

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

第21辑  
二〇二四

# 语料库语言学

CORPUS LINGUISTICS

第21辑

2024

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  
许家金 主编

语  
料  
库  
语  
言  
学

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 concordance frequency ToRCH  
iWriteBaby  
corpus-as-theory ParaConc phraseology

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

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 and Corpus Linguistics

Society of China

Published by Foreign Language Teaching and Research Press

刊名题字：崔希亮

主编：许家金

责任校对：宋瑛明 杨宇航 殷伶俐

**Journal Name Calligraphy:** Cui Xiliang

**Editor:** Xu Jiajin

**Proofreaders:** Song Yingming, Yang Yuhang,  
Yin Likai

编审委员会（按姓氏音序）

主任：

梁茂成（北京航空航天大学）

**Editorial Board** (in alphabetical order)

Chair:

Liang Maocheng (Beihang 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) 88816828

电子邮箱：bfsucrg@sina.com

投稿网址：<http://ylly.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

2024 年 第 21 辑

许家金 主编

外语教学与研究出版社

FOREIGN LANGUAGE TEACHING AND RESEARCH PRESS

北京 BEIJING

《语料库语言学》

2024 年 第 11 卷 第 1 期

目 录

冲突话语

中美媒体新冠病毒溯源话语的对抗性特征与话语策略分析..... 卫乃兴 金碧希（ 1 ）  
言语行为理论视角下伊拉克战争中的冲突性话语分析 .....李新月 马晓雷 李德坤（ 15 ）

研究论文

大语言模型辅助下基于词典的义项表生成及应用研究..... 葛晓帅 王 梦（ 26 ）  
交替现象实证研究现状与展望 ..... 王梦岚 江桂英（ 38 ）  
基于词频分布指标的莎士比亚十四行诗汉译多维标度和  
聚类分析研究 ..... 陈国兴 徐德宽（ 49 ）  
评价理论视阈下黄河流域媒体形象的构建与传播对策研究..... 王海峰 李文宇（ 68 ）  
基于语义韵强度算法的中国形象建构研究 ..... 汤慧桃 王 军（ 87 ）  
基于语料库印度媒体“人类命运共同体”新闻话语命名策略研究..... 李梦翔（ 103 ）

研究开发

frGLOBE 当代法语书面语平衡语料库的创建..... 殷俪恺（ 113 ）  
人教社中小学教材语料库（PEPTC）的研制 .....黄攀伟 胡 源 杨亚欣 张博翎（ 120 ）

书刊评介

《历时构式语法的节点与网络》述评 ..... 房印杰（ 135 ）  
《语料库设计与评估：针对语料库代表性的实用性框架》述评..... 张琮然 徐海铭（ 141 ）  
主要文章英文摘要 ..... （ 146 ）



# Corpus Linguistics

Volume 11, Number 1, 2024

## Table of Contents

**Featured column: Conflict discourse**

The analysis of the confrontational features and discursive strategies in Chinese and American media’s COVID-19 virus origin tracing discourse ..... *WEI Naixing & JIN Bixi* (1)

An analysis of conflict discourse in the Iraq War from the perspective of speech act theory ..... *LI Xinyue, MA Xiaolei & LI Dekun* (15)

**Research articles**

Research on the generation and application of sense lists assisted by large language models ..... *GE Xiaoshuai & WANG Meng* (26)

A review of the empirical alternation studies ..... *WANG Menglan & JIANG Guiying* (38)

Multidimensional scaling and cluster analysis of Shakespeare’s sonnets and their Chinese translations using word frequency indices ..... *CHEN Guoxing & XU Dekuan* (49)

The construction and dissemination strategies of the Yellow River Basin’s media image from the perspective of appraisal theory ..... *WANG Haifeng & LI Wenyu* (68)

A study of the construction of China’s image through prosodic valence analysis ..... *TANG Huitao & WANG Jun* (87)

A corpus-based study on nomination strategy of “a community with a shared future for mankind” in Indian media coverage ..... *LI Mengxiang* (103)

**New corpora, tools and methods**

Creation of the frGLOBE contemporary written French balanced corpus ..... *YIN Likai* (113)

Development of the People’s Education Press primary and secondary school textbook Corpus ..... *HUANG Panwei, HU Yuan, YANG Yaxin, ZHANG Boling* (120)

**Book reviews**

L. Sommerer & E. Smirnova (eds.). *Nodes and Networks in Diachronic Construction Grammar* ..... *FANG Yinjie* (135)

J. Egbert, D. Biber & B. Gray (eds.). *Designing and Evaluating Language Corpora* ..... *ZHANG Congran & XU Haiming* (141)

**English abstracts of major articles**..... (146)

# 中美媒体新冠病毒溯源话语的对抗性特征与话语策略分析\*

北京航空航天大学 卫乃兴 金碧希

**提要：**新冠肺炎疫情期间，关于病毒溯源的话语俨然成为国际媒体的重要话语事件。本研究基于自建语料库，探讨中美媒体的病毒溯源话语表征。研究选取2020年1月到2021年12月中美有关新闻媒体的相关报道，运用语料库分析技术，开展相关关键词的词语索引分析、搭配分析、语义趋向及语义韵分析，探索双方的话语冲突、博弈和较量特征。数据分析显示，两国媒体在新冠肺炎疫情溯源问题上采用了一系列攻防策略，呈现出3组截然不同的语义趋向和语义韵，构建出对抗性特征极强的话语博弈图景，揭示位于底层的国家立场和意识形态驱动。本文对对抗性话语的重要特征和话语策略进行了系统讨论，对话语分析具有概念建构与方法论层面的启示。

**关键词：**病毒溯源、对抗性话语、语义趋向、语义韵

## 1 引言

2020年爆发的新冠病毒席卷全球，给人类生活和全球政治、经济造成重大影响。新冠肺炎疫情话语，尤其是中美两国关于病毒溯源的话语博弈俨然成为21世纪初叶重要的话语事件。多国政府、政要、公众人物、专家学者对病毒溯源高度关注，各持己见。中美两国重要媒体也对此展开了针锋相对的博弈、攻防和激烈竞争，显示出强烈与明显的对抗性话语特征，形成了这一时期两国媒体话语交互的重要图景。基于真实语料探索中美媒体新冠肺炎疫情溯源话语表征，分析双方在相关问题上的激烈话语博弈以及针锋相对攻防策略，是话语研究不能忽视的内容。认识和理解这一特定话语的本质特征、其特有的话语运作方式、互动机理以及策略趋向对推进话语研究相关问题的探究具有现实意义。在语料库语言学的方法框架下探索这一特定话语的对抗性特征及其策略运作，有望获得在急剧变化的世界格局下有关话语竞争

\* 本文系国家社科基金项目“基于动态语料库的重大突发事件对外话语策略研究”（20BYY081）的阶段性成果。卫乃兴为本文通信作者。

作者贡献：

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

金碧希：研究方法、数据分析、初稿撰写、字数占比（40%）。

与冲突的新颖发现与见解,有助于加深理解话语攻防在国际竞争中的重要作用,并对我国的对外话语传播提供启示。

新冠肺炎疫情相关话语已引起研究者的广泛关注。*International Journal of Corpus Linguistics* 于 2021 年刊出了新冠肺炎疫情话语专刊 *Language and Covid-19*, 内容涉及新冠病毒科研论文中的渲染表达 (Hyland & Jiang 2021), 学术博客中的作者立场构建 (Curry & Pérez-Paredes 2021) 以及网络悼念话语 (McGlashan 2021) 等。然而,目前聚焦疫情溯源话语的研究仍不多见。方法论上,有学者从传播学视角出发,采用内容分析法,分析关涉溯源议题的叙事策略和叙事技巧 (张康 2022),或探讨中外媒体在报道声量、报道信源和报道框架方面的异同 (王登丰等 2022)。这些研究都在特定方面探讨了疫情话语的重要特征,但仍缺乏大量真实数据支撑的话语意义剖析。

基于语料库的话语研究已成为国际话语分析领域独具方法特色的一个重要路径。尽管语料库方法对话语探究仍存在局限,但其强大的实证数据基础以及严密的科学分析程序所带来的数据整体解释效应、话语运作图景全貌、话语建构中的型式共选机制,以及命题背后隐含态度意义的发掘,都值得进一步探索。方法论上,这一路径 (1) 严格依赖真实的文本证据; (2) 基于复现或统计学意义上的显著证据; (3) 融合词汇与语法的描写; (4) 聚焦共选型式的意义建构; (5) 缜密检查共延共文环境与社会语境下的隐含意义。在分析技术上,词语索引分析、搭配数据分析、构型分析、语义趋向和语义韵分析等成熟确立的方法依然有着极大的使用价值,可丰富经典话语分析的内涵及方法,与后者形成有意义的互补。

基于语料库的话语研究是语料库语言学的有机组成部分。它采用语料库语言学的核心理念和基本方法,高度倚重语料库的“整体解释”和“增量效应”(Baker 2006: 13)探索话语建构。Hardt-Mautner (1995) 首次提出采用语料库方法分析话语 (Baker & McEnery 2015: 7)。Teubert (2001) 通过关键词分析,对比分析耻辱词 (stigma words) 和旗帜词 (banner words) 的型式与话语意义特征,探究英国公共话语中“欧洲怀疑论”的话语表征。Baker *et al.* (2008) 基于搭配分析,探索英国媒体对移民、避难者和政治庇护寻求者的话语建构,揭示这些话语对象“众多”“失控”“无序”和“负担”等话语表征。Tang (2023) 运用话语-历史分析方法 (Wodak 2001), 分析中国驻英大使在英国主流媒体中就中美贸易摩擦展开反美国霸权的话语,建构中国负责任大国形象。

但整体而言,基于语料库的话语研究虽在国内学界行之有年,系统、严谨采用语料库方法,基于坚实证据进行缜密的话语建构分析仍有较大完善空间。

本研究探讨中美媒体的不同疫情溯源话语建构特征。具体而言,我们采用前述语料库语言学的分析框架与分析技术,尤其注重语义韵和语义趋向的对比分析,系统剖析双方话语的对抗性特征,解构双方话语博弈所用策略,挖掘话语建构背后的

国家立场与意识形态分歧,并尝试概括这类“对抗性话语”的主要特征及研究意义。整个研究严格依循基于语料库的方法和程序,以期在新的方法与视角下对新冠肺炎疫情话语,尤其是对抗性话语产生新的认识与理解。

## 2 研究方法

### 2.1 新冠病毒溯源新闻报道语料库

为进行本文研究,我们专门建立了中美媒体新冠病毒溯源新闻报道语料库。该库下含两个子库,分别为美国媒体库和中国媒体库。我们利用北京航空航天大学国际媒体话语分析平台,以标题中含有 COVID-19、coronavirus、virus、origin、origin-tracing 和 origin investigation 等词语为核心搜索词,进行文本抓取,采集自 2020 年 1 月到 2021 年 12 月的有关新闻报道。中国媒体的新闻报道选自 China Daily、Global Times 和 Xinhua News Agency<sup>3</sup> 家英文媒体,共计 382 篇文本。美国媒体新闻报道则采自 CNN、Fox News、Newsweek 和 The New York Times<sup>4</sup> 家媒体,共计 1,149 篇文本。相对而言,所选媒体受众广泛,且在相当程度上能够反映各自政府立场以及主流社会舆论。表 1 显示 COVID-19 语料库的基本信息。

表 1 COVID-19 语料库基本信息

数据	美国媒体库	中国媒体库
文本数(篇)	1,149	382
形符数	919,795	311,815
类符数	26,859	13,324
平均文本长度	800.45	816.27
类符形符比	2.92%	4.27%
标准类符形符比	42.67%	41.79%

### 2.2 分析技术与过程特征

#### 2.2.1 词语搭配分析

词语搭配是最重要的意义方式(Firth 1957),是语言共选观和短语理论的主要体现,基于共现数据分析进行(卫乃兴 2011)。搭配是两个或两个以上的词形在一定跨距内的频繁共现。基于语料库的搭配计算分为基于搭配强度和基于搭配典型性两类测量方法。前者测量的是两个共现词形的相互吸引强度,后者测量的是两个词形组合的典型性。两个方法意旨不同,各有特点,输出结果不同。本研究综合使用

Sketch Engine 平台提供的 T-score 和 MI3, 参照 Hunston (2002)、Salama (2011) 和唐丽萍 (2016) 的做法, 采用  $MI3 \geq 3$  且  $T \geq 2$  的阈值标准, 在 Sketch Engine 平台上提取两个语料库中 virus 的前 20 个最显著搭配词进行分析。

