读者交流。随后以海明威的《太阳照常升起》为例，编者探讨情态与评价性语言在叙述视角构建中的作用。海明威以其客观的写作风格闻名于世，这种客观的风格似乎缺乏叙事者的主观性和评价。本书编者用语料库方法来验证人们普遍持有的这个假设。主题词性（key-POS）分析表明，该作品中叙述者使用的评价性副词和形容词以及情态助动词相对较少。通过研读索引行，编者揭示出一些视角标志项（“let’s” “ought to” 和疑问词）在作品中被过度使用，但这些标志项主要出现在人物的直接引语中，而不是叙事本身。这一发现为人们对这部小说简洁、新闻式风格的直觉性评价提供了实证支持。

从TWT视角来看，话语呈现（discourse presentation，简称DP）是创造“世界”的信息。故第4章承接第3章，基于Leech等（2007）提出的原创性理论模型，描写小说中话语呈现的不同形式，包括言语、书写和思想（speech, writing and thought presentation，简称SW & TP），通过探讨现有的呈现人物言语、书写和思想的不同技术，着重讨论了叙述与人物话语之间的重要区别。编者利用英国19世纪小说语料库（Huddersfield, Utrecht, Middelburg Corpus of UK 19th century Fiction，简称HUM19CUK）中的例子来说明SW & TP的不同范畴，并展示了如何利用语料库方法来分析SW & TP。首先，按照SW & TP模型，手工标注伊夫林·沃（Evelyn Waugh）的短篇小说《勒夫戴先生的短暂外出》（*Mr. Loveday's Little Outing*）中的SW & TP范畴。其次，将其与兰开斯特SW & TP语料库中的小说部分（LancFic）进行比较。这种量化比较能够突显目标文本与其所属文类常态（norm）之间的种种差异，然后可用通配符搜索（wildcard search）对这些差异进行深入的质性分析，这种文体学框架与语料库方法的结合使原本极其困难的文本分析成为可能。

第5章结合语用学和会话分析的模型来分析戏剧中的对白，集中探讨了英国2016年出品的电影《我，丹尼尔·布莱克》（*I, Daniel Blake*）的剧本节选，讨论涵盖诸多语用学理论，阐明如何将这些互补的理论模型组合起来帮助理解人物之间以及作者和读者之间互动的重要性。在分析中，编者尝试运用语料库方法提取N元组合（n-grams），并结合相关索引行来展示人物对话语言之间的比较方式，以揭示主要人物语言风格特征中的显著模式，并运用礼貌原则这一语用框架进一步分析和解读这种语言模式。这种研究情境喜剧对白的语料库方法可以解释数据中的每一个样例，赋予了文体分析应有的系统性和严谨性。

在第6章中，编者秉持小说人物是一个文本和认知问题的观点。编者介绍了人物和人物刻画（characterization）的概念，并在前人研究基础上，提出了一个分析小说人物的清单；拓展了 Culpeper（2001）基于小说文本提出的分析模型的应用范围，通过对多篇散文小说中人物刻画的详细分析证明该模型也可用于分析散文和戏剧中的人物。此外，编者也具体展示了如何使用语料库方法中的主题词文内比较来分析萧伯纳的作品《卖花女》中的人物。作品中的主题词为后续分析提供了明确的聚焦点，而这一分析依托人物刻画的认知模型展开，从而凸显了贯穿全书的当代文体学三大支柱——文本、认知与语料库的持续重要性。

在第7章中，编者表明小说中修辞性语言的分析不仅依赖文本和认知方法，还可以通过语料库方法来进一步增强。编者在讨论隐喻、转喻和明喻时借用了认知语法中百科知识的概念来解释与隐喻和明喻相关的跨域映射以及转喻相关的域内映射。编者借鉴并结合了多种前人的研究方法，包括 Lakoff 等（1989）颇具影响力的概念隐喻理论。编者最后讨论了可以帮助识别始源域的语料库方法，包括使用自动赋码系统实现语义赋码。通过对语言选择过程和认知解释过程的探讨，这些分析工具有助于揭示读者理解文本中隐喻的过程。