### 2.2.2 词语索引分析

词语索引是除却原始文本之外最直接和便捷的语言使用证据, 也是语料库型式与意义研究不可替代的基础性分析技术。显著搭配词提取后, 我们对之进行词语索引观察, 概括其型式与意义特征。基于节点词 virus 的大量词语索引证据, 我们检查局部语境中的共现词语, 认定共选型式, 剖析话语建构与意义表征 (Stubbs 2001: 169)。

### 2.2.3 语义趋向与语义韵分析

语义趋向揭示的是节点词与意义的共选, 体现为“搭配词的语义特征, 或所属于的语义组” (卫乃兴 2012: 4)。语义趋向的分析概括过程即话语表征的解构过程 (Stubbs 2001; Hunston 2007)。语义韵是节点词在更大语境内由包括搭配和语义趋向在内的多种共选因素共同构筑的态度意义 (李晓红、卫乃兴 2012: 22)。语义韵揭示说话者的交际意图和态度。早期的语义韵研究一般采用三分法, 将其分为积极、消极和中性 3 个类别。本研究采用 Sinclair (1996, 2004) 的细微粒度方法, 对语义韵进行适度抽象和概括, 有助于挖掘话语参与者针对特定命题或事件的态度及其背后的立场与意识形态意义。

## 3 数据结果与分析

### 3.1 搭配数据分析

我们以 virus 为节点词, 设置搭配跨距为 5 (即 L5-R5), 进行词语搭配数据处理。采用  $MI3 \geq 3$  且  $T \geq 2$  的显著性标准, 我们提取了所建语料库中与 virus 构成显著搭配的前 20 个词形。需要说明的是, 在所有语料库研究中, 研究者都面临庞大的数据量, 也都必须聚焦分析一个可操作的数据集。根据本文的设定目的, 该语料库中前 20 个最为显著的搭配词应是首先探索的数据对象, 具体搭配词数据如表 2 所示。

表 2 中国媒体库中 Virus 的前 20 个显著搭配词

序列	美国媒体库 搭配词	频数	T-score	MI3	中国媒体库 搭配词	频数	T-score	MI3
1	originated	137	11.67	22.71	originated	27	5.18	17.90
2	escaped	60	7.72	20.21	leaked	26	5.08	17.84
3	Chinese	48	6.76	16.50	evidence	25	4.92	15.34

(待续)

(续表)

序列	美国媒体库 搭配词	频数	T-score	MI3	中国媒体库 搭配词	频数	T-score	MI3
4	Wuhan	48	6.71	16.18	deadly	22	4.68	18.39
5	leaked	44	6.61	18.89	came	20	4.44	15.90
6	dangerous	44	4.66	16.08	probe	19	4.33	15.55
7	deadly	43	6.54	19.77	new	16	3.92	13.70
8	killer	11	3.31	16.82	dangerous	15	3.87	16.85
9	variants	9	2.98	13.79	research	13	3.47	12.17
10	spread	44	6.53	16.98	global	11	3.14	11.18
11	came	18	4.12	13.48	joint	11	3.20	11.72
12	President	14	3.57	12.07	natural	10	3.14	13.90
13	lethal	13	3.60	16.39	study	10	3.03	11.26
14	Fauci	10	2.76	9.60	scientific	8	2.68	10.27
15	mutated	9	2.99	14.60	bat	8	2.81	13.25
16	contagious	10	3.15	14.43	unknown	8	2.81	13.17
17	Trump	8	2.55	9.36	animal	7	2.61	11.68
18	community	7	2.45	9.35	escaped	7	2.63	13.52
19	officials	7	2.36	8.82	fight	6	2.36	9.88
20	Pompeo	6	2.43	12.04	international	6	2.17	8.28

表 2 显示,两库共享一部分显著搭配词,如 originated、escaped、leaked 和 came。这反映了所建语料库适配设定的研究目的,而 Chinese 和 Wuhan 两个地理名称在美国媒体库中成为显著搭配词,揭示出美方话语对于病毒溯源的目标地域指向。两库均使用了一系列名词和形容词,描述病毒属性,如 deadly 和 dangerous 等。根据这些特征,我们将显著搭配词分为不同的语义组。

### 3.2 搭配词语义组

表 2 的显著搭配词可分为 3 个语义组,如表 3 所示。这三个语义组分别代表两个子库文本在溯源报道中关注的 3 个方面话题:(1)病毒来源,即对于病毒来源地的表述,主要通过 virus originated、leaked、escaped 等搭配词体现。(2)病毒属性,主要通过一系列名词和形容词修饰语对病毒属性描述。(3)溯源调查,这既包括溯源行动者或机构,如 Pompeo、Trump、Fauci 和 intelligence community 等,也包括溯源行为,如 fight、probe、evidence 和 research 等。



表 3 virus 显著搭配词语义分组

语义组	美国媒体库	中国媒体库
病毒来源	originated ( 137 ), escaped ( 60 ), leaked ( 44 ), came ( 18 ), Chinese ( 48 ), Wuhan ( 48 )	originated ( 27 ), escaped ( 6 ), leaked ( 26 ), came ( 20 )
病毒属性	deadly ( 43 ), spread ( 44 ), dangerous ( 44 ), lethal ( 13 ), mutated ( 9 ), variants ( 9 ), killer ( 11 ), contagious ( 10 )	deadly ( 22 ), dangerous ( 15 ), new ( 10 ), natural ( 10 ), unknown ( 8 ), bat ( 7 ), animal ( 7 )
溯源调查	President ( 14 ), Fauci ( 10 ), Trump ( 8 ), community ( 7 ), Officials ( 7 ), Pompeo ( 6 )	evidence ( 25 ), probe ( 19 ), research ( 13 ), global ( 11 ), joint ( 11 ), scientific ( 8 ), study ( 10 ), fight ( 6 ), international ( 6 )

### 3.3 语义趋向和语义韵

我们对上述三组不同的搭配词展开语义趋向和语义韵分析。

#### 3.3.1 病毒来源

##### 3.3.1.1 美国媒体库的病毒溯源话语表征

观察词语索引<sup>1</sup>可见,在美国媒体库中,virus 和 originated、escaped、came 及 leaked 等词形频繁共现。数据显示,originated 为 virus 最高频的搭配词,共出现 137 次。virus originated 左侧通常和 believe、suggest、allege 等认识性动词,以及 possibility、assertion、hypothesis、evidence、belief、suggestion 等认识性名词共现。这些名词和动词表达对命题或事物的认识状态判断,揭示所述命题与客观世界的一致程度或对应程度,是言说者对两者对应关系的主观设定,反映其赋予命题的信度或承诺程度(Hunston 2008, 2010; 卫乃兴、王冰昕 2022)。virus 右侧大多后接 from a Chinese lab、from the Wuhan Institute 和 in a wet market 等介词短语。认识性名词的功用在于对后续命题赋予某种认识判断,如后续命题是一种可能性主张或假设。认识性动词的功用则是提出某种认识判断,表达对后续命题的立场。将所有共现的认识性词语的意义特征整体考虑,我们不妨概括出一个“认识可能性”的语义趋向。而将整个局部文本环境的词语共选因素综合一起可认为,virus originated 传递了一个“假定的武汉实验室泄露”的语义韵。

表 3 显示,这一组的显著搭配词还有 came、leaked、escaped 等动词。检查词语索引可见,这些词左右两侧的共现词语与 virus originated 的共现型式高度一致,即 virus 左侧表达对命题或事物的认识状态或判断,右侧则是对命题判断的断言,如 from a lab。由此,这些共现词语表达了和 virus originated 同样的语义趋向和语义韵。

由上述分析可见，美国媒体库的话语参与者显示极强的趋势将武汉实验室视为病毒之源，即病毒由其可能的泄露而传播。然而，这只是一种假定和推测，缺乏真实世界的证据支持，但这种假定和推测经多次重复传播，便有可能成为一种缺省的价值预判或定见。为此，我们检查了美国媒体库中由“premodifier + virus”构成的搭配，发现较多报道直接以 Chinese virus（48）和 Wuhan virus（48）来命名新冠病毒。根据 Wodak（2011），这是一种命名策略。这种命名策略经媒体反复使用，已经将预设内嵌于话语，对受众产生一种潜移默化的认知形塑效力。

再仔细检查 Chinese virus 的词语共现，可发现该指称多和 Trump、calling、tweet、referring to 等词语共现。相关的话语事件显示，美国前总统特朗普多次在推特上使用该称谓指称新冠病毒。同样，该指称缺乏真实世界的证据支持。据 2021 年 8 月美国政府情报部门发布的报告，没有证据支持新冠病毒源自武汉的论断<sup>2</sup>。然而，当一个国家的领导人或高级责任官员不顾证据而任性地将新冠病毒之源强加于对手时，话语博弈的对抗性意旨就非常明显和强烈了。

从话语策略的角度看，可以认为美国媒体库美国媒体库的病毒溯源话语对中方实施了他者化策略。根据 Fairclough & Wodak（1997）和 Tang（2023），该策略的重要实现方式是对话语对手的负面表征或贬抑。在美国疫情形势不断恶化和中美关系紧张的大背景下，这同时又是一种责任规避策略。Weaver（1986）和 Hood（2011）指出，责任规避的一个重要途径即“找替罪羊”，将本该自己担负的责任推卸给他人。

3.3.1.2 中国媒体库的病毒溯源表征

我们检查了 virus originated 在中国媒体库中的词语索引（见表 4）。virus 左侧出现了 lies（谎言）、claim（声称）、conspiracy（阴谋）、theory（理论）、political narrative（政治叙事）、presumption of guilt（有罪推定）、unlikely（不可能）等词语。这和美国媒体库美国媒体库高频使用的认识性动词和认识性名词形成强烈对比。这些词语的语义趋向可概括为“非事实性”。综合考虑 fabricating and spreading、it is highly unlikely、cast doubt on、narrative based on conspiracy、political narrative 等共现词语意义，virus originated 显然传递了“质疑”的语义韵，是对“实验室泄露论”的回击。

表 4 中国媒体库中 virus originated 的词语索引

fabricating and spreading lies that the	<b>virus originated</b>	from a Chinese lab. When spreading
that it is highly unlikely that the	<b>virus originated</b>	from laboratory leakage, the world,
stated that it was unlikely that the	<b>virus originated</b>	in a lab, Biden ordered his
cast doubt on a theory that the	<b>virus originated</b>	in a Chinese laboratory,” based on
narrative based on a conspiracy that “the	<b>virus originated</b>	in China and it is either

（待续）



(续表)

the presumption of guilt that “the	<b>virus originated</b>	in China and the Chinese lab
spreading the political narrative that the	<b>virus originated</b>	in China. Washington has in
because Wuhan is definitely not where	<b>virus originated</b>	May
the	<b>virus originated</b>	from,” Lu said. He explained
enormous evidence to support the claim	<b>virus originated</b>	that if
that	<b>virus originated</b>	from a biomedical laboratory in
Times of India are hyping up the theory	<b>virus originated</b>	from China.” Swarajya
that the “		published an

同样，中国媒体库中的显著搭配词 *came*、*leaked*、*escaped* 也具有相同的意义实现方式，即对于假定的武汉实验室泄露说的回击，如 *twisting the facts*、*highly hyped theory*，显示了和上述词语型式一致的语义趋向和语义韵。

上述分析表明，美国媒体库和中国媒体库构建了截然对立的病毒溯源话语。美国媒体作出新冠病毒可能来自武汉实验室的有罪推定，并通过媒体反复重述，将这种主观推定变为预设，中国媒体则对之强烈质疑或断然否定、回击。

### 3.3.2 病毒属性

美国媒体库的病毒属性话语表征。该语义组主要聚焦前置修饰语所表达的病毒属性意义。在美国媒体库中，*virus* 与 *deadly*、*dangerous* 和 *killer* 共现，如例（1）—（3）所示。

（1）Fauci’s supporters also denounced the entire line of inquiry about how the **deadly** virus emerged and spread throughout the world. (Newsweek 26-Oct-2021)

（2）“... novel strains of **dangerous** viruses... poses substantially increased risks,” the statement said. (Newsweek 27-Apr-2020)

（3）Surely the decision to create **killer** viruses capable of creating global pandemics ... (Fox News 28-June-2021)

上述例句中 *virus* 的词语搭配揭示了美国媒体库对新冠病毒属性的认定：病毒是致命的，这种致命性体现于在世界范围内的大肆传播（*emerged and spread throughout the world*）；病毒是危险的，造成了诸多风险（*pose substantial increased risks*）；病毒是杀手，它能够造成全球疫情大流行（*global pandemics*）。我们可将其概括为“涌现风险”的语义趋向（*a semantic preference of emerging risks*）。

搭配词 *spread* 的词语索引显示，疫情的蔓延范围广（*to places like Africa*）；感染人数多（*rose to 7375*）；病毒传播不具蓄意性（*unwittingly spread the virus*）；病毒传

播需要加以控制 (curb virus)。我们可以将该组词语的语义趋向概括为“扩散/蔓延” (a semantic preference of spreading)。

证据显示，美国媒体库美国媒体库的病毒属性话语表征也往往将病毒的传播和“中国”或“武汉实验室”相关联。前述的话语立场预设可见一斑。

中国媒体库的病毒属性话语表征。在中国媒体库中，我们同样发现 virus 的高频搭配词 dangerous、deadly 等。但是，中国媒体库有较多的 new、unknown 等显著搭配词，将病毒表征为新型未知的病毒 (unknown/new virus)、仍需严格的科研论证 (rigorous scientific process)、尚缺乏临床经验 (lack clinical experience)，等等。据此，不妨将其语义趋向概括为“不确定性” (a semantic preference of uncertainty)，如表 5 所示。

表 5 中国媒体库中 virus 的词语索引

Among other things, it is an <u>unknown</u>	<b>virus</b>	that has no signs of genetic engineering
an unexpected attack by an <u>unknown</u>	<b>virus</b>	against human beings. China has provided
process to determine whether <u>anew</u>	<b>virus</b>	can be transmitted from person to person.
nobody noticed that we had a <u>new</u>	<b>virus</b>	.So the virus got here much earlier than we
Wuhan authorities was <u>new</u> to the	<b>virus</b>	then, and lacked clinical experience on the new

然而，virus 和 natural、bat、animal 高频共现，揭示新冠病毒和自然界的潜在联系。这些词语共现大都是引用多源信息，将病毒表征为自然发生的或源自自然界，如例（4）和例（5）所示。

（4）All available evidence to date suggests that the novel coronavirus, which has caused the ongoing pandemic, has a **natural animal origin** and is not a manipulated or constructed virus, the World Health Organization (WHO) said in a daily situation report on Thursday. (China Daily 24 April-2020)

（5）... the main point of the paper is that available data supports COVID-19 being caused by a natural virus that probably **emerged from horseshoe bats** and, much like SARS. (Global Times 22-August-2022)

上述例（5）援引世卫组织发布的报道，认为新冠病毒具有动物源性。例（6）则援引科学论文，推测病毒可能是蝙蝠引起的 (emerged from horseshoe bats)。

综上，我们可以看出美国媒体库和中国媒体库关于病毒溯源的主要话语对立在于病毒属性的表征分歧。美国媒体库通过对新冠病毒属性的表征突出其致死性和扩

散性，其传递的语义韵意在加重中国的责任，而中国媒体库则突出新冠病毒的不确定性和自然属性，表达新冠病毒源于自然界的立场。

### 3.3.3 溯源调查

#### 3.3.3.1 美国媒体库的溯源调查话语表征

该语义组揭示病毒溯源的行动者。在美国媒体库中，这些行动者大多是具有权力的机构和政治人物。与 virus 高频共现的词语有 President、Fauci、Trump、community、officials 和 Pompeo 等。这些词经常与 said、express 和 allege 等言语行为动词共现，用于引述美国政府或权威机构/人士的观点，如 Biden administration officials express that..., Fauci said..., U.S. officials have alleged..., Intelligence community has concluded that...。在这种情况下，援引权力机构或权威人物，一方面传递疫情溯源话语的美国国家立场和官方行为，另一方面意在增强命题或观点的力度和效用。这个语义趋向可以概括为“权威性”，其语义韵可以概括为“强力声言”。

#### 3.3.3.2 中国媒体库的溯源调查话语表征

与美国媒体库突显国家和官方层面的言语行为不同，中国媒体库文本强调溯源调查时应采取的行动。由表 3 可见，virus 和 evidence、research、study 等词形高频共现。以 evidence 为例，其左侧多为形容词修饰语，强调证据的充足，右侧表述具体的科学研究发现。这揭示出中国媒体库文本将病毒溯源调查看作是严肃复杂的科学问题，重视累积的多源性证据。检查 research、study 和 virus 的共现，同样发现中国媒体库话语秉持“尊重科学”的立场。据此，所体现的语义趋向可以概括为“证据性”，传递一种“严肃科学探究”的语义韵（见表 6）。

表 6 中国媒体库中 virus 与 evidence 的共现及其词语索引

mounting <u>evidence</u> shows that the	<b>virus</b>	has been circulating in the US before
sufficient clues and <u>evidence</u> that the	<b>virus</b>	was circulating in various countries before
<u>evidence</u> that a SARS-CoV-2-like	<b>virus</b>	was detected in some countries month before
there is <u>evidence</u> to suggest that	<b>virus</b>	was found in the US many months before it
all available <u>evidence</u> suggests the	<b>virus</b>	has an animal origin and is not manipulated

中国媒体库文本中的 virus 与 joint 和 international 频繁共现，强调国际社会及科学界合作的主张。这个共现型式的语义趋向可以概括为“机构团体”，语义韵可以归纳为“合作”。

上述话语构建隐含了两种不同的病毒溯源话语逻辑。美国媒体库文本趋于倚重政府高层和官方的言语行为强化立场，增加力度，以图凭借某种“权威”强力影响受众。这种话语操作不乏有实施某种话语霸权的意图，而中国媒体库文本更多地注

重多元证据的作用，认为溯源是一个科学探索过程。

综上，词形 virus 的搭配分析、词语索引分析、语义趋向和语义韵分析表明，中美两国媒体的新冠肺炎疫情溯源话语呈现 3 组针锋相对的语义表征和态度立场，显示出了中美间的对抗性话语特征。就病毒属性问题，美国媒体强调新冠病毒的致命性和扩散性，意在加重中国的责任；中国媒体强调病毒的不确定性和自然属性，传递病毒源于自然界的立场。美国媒体倚重政府高层人物言语行为，传递一种强力国家立场；中国媒体强调科学证据和机构团体合作，传递病毒溯源是严肃科学问题和提倡全球合作的态度立场。

## 4 讨论与结论

本文基于自建的 COVID-19 语料库，对比分析了中美有关媒体的病毒溯源话语建构与意义表征。我们发现，两组语料库证据呈现出 3 组截然不同的语义趋向和语义韵意义，从多个维向上揭示了中美两国媒体在病毒溯源问题上的话语冲突。

美国媒体库和中国媒体库文本的新冠肺炎疫情溯源报道呈现出极强的对抗性话语特征，映射出对立双方激烈的话语博弈和较量。这些对抗性话语不同于一般的话语交流，显示出一些独特的建构方式及表征。其一，对抗性话语的双方倚重区别明显的语义趋向，传递针锋相对的语义韵意义。这些针锋相对的意义表征背后是截然不同的国家利益、立场和意识形态。语义趋向和语义韵是话语意义表征、传递国家意识形态和立场的重要手段。它们从一个侧面揭示出对抗性话语的意旨与建构方式。其二，对抗性话语不同于一般的话语实践，依循另一套原则。Grice (1975) 认为，一般的话语交流要遵守数量准则、质量准则、关系准则和方式准则等准则。但美国媒体库的病毒溯源话语建构显然缺乏真实证据支持，违反了质量准则。两个库的话语建构也并不旨在寻求共识和注重协商，而在于攻防、博弈。根据 Leech (1983) 的礼貌原则，话语交际双方无论意见是否一致，皆应最大限度地增加对方的受益、减少对方的受损，保护对方的面子。但是，本文研究的新冠病毒溯源话语却显示与之完全不同的话语方式。话语博弈的双方并不在意对方面子，而较多地使用直接指控、驳斥等话语行为。其三，对抗性话语的内在逻辑尤为表明，“话语事件”与“真实世界事件”没有直接的对应关系。驱动话语叙事的是背后的意识形态与国家立场。这些都可从美国媒体库文本不顾证据，将病毒之源强加于中国的做法得到明证。

本文探讨的对抗性话语与目前中美两国的激烈竞争的时代背景密不可分，是其重要的构成。中美关系过去几十年经历了从战略伙伴到竞争者和对手的剧烈变化，从克林顿时代的“战略伙伴关系”，到小布什和奥巴马时期的“接触加遏制”并行，再到特朗普时期中美关系急剧恶化与几乎全方位对抗，直至拜登政府将中国视为“唯一有能力挑战美国主导世界秩序的力量”<sup>3)</sup>。本文研究的对抗性话语是真实世界里中

美关系现实势态的话语再现。在这个意义上,话语世界与真实世界互为构成。激烈的话语对抗是现阶段以及未来较长时期内以竞争、对抗为主要特征的中美关系的构成部分,同时又强化、形塑或重构着双方的竞争、冲突或对抗。

本文揭示出语言使用与话语效力、话语霸权的链条关系。话语实践中反复出现的语言型式,如前文剖析的语义趋向和语义韵,在相当程度上增强着话语效力。它们经媒体的频繁使用,对受众认知产生潜移默化的影响。这一点可由 Hoey (2005: 8) 提出的词汇启动 (lexical priming) 理论来解释。词汇驱动理论认为,词汇在不断被接触获得的过程中,它的各种上下文和语境也会被累积性地获得。而人们的词汇知识也就包括了它在特定情境下与某些词共现的事实。据此,受众反复、持续接触特定的词汇,极有可能产生一种心智触发器 (mental trigger), 启动相关的型式与意义,形成认知语言学所说的形式——意义配对固化 (entrenchment) (Langacker 1987: 59)。这就是说,频繁的形式使用导致特定的话语传播效力。

从另一角度看,语言使用的累积效应、型式一意义的固化关系在话语博弈中起着双重角色作用,既借助话语霸权广为传播,又反过来增强话语霸权。根据 Gramsci (1971), 话语霸权依赖强制 (coercion) 和劝诱 (persuasion) 两条途径。前者主要通过机构性压制、训诫来维持,后者主要通过强势意识形态、价值观、社会风俗的作用达成。在本文中,被研究的美国媒体受众极广、影响力巨大,拥有不可忽视的话语霸权。美国媒体库文本就病毒溯源提出的指控,尽管与真实世界事件相去甚远,也会产生较大的蛊惑煽动影响。这是中方媒体在博弈和较量中需要认真对待和破解的课题。另一方面,前文分析的特定高频话语方式和语言型式皆指向特定话语意义,依赖其固化的型式一意义又潜移默化地操控受众的认知,强化话语霸权。

本文的研究对话语分析可提供概念建构与现实研究取向方面的启示。概念建构层面,我们提出了对抗性话语的概念,在型式调用、意义方式、话语策略等层面揭示了中美媒体在病毒溯源问题上的激烈话语对抗。我们认为,对抗性话语应是未来长时期内中美话语,乃至国际政治舞台多方话语交互的主要形态,其主要特征、意义方式、话语策略、话语效益都值得深入探究。现实取向方面,国内话语研究者似应更多地关注国际交流中的话语对抗。相关研究应在话语意旨、意义方式、话语策略等层面对中国政府以及传播业界提供有意义的建议和参照,助益我国的话语传播效益提升。

## 参考文献

- BAKER P. Using corpora in discourse analysis[M]. London: Continuum, 2006.
- BAKER P, GABRIELATOS C, KHOSRAV INIK M, et al. A useful methodological synergy? Combining critical discourse analysis and corpus linguistics to examine discourses of refugees and asylum seekers in the UK press[J]. Discourse & Society, 2008, 19(3): 273-306.