第8章总结了50年来的思维风格相关研究，将思维风格定义为表明人物特定思维方式的语言特征的累积。编者描写了传统上与思维风格联系在一起的词汇语法特征以及小说思维风格研究中的语用学和认知方法。本章最后创新性地将语料库方法与认知语法融合在一起：通过主题词文外比较识别出福克纳的作品《喧哗与骚动》中班吉的多用词和少用词，并结合现有的原创质性研究，重新审视班吉的思维风格。这种运用语料库方法来重新审视文本的既定理解阐明了本书自始至终鼓励的做法——文体学分析要致力于方法论上的可检索性、严谨性和可复制性。

对文体研究者而言，本书第9章可视为对言语幽默的首次研究综述，内容覆盖广泛。编者首先解释了歧义、不一致性及其消解何以成为幽默话语的基本特征，接着介绍了言语幽默普遍理论（the general theory of verbal humour, 简称 GTVH）。重要的是，通过重组或修订，编者阐明了 GTVH 与文体分析的相关性，并将其整合到了文体研究范式中。本章还结合前面章节中介绍的概念和模型，表明如何用它们来研究幽默话语，同时也阐明了幽默在打造叙事声音、塑造人物以及理解人物关系上具有的特殊作用。在章末编者概述了一些使用语料库方法研究幽默的可能路径，同时也阐释了如何运用语料库方法来检验研究者关于正常语言用法的直觉，以证实关于不一致性或意外语言使用的断言。

在最后一章中，编者将前面章节中介绍的工具进行打包，讨论如何将该工具包用于文本分析中，可谓提供了一个完成文体研究项目的有用指南，使文体研究工作进一步明确。值得注意的是，编者介绍了证伪的概念，并认为文体研究中的任何工作都应该是可证伪的，即要以清晰的假设或可检验的研究问题为中心。此外，编者

以焦点框的形式概括了文体研究中应避免的陷阱。最后一部分则是对文体研究中关键原则的概述。

### 3 评价

本教材两位编者基于对认知科学和语料库方法的深刻理解，围绕文体学热点话题及其理论框架的讨论，为读者提供了一整套理论性和实践性兼具的文体分析工具包。它不仅夯实了当代文体研究的理论基础，还为文学文本和非文学文本的文体分析提供了成熟的分析模型和可操作性强的语料库描写工具。以下笔者将从理论整合创新、研究方法创新和编写体例创新三个层面评价本书的特色和价值。

本教材的第一大特色是编者在创新性整合上的理论自信。与新弗斯学派中反对语料库语言学和认知语言学结合的学者不同，Jane Lugea 和 Brian Walker 基于对认知科学和语料库方法的深刻理解，从兰卡斯特学派的 Leech、Short、Culpeper、Semino 等学者处汲取了不少灵感，有机整合了语言学各个领域的精髓，创新性地提出：文本、认知与语料库三者之间是互补关系，构成了当代文体研究的三大支柱，共同指向文本意义的建构与阐释。一方面，编者把文本世界理论作为文体研究的开端，有助于厘清文本与其话语语境之间的关系，也强调了读者的认知在创造文本理解中的作用，既促进了文本的语言描写及其取得效果的解释，也有助于了解文本风格的创造过程。另一方面，编者对幽默研究首次进行了综述，尤其是通过重组或修订将 GTVH 整合到了文体研究范式中，将 TWT 运用到笑话分析中，也把 TWT、共文（co-text）及相关语境融入幽默话语的加工过程。

本书的第二个特色在于语料库方法在文体分析中的创新性应用。书中几乎每一章都有语料库方法的应用介绍，并在最后一章概括了语料库用于文体研究的两种主要方式。其一，现有的大型语料库可以作为参照库，为读者对文本中的语言选择可能产生的语言直觉提供经验证据。其二，语料库工具可以直接对被调查文本进行多种方式的分析。语料库工具可赋能文本组、单个文本或文本部分间的定量语言比较，帮助读者识别作者、叙述者和人物风格的差异。即便在分析需要人工干预的语用功能方面，对语料库进行标注也有助于记录人工分析的结果，使其更容易追踪和量化，从而使文体分析更具系统性和严谨性，展示出一种综合性方法所提供的广泛适用性。