- BAKER P, MCENERY T. Corpora and discourse studies: integrating discourse and corpora[M]. Basingstoke: Palgrave Macmillan, 2015.
- CURRY N, PÉREZ-PAREDES P. Stance nouns in COVID-19 related blog posts: a contrastive analysis of blog posts published in the conversation in Spain and the UK[J]. *International Journal of Corpus Linguistics*, 2021, 26(4): 469-497.
- FAIRCLOUGH N, WODAK R. Critical discourse analysis[C]//VAN DIJK TA. *Discourse as social interaction*. London: Sage, 1997.
- FIRTH J R. *Papers in Linguistics, 1934-1951*[M]. Oxford: Oxford University Press, 1957.
- GRAMSCI A. *Selections from the prison notebooks*[M]. New York: International Publishers, 1971.
- GRICE H P. *Logic and conversation*[C]//COLE P, MORGAN J. L. *Syntax and semantics*. New York: Academic Press, 1975: 41-58.
- HARDT-MAUTNER G. Only connect: critical discourse analysis and corpus linguistics[R]. UCREL Technical Paper 6. Lancaster: University of Lancaster, 1995.
- HOEY M. *Lexical priming: a new theory of words and language*[M]. London: Routledge, 2005.
- HOOD C. *The blame game: spin, bureaucracy and self-preservation in government*[M]. New Jersey: Princeton University Press, 2011.
- HUNSTON S. *Corpora in applied linguistics*[M]. Cambridge: Cambridge University Press, 2002.
- HUNSTON S. Semantic prosody revisited[J]. *International Journal of Corpus Linguistics*, 2007, 12(2): 249-268.
- HUNSTON S. Starting with the small words: patterns, lexis and semantic sequences[J]. *International Journal of Corpus Linguistics*, 2008, 13(3): 271-295.
- HYLAND K, JIANG, K. The Covid infodemic: competition and the hyping of virus research[J]. *International Journal of Corpus Linguistics*, 2021, 26(4): 444-468.
- LANGACKER R W. *Foundations of cognitive grammar: theoretical prerequisites* (Vol. 1)[M]. Stanford: Stanford University Press, 1987.
- LEECH G. *Principles of pragmatics* [M]. New York: Longman, 1983.
- MCGLASHAN M. Networked discourses of bereavement in online COVID-19 memorials[J]. *International Journal of Corpus Linguistics*, 2021, 26(4): 557-582.
- SALAMA A. Ideological collocation and the recontextualization of Wahhabi-Saudi Islam post-9/11: a synergy of corpus linguistics and critical discourse analysis[J]. *Discourse & Society*, 2011, 22(3): 315-342.
- SINCLAIR J. The search for units of meaning[J]. *Textus*, 1996, 9(1): 75-106.
- SINCLAIR J. *Trust the text: language, corpus and discourse*[M]. London: Routledge, 2004.

- STUBBS M. Words and phrases: corpus studies of lexical semantics[M]. Oxford: Blackwell Publishing, 2001.
- TANG L. Othering as mediated soft-power practice: Chinese diplomatic communication of discourse about China-US trade war through the British press[J]. Discourse, Context & Media, 2023, 51(3): 100669.
- TEUBERT W. A province of a federal superstate, ruled by an unelected bureaucracy: Keywords of the Euro-sceptic discourse in Britain[C]//MUSOLFF A, GOOD C, POINTS P & WITTLINGER R. Attitudes towards Europe: Language in the Unification Process. Aldershot: Ashgate, 2001: 45-86.
- WEAVER R K. The politics of blame avoidance[J]. Journal of Public Policy, 1986, 6(4): 371-398.
- WODAK R. The discourse-historical approach[C]//WODAK R, MEYER M. Methods of critical discourse analysis. London: Sage, 2001: 63-94.
- 李晓红, 卫乃兴. 汉英对应词语单位的语义趋向及语义韵对比研究[J]. 外语教学与研究, 2012 ( 1 ): 20-33.
- 唐丽萍. 美国大报之中国形象的语料库语言学方法辅助下的批评话语分析[M]. 北京: 高等教育出版社, 2016.
- 王登丰, 王安然, 苏岩, 等. 反霸权新闻在新冠病毒溯源报道政治化过程中的作用[J]. 全球传媒学刊, 2022 ( 3 ): 58-75.
- 卫乃兴. 再探经典短语学的要旨和方法: 模型、概念与问题[J]. 外语与外语教学, 2011 ( 3 ): 29-34.
- 卫乃兴. 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, 2012 ( 1 ): 1-6.
- 卫乃兴, 王冰昕. 学术话语探索: 局部语法、话语行为与知识构建[M]. 北京: 外语教学与研究出版社, 2022.
- 张康. 话语博弈下中国外交部例行记者会叙事策略研究——以新冠溯源议题为例[D]. 山东大学, 2022.

通信地址: 100191 北京市 北京航空航天大学外国语学院

# 言语行为理论视角下伊拉克战争中的冲突性话语分析\*

国防科技大学 李新月 马晓雷 李德坤

**提要：**本研究统计了国内外新闻报道中不同群体在伊拉克战争进程中的各种话语，并在 Searle 言语行为理论的框架下对纷繁复杂的话语进行分类。随后，本文梳理了伊战中的冲突性话语，并分析美方使用特定言语行为类型的话语意图。研究发现，战争中的冲突性话语均围绕一个特定的冲突焦点。美军在伊战中频繁使用不同的言语行为类型，以此掌控战争进程，推行霸权主义和强权政策。

**关键词：**言语行为理论、伊拉克战争、冲突性话语、语料库

## 1 引言

伊拉克战争是以美英军队为主的联合部队对伊拉克发动的军事行动。不论是其战争合法性，参战方对战争进程的描述，还是战后伊拉克政府重建问题，都引发了国际社会的不同争论，构成了复杂的冲突性话语立场。从语言学的角度切入，运用 Searle 的言语行为理论分析伊战中的冲突性话语特点，有助于我们从微观语言层面探索和审视冲突性话语在战争中的作用。

## 2 文献综述

冲突性话语是一种普遍存在于人们日常言语交流中的语言现象，因而也成为语言学领域的一个重要研究对象。国外学者在 20 世纪 80 年代取得了一些前期成果，研究角度涉及多类学科，如人类学、社会学、心理学、语言学等。语用学对冲突性话语研究的逐步重视主要体现在 *Journal of Pragmatics* 于 2002 年和 2007 年分别出版的以冲突为主题的两个专刊中。它们从语用的角度探讨了不同的冲突类型、不同语域中的冲突现象、冲突的文化敏感性等，突出了冲突性话语的建构作用。此外，

\* 本文系湖南省社科基金项目“基于文本挖掘的中美主流媒体贸易战报道对比研究”(19YBA047) 的阶段性成果。马晓雷为本文通讯作者。

作者贡献：

李新月：研究方法、数据分析、初稿撰写、字数占比(80%);

马晓雷：选题构思、研究方法、过程研讨，字数占比(20%);

李德坤：数据收集、数据分析。



一些学者研究了各种情景下冲突性话语的表现、形成与管理等。

近年来,国内学者对冲突性话语的关注也显著上升,相关研究多种多样。冉永平(2010)对冲突性话语的语用学研究进行了系统的概述,就冲突性话语的定义、正面性、反面性等主要特征展开讨论,并在文末提出了这一领域的探索范围和主要议题,为之后的研究者们提供了思路。赵英玲(2004)采用结构分析法,对汉语冲突性话语的起始话步、冲突话步及结束话步进行了较详细的模式分析,以此揭示冲突性话语的构建,深化话语分析研究。龚双萍(2011, 2014, 2018, 2020)则针对冲突性话语回应策略与权势问题,冲突性军事外交外语中合法化问题及中美冲突性国防话语文化话语分析问题在语用学上进行了深入分析。此外,在近几年引起热议的网评领域,陈崇国、董保华(2022)对其冲突性展开立场方面的研究。另有赖小玉(2016)结合人际和谐管理模式,探讨家庭冲突性话语的语用效应。

尽管上述学者从多种角度研究了冲突性话语,但是目前还未曾出现针对战争中冲突性话语的相关研究。本文据此尝试在这一空白区域有所探讨,为后续研究提供一点思路。

### 3 理论框架

20世纪50年代,Austin在牛津大学的一次讲座中提出了言语行为理论(Speech Act Theory)。他将言语行为划分为言说行为(locutionary act)、施事行为(illocutionary act)和取效行为(perlocutionary act)。言说行为是指说出合乎语言习惯的、有意义的话语。施事行为意为在特定语境中赋予有意义话语的一种语力。取效行为指得是言说行为或施事行为在听者身上所产生的某种效果。在这3种言语行为中,Austin更偏重施事行为,却鲜少论及其余两种言语行为,以至于施事行为甚至成为言语行为的代名词。换言之,他对施事行为的分类即是对言语行为的分类。依照施事动词的不同,他将言语行为划分为评判、施权、承诺、表态和论理5种类别(Austin 1962)。

后来,Searle(1969, 1973, 1979)在批判继承Austin言语行为理论的基础上,重新确立了言语行为分类的标准。他主张,影响言语行为分类的三个最主要因素是言语行为目的、言辞与外界之间适合方向以及心理状态。除此之外,还有9个次因素,分别是:目的在表现上的强弱程度、言语双方的社会地位、说话者与听话人的利益关系、与话语其他部分的联系、命题内容、言语行为与非言语行为、言语之外的条件、施事动词的使用和言语行为的方式。据此,Searle提出了话语目的(基本条件)、心理状态(真诚条件)、适合方向(先决条件)和命题内容(命题条件)这4条言语行为分类的标准,并将施事行为重新划分为阐述、指令、承诺、表态和宣告5类。本文在此种分类方法的基础上,结合语料自身特点,增加了“预测”一类,对伊战中的言语行为进行了归类,下面将逐一介绍这6种言语行为类型。

(1) 阐述类:说话人描述或陈述其认为是正确的、真实的命题和话语,比如陈

述、告诉、断言、声称、坚持主张等；

(2) 指令类：说话人发出指令希望或试图使听话人做一些事情，比如要求、请求、命令、建议、警告、指示、威胁、坚持要某人做什么事等；

(3) 承诺类：说话人自己承诺采取一些将来的行动，表达说话人自己打算要采取某种行动的意图，比如答应、承诺、发誓、主动提供、允许、保证等；

(4) 表达类：说话人表达自己对某人或某事物的感情态度或传递一种心理状态，比如感谢、祝愿、希望、欢迎、赞扬、拒绝、悼念、后悔等；

(5) 宣告类：说话人通过言语使客观世界上的事情状况即刻发生改变的言语行为，比如命名、任命、祝福、宣判、辞职、判刑等；

(6) 预测类：说话人对将来有可能发生的事情进行预先测定和推测的过程，比如：可能、如果、一旦、终将、预计、也许等。

此后，又有诸多语言学家对言语行为理论进行了丰富和完善。Strawson (1971) 提出以意图为中心的言语行为理论。Bach & Harnish (1979) 在继承 Strawson 思想的基础上，将施事行为划分为规约行为和交际行为。Cohen & Perrault (1979) 提出基于计划的言语行为理论，将言语行为描述为“计划算子”。这一贡献对基于言语行为理论的人工智能研究产生了重大影响。

## 4 研究设计

### 4.1 研究方法

本研究系统搜集了伊拉克战争期间《人民日报》《新华每日电讯》等国内官方主流媒体发表的新闻报道约 500 篇。过滤掉内容重合的报道后，留下共计 379 篇新闻报道。文本细读后，研究者根据发声时间、信息来源、发声者身份、发声内容、发声意图这 5 条标准对伊战中出现的各种话语进行了归类整理，最后获得共计 741 条言语信息，形成本研究的小型语料库。

### 4.2 研究问题

言语贯穿战争进程始终，马晓雷等 (2022) 认为，在战争叙事的总体框架下，语言主要发挥着实施战争行为、描述战争进程、组织战争行为和影响受众认知这 4 种功能。不仅如此，其也与战争发动意图、发动程序、国际人道主义救援、战后重建等问题息息相关。本文梳理了特定时间范围内伊拉克战争中错综复杂、包罗万象的言语行为，旨在回答以下研究问题：

(1) 伊拉克战争中有哪些冲突性话语？

(2) 伊拉克战争中各方话语属于什么言语行为类型？呈现出什么特点？说明了

什么问题？

（3）发声者在战争情境下使用特定言语行为类型话语的意图是什么？可以达到什么样的效果？

4.3 研究结果

4.3.1 伊拉克战争冲突性话语类型

冲突与缓和是言语交流中普遍存在的两种情形。狭义的冲突指的是通过展现彼此的对立而试图保持各自立场的言语行为；广义的冲突进一步包括对立的引发、开始、发展和管理的整个过程。冲突性话语则指交际主体出于立场、角度、观念上的差异，而对同一事件产生了不同的看法，进而出现争论、反驳、反对甚至争吵等对抗性的言语事件和行为。其最典型的特征是对抗性，这一点在战争中体现得尤为明显。

伊拉克战争中针对不同事件而引发的冲突性话语数量繁多，种类繁多。本文主要以战争意图、战争发动程序、战争新闻管控、战后重建及战场外冲突 5 个类目作为切入点，选取了盟军、伊方（叙方）以及国际社会 3 方的典型冲突性话语进行分析，详见表 1。

表 1 伊拉克战争各类型冲突性话语统计

冲突焦点	阵营	冲突性话语示例
战争意图	盟军	美国国务院发言人鲍彻：伊拉克研制出了一种类似南非制造的 <b>集束炸弹的武器</b> ，可以用来对一个目标施放化学武器。
	伊方	伊拉克总统萨达姆：“我们不是武器收藏家”“在两伊战争期间以及在受到以色列威胁时，我们曾拥有这些武器用于自卫”“但是 <b>现在伊拉克已经不再拥有大规模杀伤性武器</b> 。”他还说，“伊拉克确实希望消除本地区及全世界的所有大规模杀伤性武器，但是美国应该自己作出表率首先销毁自己的大规模杀伤性武器。”
	国际社会	伊朗议员拉马赞·瓦希迪：对美国来说，伊拉克将变为第二个越南。美国对伊拉克的进攻使许多国家认识到， <b>华盛顿是一个倾向于“暴力原则”和“武力原则”的潜在侵略者</b> 。
战争发动程序	盟军	美国总统国家安全事务助理赖斯：伊拉克总统萨达姆所玩弄的“游戏”该结束了。 <b>无论联合国安理会是否通过一项授权对伊动武的新决议，美国都将联合一些盟国以武力来解除他的武装</b> 。
	伊方	伊拉克常驻联合国代表杜里：在致函联合国秘书长安南的信件中 <b>谴责美国攻打伊拉克违反了国际法准则</b> ，并要求安理会就此展开辩论。他还说， <b>美国针对伊拉克的军事行动“是侵略战争，是对国际法准则的侵犯</b> 。”他要求联合国拿出实际行动捍卫《联合国宪章》。

（待续）

(续表)

冲突焦点	阵营	冲突性话语示例
战争发动程序	国际社会	联合国秘书长安南：不经安理会授权动武，对伊拉克军事打击的合法性将受到严重损害，并将违反联合国宪章。
	盟军	英国工党主席里德：指责英国 BBC 记者奥马尔报道了巴格达市场受美军导弹袭击后的现场，看起来“就像在为阿拉伯做宣传”
战争新闻管控	伊方	伊拉克新闻部长萨哈夫：否认了美军攻入巴格达市中心的报道。他说，美军只是企图靠近巴格达市区，这无异于自杀行为。萨哈夫要求记者们不要相信美英的谎言。称美方的说法只是为了转移人们对 6 日晚多名美军士兵在巴格达郊区被击毙的注意力。
	国际社会	埃菲社：美国电视台对伊拉克战争的报道内容“进行了谨慎的选择，因此，在美国国内看到的战况与国际社会在世界其他地方看到的有很大的不同”。
战后重建	盟军	美国国务卿鲍威尔：美国不会在战后把对伊拉克的控制权交给联合国。由美国领导的联盟将发挥主要领导作用，直至新政府产生。
	伊方	“伊拉克伊斯兰革命最高委员会”发言人巴亚提：他们抵制会议是因为不满会议安排的方式。因为“这个会议是由弗兰克斯将军（美军中央司令部司令、伊拉克战争总指挥）安排的，我们不能参与一个由美国将军为伊拉克人安排的会议，也不会接受一个由外国人强加于伊拉克的政府。”
	国际社会	埃及总统穆巴拉克：在萨达姆政权倒台后，新政府的组建上应该尽早允许伊拉克人民自己来管理自己的国家。
	盟军	美国国防部长拉姆斯菲尔德：叙利亚在过去的 15 个月中曾进行过化学武器试验。根据美方掌握的情报，一些协助伊拉克军队对美军作战的“外国志愿者”是通过叙利亚入境的。萨达姆政权的部分高官已逃往叙利亚。
战场外冲突	叙方	叙副总理兼外长沙雷：美国领导人指责叙利亚意在“中断叙伊人民的历史和民族联系”，并强调：“这是不可能的。”他指出，连美国人自己都知道，针对叙利亚的上述种种指责是“毫无根据的”，而且至今美国没有拿出任何相关证据。美国此举的目的是转移国际社会的注意力。
	国际社会	阿拉伯国家联盟秘书长穆萨：如果美国在占领伊拉克之后，还要将矛头指向其他中东国家。在中东地区试图推行多米诺骨牌效应，那无异于“火上浇油”。

观察上表可以发现，每一类冲突性话语背后都隐藏着一个类似导火索的冲突焦点。例如：在“战争意图”这一类别中，引发各方冲突性话语的焦点为对发动伊拉克战争原因的认识。具体来说，以美国为首的盟军坚持认为伊拉克已经制造出了大

规模杀伤性武器，足以危及其他国家的安全和世界和平。因此，发动伊战旨在找到并销毁伊拉克拥有的大规模化学武器。而伊方和国际社会则认为美国扮演了侵略者的角色，因为伊拉克现在已经不再拥有美方所声称的“大规模化学武器”，美方此次对伊动武实则是打着“维护世界和平”的幌子另有图谋。

再比如，在“战后重建”过程中如何组建伊拉克新政府这一问题上，各方产生了分歧和争论。美方明确表示，萨达姆政府倒台后将由美方控制伊拉克全境直至新政府产生。伊方则表示拒绝参加由美国将军安排的会议，反对美国强加给伊拉克的“新政府”。国际社会方面，埃及总统穆巴拉克也认为应该由伊拉克人民自己组建新政府，自主管理国家。

4.3.2 伊拉克战争中各方言语行为类型

本文参照 Searle 言语行为理论中对施事行为的划分标准，新增“预测”一类，共 6 类，分别是：阐述类、指令类、表态类、承诺类、宣告类和预测类，然后对搜集到的 741 条伊战信息进行了归纳整理，见表 2。

表 2 伊拉克战争各方言言语行为类型统计

言语行为类型（总计）	阵营	数量	比例（%）	示例
阐述 (295)	盟军	161	54.6	美军中央司令部：美军两支特遣坦克部队突入巴格达城中心。目的是向伊拉克军民传达联军不仅已逼近首都，而且还有能力随时“自由进出”。
	伊方	31	10.5	伊拉克新闻部长萨哈夫：美军战机当天轰炸两辆赴伊拉克充当“人体盾牌”的国际志愿者乘坐的公共汽车，炸伤数人。他还说。过去一天里至少有 56 名伊拉克平民在美英实施的空袭中丧生。萨哈夫同时宣布，伊军在伊拉克中部击毁了 7 辆敌军坦克。
	国际社会	103	35	俄罗斯总统普京：“如果拳头可以替代国际法，强国总是正确，强国可以为所欲为。不受任何限制。那么，作为国际法准则的国家主权不容侵犯的原则将受到动摇。
指令 (81)	盟军	46	56.8	美国中央司令部司令弗兰克斯。规定了“三不政策”： <b>不得问</b> 英联军伤亡情况； <b>不得提</b> 与目前正在进行的军事行动有关的问题； <b>不得问</b> 与今后军事行动计划有关的问题。
	伊方	7	8.6	伊拉克贸易部部长萨拉赫： <b>要求</b> 联合国恢复对伊援助运输。他还谴责美英等中止石油换食品计划。

（待续）



(续表)

言语行为类型 (总计)	阵营	数量	比例 (%)	示例
指令 (81)	国际社会	28	34.6	俄罗斯外交部: 俄驻伊拉克使馆所在居民区数次遭到轰炸, 直接威胁到俄外交宫的安全。俄外交部 <b>要求</b> 美国当局立即采取有效措施, 以避免类似“危险的和不能接受的事件”再次发生。
	盟军	58	29.7	美国国务院发言人鲍彻: 他 <b>担心</b> 许多阿拉伯媒体一直在曲解美军行动并煽动阿拉伯人的反美情绪。
表态 (195)	伊方	38	19.5	伊拉克驻约旦大使亚辛。伊拉克 <b>拒绝</b> 接受来自所有阿拉伯国家的人道主义援助。相比于前者, 伊拉克更 <b>希望</b> 阿拉伯国家采取一个统一的立场制止美英联军对伊拉克人民的侵略。
	国际社会	99	50.8	俄副总理、财政部长库德林: 伊拉克战争结束后债务的延续性对我们来说是最主要的问题。 <b>希望</b> 战后新成立的伊拉克政权能够遵守国际法, 履行还债义务。
承诺 (83)	盟军	44	53	美国国防部部长拉姆斯菲尔德: 美国 <b>计划派遣</b> 1200 名警察顾问和司法专家前往伊拉克, 协助在当地建立安全社会环境。
	伊方	15	18.1	萨达姆长子乌代: 伊拉克已开始动用防空炮火向美国战机还击。 <b>伊还将派出</b> “萨达姆敢死队”回击美英的进攻。
宣告 (3)	国际社会	24	28.9	芬兰总统哈洛宁: 芬兰 <b>将不参与</b> 美可能对伊拉克发动的战争。
	盟军	3	100	美军中央司令部对伊拉克战争指挥部副指挥布鲁克斯。 <b>伊拉克战争的主要军事行动已经结束, 美国和英国联军“已控制了伊拉克全境”, 伊拉克的军事能力已被摧毁。</b> 但今后仍会有“小的激战”, 因为在伊拉克一些地方美军尚未根除所有军事活动, 还有些地方美军根本没有进驻。
预测 (88)	盟军	23	26.1	美国国防部部长拉姆斯菲尔德: 伊拉克战争 <b>可能</b> 持续 6 天、6 周, 但他怀疑会持续 6 个月。
	伊方	4	4.5	巴格达医生塔拉勒大夫: “我 <b>估计</b> 伊拉克要流行瘟疫甚至包括霍乱。这主要因为伊拉克的大量基础设施遭到破坏。”
	国际社会	57	64.5	瑞典首相佩尔松: <b>如果</b> 美英两国在未获联合国授权的情况下强行对伊动武, 它们将为此付出“高昂的政治代价”。

我们可以发现, 在 6 种言语行为类型中, 阐述类是伊战各方使用频率最高的言

语类型,其次是表态类。预测类、承诺类和指令类使用频率基本持平。使用频率最低的是宣告类言语行为类型,仅有3条,且均为盟军所用。

阐述类主要用于说话人陈述或描述其认为是正确的现实状态和情况。不论是战争中哪一方还是国际社会,都难免会对战争进程、战场态势进行描述和评论,这就使得阐述类言语行为类型数量最多。值得注意的是,在利益对立达到顶峰的战争情境下,战争各方尤其是被卷入到战场中的言语交际主体,他们对战争进程、战斗伤亡、战斗力水平等情况的阐述是否真实,究竟有多少可信度?由此也牵扯出了战场假新闻的问题,这也是战争中阐述类言语行为的鲜明特点。

整场战争中,宣告类言语行为类型使用频率最低。究其原因,主要是由战场决定权,或者说参战各方的军事实力乃至综合国力决定的。伊拉克战争中,美国显而易见符合上述标准。论军事实力,它拥有敌方没有的各种高科技武器装备和精锐的战斗力量;论经济实力,美元在世界上的优越地位无可比拟。因而,从战争伊始,伊战的主动权和决定权就掌握在它手里。所以,不论是2003年3月20日,布什总统在电视讲话中宣布伊拉克战争开始,美国白宫发言人弗莱舍4月11日宣布以萨达姆为首的伊拉克政府不复存在,还是美军中央司令部对伊战争指挥部副指挥布鲁克斯准将在4月15日向世界宣告伊拉克战争的主要军事行动已经结束,控制整场战争关键节点的一方始终是以美方为代表的盟军。

同理,在指令类这一言语行为类型中,使用频率最高的依然是美方。正因为有强大的军事实力和坚实的综合国力作支撑,它才有底气围绕伊拉克拥有“大规模化学武器”这一点大做文章,说出“限令萨达姆父子48小时内离开伊拉克,否则将面临以美国为首的军事打击”这样强硬和有震慑力的话语,抑或是在美军在战场上失利之时,凭空捏造出一个个虚假的指控。比如:指责叙利亚政府是奇怪的政权,命令叙政府交出其容留的伊拉克高官,责备俄罗斯政府为伊方提供抗干扰设备,阻碍美军导弹精准打击目标物等。