本书的第三个特色在于编写体例上的创新。作为一本文体学教材，本书面临的一大挑战就是通俗、清晰、简明地陈述核心概念。本书精心选择文体学核心研究议题，每一章都以一个简短的理论概述开始，然后是说明性的文本分析及语料库方法的应用，最后进行总结。令人耳目一新的是，书中提供了丰富多样的契合特定主题的活动、呈现更多细节内容的焦点框及语料库索引行。此外，书中每一章都有电子

补充材料，包括书中活动题的答案和以语料库索引行形式呈现的简短的任务说明，读者可从本书在Palgrave和SpringerLink网站上的网页进行下载。可以说，两位编者出色地实现了教材编写的目标，并受到了相关专家的高度肯定。

本书尽管特色纷呈，但也有少许瑕疵。作为一本文体学教材，本书主要关注文学体裁（散文、小说、诗歌、戏剧）的风格研究，而对更具实用性的非文学体裁（如新闻报道、平面广告等）没有涉及。书中关于语料库方法在文体研究中的应用往往局限于文本的表层形式，缺乏对内容的深度探讨，对于熟悉语料库研究的读者而言可能仍显不足。此外，书中第53页在讨论情态动词使用标准差时，未对重要数据14.37的来源及其与上文每千词平均使用频率14.64之间的关系作出说明，同时也存在极个别的打印错误。

总体上看，瑕不掩瑜。本书在理论框架的整合、研究方法的综合和内容编写的体例上都有所创新，可圈可点，在我国国内既可作为学术专著，供从事文体学、文学、语料库语言学、认知语言学等相关领域研究的学者仔细研读，也可作为教材，供文体学相关课程的师生参考和选用。

## 参考文献

- Carter R, Simpson P, 1989. Introduction[M]//Carter R R, Simpson P P. Language, discourse and literature: an introductory reader in discourse stylistics. London: Routledge: 1-21.
- Culpeper J, 2001. Language and characterisation: people in plays and other texts[M]. Harlow: Longman.
- Jakobson R, 1960. Closing statement: linguistics and poetics[M]//Sebeok T A. Style in language. Cambridge: MIT Press: 350-377.
- Lakoff G, Turner M, 1989. More than cool reason: a field guide to poetic metaphor[M]. Chicago: University of Chicago Press.
- Leech G, Short M M, 2007. Style in fiction: a linguistic introduction to English fictional prose[M]. Harlow: Longman.
- Lugea J, Walker B, 2023. Stylistics: text, cognition and corpora[M]. London: Palgrave Macmillan.
- Semino E, Culpeper J, 2002. Cognitive stylistics: language and cognition in text analysis[M]. Amsterdam: John Benjamins.
- Semino E, Short M M, 2004. Corpus stylistics: speech, writing and thought presentation in a corpus of English writing[M]. London: Routledge.
- Turner M, 1991. Reading minds: the study of English in the age of cognitive science[M]. Princeton: Princeton University Press.

## 通信地址

201306 上海市 上海电机学院外国语学院

（责任编辑：刘若冰）

# 语料库语言学

## CORPUS LINGUISTICS

《语料库语言学》由教育部人文社会科学重点研究基地“中国外语与教育研究中心”创办，于2014年首次出版，是亚洲地区语料库语言学领域较早发行的专业学术出版物，聚焦基于大规模语言数据的理论探索、方法创新与跨学科应用。长期关注该领域的新动态、新议题、新方法，欢迎基于语料库和生成式人工智能技术的词汇语法、话语语用、翻译研究、语言教学、社会语言学及认知语言学等创新研究，鼓励交叉学科方法，以学术性、前沿性和国际性为宗旨。

外研社·期刊出版分社

电话：010-88819267

E-mail: qkzx@fltp.com

网址: www.bfsujournals.com



记载人类文明  
沟通世界文化  
www.fltp.com



北外学术期刊



HSS Online平台

责任编辑：赵 雪  
责任校对：王雨舟 白小羽  
封面设计：锋尚设计



定价：60.00元