除此之外,在表态类言语行为类型中,国际社会主要表达了反对盟军擅自发动伊拉克战争,谴责盟军伤害无辜平民及新闻工作者,同情战争中的伊拉克难民,以及希望伊拉克重新建立起代表本国民众利益的新政府等各种复杂情感。

承诺类言语行为意指说话人将来要采取哪些行动,常见的语言实现形式有发誓、保证等。使用该种言语行为类型频率最高的依然是盟军。伊战前期,美军就曾频繁公开承诺一系列未来的作战安排。比如:动用“小鹰号”航母,“尼米兹号”航母等先进武器装备,向海湾地区增派兵力,在伊战中投入号称美军“重拳”,也是美军第一支“数字化”师的第4机步师等。这些话语无疑成为美军“震慑”战略的一部分。到了后期,随着萨达姆政府垮台,美军在伊军事行动基本结束,其主要使用承诺类言语行为用以预告美方在伊拉克战后重建中的动作。比如:派遣相当数目的警察,法律顾问等协助维持当地治安环境,召开旨在讨论筹建伊拉克新政府的会议等。纵

观整个过程,无不体现出美国一手遮天,无视国际准则,干涉别国主权的丑陋面目。

最后,相比于上述可以鲜明体现出权势差异的言语行为类型,预测类言语行为显得更具一般性和普遍性。伊战各方对该类言语行为的使用主要聚焦于对事件走向和后果的猜测和推断。比如:对战争持续时间长短的猜测,对伊战影响的推断等。

#### 4.3.3 发声者在战争情境下使用特定言语行为类型话语的意图

回看伊拉克战争,各方使用特定言语类型传递信息透露出其独特的话语意图。本文主要分析频繁并灵活使用不同言语行为类型为己方造势的美方。

首先,在战争准备阶段,美方擅长使用表态类言语行为,以此拉拢更多盟友提出并通过对伊动武的提案,从而将伊拉克战争合法化。如:布什总统在2002年9月就曾先后列举出自海湾战争后伊拉克向国际社会背弃承诺的事例,进而呼吁联合国采取行动迫使伊拉克销毁大规模杀伤性武器。

其次,在盟军战场行动受挫,导弹误击友军,国内外对美军战斗实力质疑的声音此起彼伏之时,美军“枪口”一转,通过巧妙运用阐述类言语行为开辟了“第二战场”,自导自演了一场捕风捉影、无中生有的闹剧。其中,最为严重的就是针对叙利亚的指责。例如:美国防部长拉姆斯菲尔德曾说:“美国方面‘有情报证实’,正有一些包括夜视镜在内的军需物资通过叙伊边境被运进伊拉克。”而后又一口咬定叙利亚拥有化学武器。不仅如此,美方甚至扬言不排除对叙采取军事手段。无独有偶,其对伊朗、俄罗斯的指责也和前述方法如出一辙。但是,归结到一点,美方如此大费周章制造舆论焦点实质上是为了掩盖其在战场上的失败,转移国际社会视线,舒缓国内外民众的反战情绪。

再次,美方指令类言语行为类型的运用实质上是对其所“阐述”谣言的二次加深。表面上看,美方似乎是基于客观存在的事实,才敢如此自信地发表激烈言论,给国际社会营造出一种假象。通过“警告叙利亚不要窝藏伊拉克领导人,要求叙与美采取合作态度”“指责俄罗斯向伊出售违禁军用设备,使联军遭受损失,要求俄方立即停止有关军售”等话语,美方成功实现了混淆群众视听、减少战场舆论压力的目的。

最后,宣告类和承诺类言语行为类型均是为了展示美军强大的军事实力和雄厚的综合国力。具体来说,美方主要通过承诺类话语表达其必将推翻萨达姆政权,赢得伊拉克战争的信心。宣告类言语则是彰显其世界强国和霸权政策的有力证明:在一场颠覆其他国家主权的战争中,宣布战争开始,萨达姆政府倒台,美军取得胜利等一系列具有重大意义的战争事件的都是美国一方。此外,从长远来看,这也起到了“杀鸡儆猴”的效果:通过宣扬和巩固其超级强国的地位,清除不服从“管教”的异己,从而建立起单极、“绝对”霸权的“新”全球政治、经济秩序,最终实现一超主导世界。



## 5 讨论

本研究主要从语言层面总结了伊拉克战争中冲突性话语的特征,在大量语料信息的基础上揭示了战争各方的话语意图。总体来说,研究问题和研究设计部分还都比较浅显,缺乏一定的学术深度。在未来的研究中,研究者可以尝试融合贯通国际政治、军事学、国家战略、国家安全等方面的学科知识,通过实例分析,总结概括出战争中冲突性话语分析的语用框架,从更系统全面的理论角度探讨战争中的意识形态和国家利益问题。对比分析同一言语交际主体在不同战争中引发、回应和缓解冲突性话语的言语策略也是一个有价值的思路,值得关注和探讨。

## 6 结语

本文在 Searle 对施事行为分类标准的基础上,增加了“预测”一类言语行为类型。然后,利用这个划分标准对伊战各方的冲突性话语和言语行为类型进行了统计整理,并引用相关示例解读了美方在战争中使用特定言语行为类型的话语意图和利益考量,揭露了美方打着“帮助伊拉克建立民主政权”的幌子,推行霸权主义和强权政策的丑陋嘴脸。因而,本研究在一定程度上扩充了冲突性话语的研究范围,对后来者具有些许参考意义。但本文尚属于一个比较笼统的探索性研究,还存在诸多不足之处。

### 参考文献

- AUSTIN J L. How to do things with words[M]. New York: Oxford University Press, 1962.
- BACH K, HARNISH R M. Linguistic communication and speech acts[M]. Cambridge: MIT Press, 1979.
- COHEN P R, PERRAULT C R. Elements of a plan-based theory of speech acts[J]. Cognitive Science—A Multidisciplinary Journal, 1979, 3(3): 177-212.
- SEARLE J R. Speech acts: an essay in the philosophy of language[M]. Cambridge: Cambridge University Press, 1969.
- STRAWSON P F. Intention and convention in speech acts[M]//SEARLE J R. The Philosophy of language. London: Oxford University Press, 1971, 23-38.
- SEARLE J R. Austin on locutionary and illocutionary acts[C]//BERLIN I. Essays on J. L. Austin. Oxford: Oxford University Press, 1973, 141-159.
- SEARLE J R. Expression and meaning: studies in the theory of speech acts[M]. Cambridge: Cambridge University Press, 1979.
- 陈崇国,董保华.冲突性网评中立场表述的语用分析[J].外语学刊,2022(2):8-35.
- 龚双萍.冲突性话语回应策略与权势的语用分析[J].外语学刊,2011(5):76-81.

- 龚双萍. 冲突性网评中情感立场的语用分析[J]. 现代外语, 2014 ( 2 ): 168-178.
- 龚双萍. 中美冲突性国防话语的文化话语分析[J]. 云梦学刊, 2018 ( 2 ): 8-11.
- 龚双萍. 冲突性军事外交话语中合法化的语用研究[J]. 解放军外国语学院学报, 2020 ( 6 ): 26-33.
- 马晓雷, 文月, 张韧. 战争叙事视角下的军事语言协同能力研究[J]. 语言战略研究, 2022 ( 6 ): 34-43.
- 冉永平. 冲突性话语的语用学研究概述[J]. 外语教学, 2010 ( 1 ): 1-6.
- 赵英玲. 冲突话语分析[J]. 外语学刊, 2004 ( 5 ): 37-42.

**通信地址:** 410073 湖南省长沙市 国防科技大学军事外语系

# 大语言模型辅助下基于词典的义项表生成及应用研究\*

广东外语外贸大学/山东农业大学 葛晓帅 山东农业大学 王 梦

**提要：**本文提出了基于词典的义项表的概念和设想。义项表是对语料库词表在语义层面的拓展。文章探究了义项表的呈现形式，并验证了 ChatGPT 大语言模型辅助生成义项表的可行性。研究发现 ChatGPT 模型对义项的区分准确率高、泛化能力强，可以用于辅助生成基于词典的义项表。最后，本文探讨了义项表的价值，并以实例展示了义项表的实际应用。

**关键词：**义项表、语料库、大语言模型、ChatGPT、词典

## 1 引言

语料库是语言研究的重要手段，语料库及语料库检索工具随技术不断发展。电子语料库规模从最早的 100 万词次，到现在上百亿词次。检索工具从早期的专用工具，如 BNC 语料库专用的 Xaira，发展到现在常用的第三代通用检索软件，如 AntConc、WordSmith Tools、LancsBox，再到基于网络的第四代通用检索平台，如 CQPweb、Sketch Engine。

语料库工具中包含的功能也在不断发展。最早的工具中仅有索引行、词表、多词序列列表等功能，后续则添加了关键词表、词频分布图、词云、词汇网络等功能。功能在不断丰富，但语料库研究基于频率的核心思想没有改变。

语料库中的各种词表和统计都是基于频率的。频率反映一个形符在语料库中出现的次数。需要注意的是，形符不一定是单词的形式。许多语料库检索工具允许自行定义形符，即形符能够包含的元素。对于词性赋码的语料，将单词和词性码整体视为一个形符，能够更加精确地统计不同词性的词频。

基于频率，语料库工具通常提供多种统计列表，如词表、词元表、多词序列列表、搭配词表、主题词表等。不同列表从不同角度反映语料库中语言单位的凸显程度。

\* 本文系国家社科基金重点项目“基于融合媒体的多模态英汉学习词典数据库研究”（20AYY012）的阶段性成果。王梦为本文通信作者。

作者贡献：

葛晓帅：选题构思、数据收集、研究方法、初稿撰写、字数占比（60%）；

王梦：选题构思、数据收集、研究方法、数据分析、讨论结论、修改润色、字数占比（40%）。

词表通常按照词频由高到低的顺序反映语料库中不同词形的使用频率。词元表则是在词表的基础上对词形进行归元处理（如 am、is、are、was、were 都归为 be），统计的是词元的累计词频。多词序列表与词表相似，区别在于多词序列表统计的是连续的多个词形的频率。搭配词表和主题词表都涉及计算。搭配词表呈现的是在一定范围内与检索的节点词共现强度高的词。主题词表呈现的是两个语料库对比后，在一个语料库中多用或少用的词。

上述各种列表对于我们了解语料库中的词汇分布起到重要作用，同时我们也发现，这些词表大部分是在形态层次上进行统计处理。词性附码表可以反映一些句法层面的信息，但在语义层面，尚未有相关列表出现。这并不意味着学界没有开展语料库的语义研究，相反，语料库相关的语义研究十分丰富。

## 2 语料库相关的语义研究

语料库相关的语义研究集中在两个领域，一是自然语言处理，二是语料库语言学。

自然语言处理领域对语义的关注可以追溯到 20 世纪 40 年代末，最早是为了进行机器翻译（Agirre & Edmonds 2007）。无论是机器翻译还是自然语言理解，难点都在于词的多义性。根据 WordNet，开放词类中 82.3% 的词都是多义词（Miller *et al.* 1994）。结合语境来判断多义词在语境中使用的义项的任务在自然语言处理领域被称为词义消歧。词义消歧相关研究主要是算法研究，即通过统计、机器学习等算法（Segond *et al.* 1997; Yarowsky 2000; Ide 2006; Parameswarappa *et al.* 2013; Lau *et al.* 2014）提升词义消歧的准确率。

词义消歧算法需要义项标注语料库作为训练和验证数据集。义项标注语料库包括 SemCor、SenseDefs、Senseval-2 等（Agirre & Edmonds 2007; Pasini & Camacho-Collados 2020）。这些语料库通过人工或算法自动根据 WordNet、FrameNet 等数据集中提供的义项对语料库中的词汇进行义项标注。WordNet、FrameNet 等数据集中的义项划分主要是面向机器的，利用面向人的词典开展词义消歧的研究（Lesk 1986）则十分少见。

近年来语义标注也逐渐受到自然语言处理学家的关注。格罗宁根大学的学者（Bjerva *et al.* 2016; Abzianidze & Bos 2017）提出了一种通用的语义标注标签集。该标签集包括回指、言语行为、属性、比较、逻辑等 13 类共 80 个语义标签。语义标注可以提升词性附码、依存语法标注、自然语言推理等各项任务的表现（Abdou *et al.* 2018）。语义标注的思想和实践在语料库语言学领域更早就已出现。

语料库分析工具 Wmatrix 的前身 Matrix 已经使用了 USAS 语义分析系统。该系统把词汇按照语义场分为 20 个大类，232 个小类（Rayson 2003）。这一系统相较于

格罗宁根大学学者提出的分类系统更精细,但依然停留在语义场层次。更加精细的语义分析可以在海词词典中找到。海词词典为词条提供义项常用度信息,如图 1 所示。



图 1 海词词典 bank 词条

这一信息能够帮助使用者了解多义词中不同义项的常用度。但仔细研究会发现,海词词典的释义常用度分布图存在不少问题。首先,其中的义项常用度统计数据来源不明,如在 bazar 词条中出现了“=bazaar.”义项,这显然是乱码造成的。其次,义项分布没有区分词类,图 1 中就可看出动词和名词义项常用度混合在一起。最后,义项常用度信息不全,如“abstract”词条就缺少义项常用度信息。

综观语料库相关的语义研究可以发现,自然语言处理领域常用机器可读的数据集对语料进行词义消歧标注。这些数据集中对词汇的义项划分被认为粒度太细 (Atkins 2003; Agirre & Edmonds 2007),而语义标注的粒度又太粗。词典的义项划分适于人类阅读,但相关研究不足。海词词典的思路与我们将要提出的义项表思路非常接近,但其义项常用度分布没有区分词类,常用度统计可靠性存疑。鉴于此,我们提出基于词典的义项表。

义项表是对词表和词性赋码表在语义层面的延伸。语料库中的词表、词性赋码表等是了解语料库的重要工具,但迄今为止,这些词表都没有涉及语义层面。这是因为语义分析一直以来是一项机器难以胜任的任务。虽然机器能够区分同形异义词 (Agirre & Edmonds 2007),但多义词的识别准确率一直在 90%以下 (Kocoń *et al.* 2023),实际使用价值不高。通过人工进行义项标注可以提升准确率,但需要大量的人力和时间投入,不可能实现大规模应用。大语言模型在语义理解方面的突破性进展给我们带来了希望。能否利用大语言模型帮助我们实现基于词典的义项表生成是本研究的基本问题,可细分为如下 3 个具体研究问题:

- (1) 应以何种形式呈现基于词典的义项表?
- (2) 大语言模型能否用于生成义项表?
- (3) 生成的义项表有何具体应用?

3 义项表的形式

无论是基于形态的词表，还是添加了语法信息的词性赋码表，在语料库检索工具中都被视为字符串。只要字符串相同，则合并统计频率。在生语料库生成的词表中，作名词的 hope 与作动词的 hope 字符串相同，会被视为同一种形符，合并统计词频。在词性赋码语料库生成的词性赋码表中，单词后会添加词类信息，hope 的名词形式为 hope\_NN”，动词形式为“hope\_VVP”。由于赋码后的名词和动词字符串不同，则分别统计词频。Wmatrix 的语义标注也是同样的思路，即通过在单词后添加语义分类标签来区分单词在不同语义场的用法（Rayson 2008）。

参考上述思路，我们认为义项表中的义项形式应当以“token\_lemma\_POS\_senseNo.”的形式呈现。其中“token”是形符，“lemma”是归元后的词头，“POS”是词性码，“senseNo.”是义项编号。中间添加词性码是因为多义词在不同词性下的义项不同，不应将不同词性的义项混为一谈。以句子“They neither liked nor disliked the Old Man.”为例，经义项标注后为：

They\_they\_pron\_1 neither\_neither\_adv\_2 liked\_like\_verb\_1 nor\_nor\_conj\_1 disliked\_dislike\_verb\_1 the\_the\_det\_1 Old\_old\_adj\_5 Man\_man\_noun\_6 \_.\_sent\_0

其中第一个形符“They”的词头为“they”，词类为代词（pron），采用的是词典中的第 1 个义项；第二个形符“neither”的词头为“neither”，词类为副词（adv），采用的是第 2 个义项等。

添加词头信息和词类标签是为了对屈折变化的词形进行合并统计。我们没有采用 TreeTagger 或 CLAWS 词类标签集，而是采用了与词典相同的简化的词类标签。这是因为 TreeTagger 和 CLAWS 标签集对同一词类进行了非常细致的划分，如 TreeTagger 标签集包括 58 个标签，其中名词就有 NN（单数或不可数）、NNS（复数）、NP（专名单数）、NPS（专名复数）4 种，实义动词根据时态和数的不同有 6 种标签（Schmid 1999）。这种细致的性、数、体标签对于义项的区分意义不大，且会造成同一词性同一义项的字符串形式不同，影响后续义项频率统计，因此我们采用与词典一致的简化词类标签。以《牛津高阶英语学习者词典（第 10 版）》（Oxford Advanced Learner’s Dictionary 10th Edition，以下简称 OALD10）为例，基于该词典的义项表应包含以下词类标签（见表 1）。

表 1 基于 OALD10 的词类标签

标签	词类	标签	词类
noun	名词	det	限定词
verb	动词	art	不定冠词

（待续）

(续表)

标签	词类	标签	词类
adj	形容词	infinm	不定式 to
adv	副词	linkv	系动词
prep	介词	modal	情态动词
phv	短语动词	number	基数词
exlm	叹词	ordnu	序数词
abbr	缩写词	suff	后缀
auxv	助动词	pref	前缀
conj	连词	pron	代词
defa	定冠词	sym	符号

经过义项标注的语料，可以用来统计每个单词在不同词类下各个义项的使用频率，按照需求排序后便是义项表。在对一篇 351 词的英文短文进行义项标注后，我们得出以下义项表（见表 2）。

表 2 英文短文义项表前 15 位

序号	义项	频次
1	the_the_det_1	22
2	The_the_det_1	5
3	in_in_prep_1	6
4	in_in_prep_2	1
5	in_in_prep_5	1
6	in_in_prep_9	1
7	in_in_prep_11	1
8	and_and_conj_1	8
9	And_and_conj_3	1
10	a_a_art_1	8
11	was_be_link_3	5
12	was_be_link_2	1
13	was_be_aux_1	2
14	were_be_verb_1	1
15	of_of_prep_4	6



如果更加关注语义层面，可以去除义项表中的形符，仅保留“lemma\_POS\_SenseNo.”的形式。缩减后，上述义项表可以呈现为表 3 的形式。

表 3 短文义项表前 14 位（去除形符）

序号	义项（义位）	频次
1	the_det_1	27
2	in_prep_1	6
3	in_prep_2	1
4	in_prep_5	1
5	in_prep_9	1
6	in_prep_11	1
7	and_conj_1	8
8	and_conj_3	1
9	a_art_1	8
10	be_link_3	5
11	be_link_2	1
12	be_aux_1	2
13	be_verb_1	1
14	of_prep_4	6

表 3 中的义项形式不再关注原形符的形式。如表 2 中的第 1 和第 2 义项形式属于同一义位，在表 3 中合并为第 1 义位。

义项表的排序可以根据需求进行调整，对词头、词类、义项编号设置不同优先级进行排序可以满足不同的研究需要。表 2 和表 3 中是优先按照同一词头的总频次排序，然后按照同词头不同词类频次排序，再次按照义项频次排序。也可根据研究需求优先按照义项等频次排序。

表 2 和表 3 的义项表基于人工标注的英文短文。处理大规模的语料，需要自动化。大语言模型在语义理解方面的优秀表现（King *et al.* 2023）自然促使我们思考能否将其用于义项标注。

## 4 大语言模型辅助的义项标注

大语言模型，尤其是 ChatGPT 在 2022 年底发布后，受到了广泛关注。与之前人工智能模型不同的是其在语义推理和上下文对话中表现出来的能力（Ray 2023；



King *et al.* 2023)。为了测试大语言模型能否辅助实现义项标注的任务。我们选取 ChatGPT 4.0 模型，义项库则选自 OALD10 词典。

选择 ChatGPT 4.0 是因为它是本文写作时最先进的大语言模型。我们在先导研究中发现 4.0 版的模型表现要远优于 3.5 版。因此，选取 ChatGPT 4.0 模型可以最好地反映当前大语言模型的能力。

全部词汇的义项集合在词义消歧任务中称为义项库，最常用的义项库是 WordNet (Agirre & Edmonds 2007)。WordNet 义项划分过细，不适于人类理解。我们选择 OALD10 是因为 OALD10 是网络上最受欢迎的人读英语词典之一，且前人在词义消歧研究中采用过牛津高阶英语学习者词典 (Lesk 1986)。

在先导研究中我们发现 ChatGPT 4.0 模型可以完成义项标注，但为了测试其准确性，有必要进行系统性测试。测试语料选自 Brown 语料库。基本思路是人工先判断单词的义项，之后用 ChatGPT 对同一个句子同一个单词进行义项判断，之后对比模型的判断与人工的判断是否一致。

#### 4.1 多义词的选取

因单义词不涉及义项标注的问题，所以首先要选出多义词。为了使测试结果更具有代表性，我们根据不同词类，各随机选择 100 个多义词，包括以下 4 个步骤：

(1) 对 Brown 语料库进行词性赋码。

(2) 根据词性赋码后的形符生成词性附码表。

(3) 根据 OALD10 的义项划分，对每个单词的义项数量进行标注，筛选出义项数量大于 1 的单词。

(4) 对词性附码表进行分析筛选。词性附码表生成后，首先按照词类进行筛选，在每个词类下，按照频次分层随机选取 100 个单词。

#### 4.2 人工判断

人工判断可以作为机器判断的标准 (Bjerva *et al.* 2016)。为了尽量客观，我们采取双人判断模式。两名判断者均受过词典学专业训练，且长期从事英语教学研究。具体步骤如下：

(1) 从语料库中为多义词随机选取一个句子。

(2) 根据 OALD10 的义项划分，人工判断句子中多义词的义项，并记录该义项。

(3) 两名判断者各自独立判断义项，判断一致的句子保留，否则回到第 1 步。

#### 4.3 模型判断

ChatGPT 的判断表现与提示词质量密切相关。提示词要上下文清晰，让模型扮

演特定角色，指令明确，并提供示例（Shom 2023）。经综合测试效果后，采用如下提示词对 ChatGPT 进行提问：

你是一名词典学家 (lexicographer)、语言专家 (language expert)、自然语言处理专家 (NLP expert)、语境专家 (contextual knowledge expert)、语料库语言学家 (corpus linguist)、语义标注专家 (semantic role labeler)，帮我判断下面句子中 new 一词是词典中的第几个义项。

句子：Surely there is nothing new about artificial intelligence.

词典数据：{形容词 new 在 OALD10 中的数据，Python 字典格式}

判断结果以 Python 字典的形式返回，如 {'sense\_number': 1, 'sense\_definition': 'the actual definition', 'reasoning': 'the reasoning'}，其中 sense\_number 是义项编号，整数；sense\_definition 是义项的内容，字符串。reasoning 是推理过程，字符串。

提示词中的词典数据采用了 Python 字典格式。我们发现提示词用机器可读的格式有助于模型的理解和最终表现。ChatGPT 会返回 Python 字典格式的结果数据，将返回的结果进行记录。

4.4 结果对比

将 ChatGPT 判断的义项编号与人工判断的编号进行对比，判断一致的，认为 ChatGPT 判断准确。对于不一致的判断，我们发现，其中有一些是可以接受的，另一些是错误的。如：在句子 “Most of us brush our teeth by hand.” 中，hand 一词人工判断为第 8 个义项，即 by a person rather than a machine，而 ChatGPT 则判断为第 1 个义项，即 “the part of the body at the end of the arm, including the fingers and thumb”。缺乏语境的情况下，两种判断均可接受。因此，我们参考 Palmer *et al.*（2007），提出两项衡量 ChatGPT 判断准确率的指标：一是一致率：模型判断与人工判断义项一致的词汇数量占总词汇量的比例。二是可接受率：模型判断的词汇义项，被认为可接受的数量，占总词汇数量的比例。结果如表 4 所示。

表 4 ChatGPT 判断结果

词性	一致率	可接受率
名词	93%	96%
动词	95%	98%

（待续）

(续表)

词性	一致率	可接受率
形容词	96%	97%
副词	90%	95%
连词	74%	95%
介词	88%	93%
冠词	71%	74%

从表 4 可以看出, ChatGPT 对义项的判断总体准确率较高, 尤其是实词, 一致率和可接受率均在 90% 以上。连词的一致率较低是因为连词的义项差别较小, 难以区分。介词的一致率稍低, 可能是由于介词本身释义模糊或配例与释义矛盾造成 (Lindstromberg 2001)。冠词的判断一致率和可接受率均远低于其他词性, 观察 ChatGPT 的判断发现其将所有的冠词均判断为第 1 义项, 可能是由于冠词被视为停用词, 在模型中表示不足的缘故。

从综合表现来看, ChatGPT 的判断准确率较高, 且泛化能力强, 可以辅助实现基于词典的义项判断。基于 ChatGPT 的语义判断生成的义项表具有实用价值。

5 义项表的应用

义项表是词表在语义层面的拓展, 因此义项表对于意义研究具有重要的价值。从时间维度上看, 对不同时期的语料库进行义项标注并生成义项表, 可以对比发现词汇意义的历时变迁。从空间维度上看, 对同一语言不同地区的语料库生成义项表, 可以对比发现语言变体中词汇的意义区别。

义项表在不同的学科中也可起到重要作用。在词典编纂中, 义项表可以帮助决策义项排序。在语言教学领域, 义项表可以辅助生成教材词汇表。在二语习得领域, 对学生作文生成义项表可以了解学生的词汇习得过程。不同语料库之间的义项表进行对比分析后, 还可以生成主题义项表, 反映语料库中的义项凸显度。因此义项表还可以用于词汇研究、话语分析等方向。下面用词汇语义变迁的实例展示义项表的应用。

语义变迁是词汇演变的常见现象。下面以名词 screen 为例, 展示义项表在语义变迁研究中的具体应用。我们选取 Brown、Frown 和 Crown 语料库, 分别代表 20 世纪 60 年代, 20 世纪 90 年代和 21 世纪 10 年代的美国英语 (章柏成、许家金 2013)。3 个语料库都属于布朗家族, 语料的选取规则和规模一致, 可以直接对比。

对三个语料库中出现的名词 screen 进行义项标注, 生成义项表, 结果如表 5 所示。

表 5 名词 screen 在不同语料库中的义项分布

screen 的义项	Brown 频次	Frown 频次	Crown 频次
[c]屏幕；荧光屏；荧屏	13	0	61
[c]银幕	12	4	5
[sing][u]（统称）电影，电视	2	14	7
[c]（电脑屏幕上）一屏的内容；画面	0	19	0
[c]隔板；屏风；幕；帘；帐	0	2	2
[c]~（of sth）掩蔽物；掩护物；屏障；庇护	9	6	5
[c]纱窗；纱门	8	5	1
[c]（拦在圣坛或唱诗班四周的木质或石砌）围屏	0	0	0

为了直观对比展示义项的变化，我们将频次信息转换为百分比，并生成柱状图，如图 2 所示：

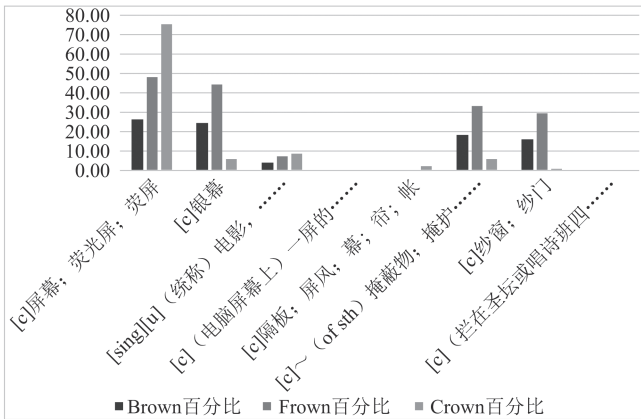


图 2 screen 在不同语料库中的义项占比

从图 2 可以看出，screen 一词各个义项在不同时期的占比不同。如表示“屏幕”的第 1 义项，从 Brown 语料库中的占比 26%，到 Frown 语料库中占比 48%，再到 Crown 语料库中占比 75%。这一变化反映了从 20 世纪 60 年代到 21 世纪 10 年代电子设备普及率提升，各种屏幕在人们生活中越来越常见的现实。与之相对的是表示“纱窗”的第 7 义项，在 Brown 语料库中占比 16%，在 Frown 中占比 29%，但到了 Crown 中，只占 1%。这说明 screen 在新世纪的语境中更少用于表达纱窗这一含义。

由于 3 个语料库规模都较小，有两个义项没有出现在其中，但上述示例表明利用义项表来展示语义变迁具有可行性，利用图表展示，结果非常直观。

## 6 结论

语料库研究依赖各种形式的词表,本文提出的义项表是对词表在语义层面的拓展。通过回顾相关研究,提出了义项表的呈现形式。在对 Brown 语料库中的多个词类进行抽样测试后,本文发现 ChatGPT 模型对语境中的义项判断准确率高,泛化能力强,可以辅助生成基于词典的义项表。义项表在词典学、词汇学、外语教学、话语研究等方面有较高的应用价值。本文还通过具体实例展示了义项表在语义变迁方面的应用。本研究也存在一定局限。受限于人工判断的限制,测试规模较小,且仅测试了 ChatGPT4.0 模型对英语的语义理解,词典数据也仅限于 OALD10。这些局限有待未来研究去突破,但随着大语言模型的不断进步,利用大语言模型对语料库进行义项标注的可行性会越来越高,义项表的应用也将更加广泛。

### 参考文献

- ABDOU M, KULMIZEV A, RAVISHANKAR V, et al. What can we learn from semantic tagging?[J]. arXiv Preprint, 2018. <http://arxiv.org/abs/1808.09716>.
- ABZIANIDZE L, BOS J. Towards universal semantic tagging[J]. arXiv Preprint, 2017. <http://arxiv.org/abs/1709.10381>.
- AGIRRE E, EDMONDS P. Word sense disambiguation: algorithms and applications[M]. New York: Springer, 2007.
- ATKINS S. The Contribution of framenet to practical lexicography[J]. International Journal of Lexicography. 2003, 16: 333-357.
- BJERVA J, PLANK B, BOS J. Semantic tagging with deep residual networks[J]. arXiv Preprint, 2016. <http://arxiv.org/abs/1609.07053>.
- IDE N. Making senses: bootstrapping sense-tagged lists of semantically-related words[J]. Computational Linguistics and Intelligent Text Processing. 2006, 13-27.
- KING E, YU H, LEE S, JULIEN C. “Get ready for a party”: exploring smarter smart spaces with help from large language models[J]. arXiv Preprint, 2023. <http://arxiv.org/abs/2303.14143>.
- KOCOŃ J, CICHECKI I, KASZYCA O, et al. ChatGPT: jack of all trades, master of none[J]. Information Fusion. 2023, 101861.
- LAU JH, COOK P, MCCARTHY D, et al. Learning word sense distributions, detecting unattested senses and identifying novel senses using topic models[R]. Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, 2014, 1: 259-270.
- LESK M. Automatic sense disambiguation using machine readable dictionaries: how to tell a pine code from an ice cream cone[R]. Proceedings of the 5th annual international conference on systems documentation-SIGDOC' 86. 1986: 24-26.
- LINDSTROMBERG S. Preposition entries in UK monolingual learners' dictionaries:

- problems and possible solutions[J]. *Applied Linguistics*. 2001, 22: 79-103.
- MILLER GA, CHODOROW M, LANDES S, et al. Using a semantic concordance for sense identification[R]. *Human Language technology: proceedings of a workshop held at Plainsboro*, 1994.
- PALMER M, NG HT, DANG HT. Evaluation of WSD systems[C]//AGIRRE E, EDMONDS P. *Word sense disambiguation*. New York: Springer, 2007: 75-106.
- PARAMESWARAPPA S, NARAYANA VN, YAROWSKY D. Kannada word sense disambiguation using decision list[J]. *International Journal of Emerging Trends & Technology in Computer Science (IJETTCS)*. 2013, 2: 272-278.
- PASINI T, CAMACHO-COLLADOS J. A short survey on sense-annotated corpora[R]. *Proceedings of the twelfth language resources and evaluation conference*, 2020: 5759-5765.
- RAY PP. ChatGPT: a comprehensive review on background, applications, key challenges, bias, ethics, limitations and future scope[J]. *Internet of Things and Cyber-Physical Systems*. 2023, 3: 121-154.
- RAYSON P. *Matrix: a statistical method and software tool for linguistic analysis through corpus comparison*[D]. Lancashire: Lancaster University, 2003.
- RAYSON P. From key words to key semantic domains[J]. *International Journal of Corpus Linguistics*. 2008, 13: 519-549.
- SCHMID H. Improvements in part-of-speech tagging with an application to German[C]//ARMSTRONG S, CHURCH K, ISABELLE P, et al. *Natural language processing using very large corpora*. Dordrecht: Springer, 1999: 13-25.
- SEGOND F, SCHILLER A, GREFFENSTETTE G, et al. An experiment in semantic tagging using hidden markov model tagging[C]//VOSSEN P, ADRIAENS G, CAKZIKARI N, et al. *Automatic information extraction and building of lexical semantic resources for NLP Applications*, 1997.
- YAROWSKY D. Hierarchical decision lists for word sense disambiguation[J]. *Computers and the Humanities*. 2000, 34: 179-186.
- SHOM. 驾驭 ChatGPT: 学会使用提示词[M]. Boyan, 译. 北京: 电子工业出版社, 2023.
- 章柏成, 许家金. 基于布朗家族语料库的英语现在进行体的历时考察[J]. *外语教学*. 2013, 34: 42-47.

通信地址: 271018 泰安市 山东农业大学外国语学院



# 交替现象实证研究现状与展望<sup>\*</sup>

厦门大学 王梦岚 江桂英

**提要：**交替现象实证研究体现了“语言学+”模式的新型研究进路。语言研究与统计知识融合、量化分析与认知探索并驱，推动了交替研究向纵深发展。本文首先回顾了交替研究的理论视角、主要话题和研究方法，总结得出交替研究的主要进展包括制约因素的对比研究、心理现实性研究以及高阶统计方法的应用与探索。未来研究应在顺应趋势的基础上立足母语中的交替现象，拓宽研究广度、丰富影响因素群、采用在线实验范式、强化 R 语言的统计引领作用。

**关键词：**交替、制约因素、统计建模、心理现实性

## 1 引言

交替 (alternation) 是一种常见的语言现象，如词汇交替、句法交替等。交替结构 (alternating constructions) 指的是同一语言中的两种或多种结构，从直觉上看，它们可以用来表达同一意义 (Belligh & Willems 2022)。伴随着形式语言学、功能语言学、认知语言学的发展，既有研究从理论、描写等视角对交替现象进行了诸多阐述 (Medina 2011)，但主要停留在内省式探讨层面。交替往往被视作一种边缘语言现象，其实证研究未引起国内学界的足够重视 (董晓明、林正军 2018)。随着大数据时代的到来，传统语言现象研究迸发出新的活力。在大规模自然语言数据及多样化统计方法的驱动下，交替研究开始走出结构间关系探究的窠臼，也不再囿于交替结构是否等价表达的讨论，实证研究视角日趋丰富，交替结构特征、制约使用者交替行为的因素类型及特点、制约机制等方面得到更加科学、客观的认识。虽然已有研究持续性引介语言研究中的多因素分析及相关统计方法 (房印杰 2016; 许家金 2020, 2023)，但针对交替现象实证研究的系统性述评仍显不足。因此，本文拟基于文献，聚焦数字化、跨学科背景下的交替研究，呈现其理论、话题、方法等方面的贡献与局限，凸显语言研究与统计知识融合、量化分析与认知探索并驱的两大趋势，以期为进一步促使交替研究向纵深发展、扩大语料库语言学的研究范围提供参考。

<sup>\*</sup> 江桂英为本文通信作者。

作者贡献：

江桂英：选题指导、框架设计、讨论结论、修改润色、字数占比 (30%)；

王梦岚：文献收集、研究方法、初稿撰写、字数占比 (70%)。

## 2 交替研究现状

### 2.1 理论探索及应用

当前交替研究虽偏重数据分析及结果讨论,理论深度渐有提升。一是理论探索。概率语法理论(Bresnan 2007)及概率性语境共选的观点(许家金 2020)异曲同工,主张语言形式选择具有概率性,受制于多种语境因素,均突出了基于使用的语言观。从这一视角可知,使用者的交替行为并不随意,而是基于对语境信息在语言知识中的分布规律作出判断的结果。上述框架扩展了交替知识范围,除交替结构特征、结构间关系之外,还包括概率性语境知识。二是理论应用。一方面,研究基于现有理论探讨交替结构选择的制约因素,如变异社会语言学视角下的教育、职业、社区等社会及地域因素(Rothlisberger & Tagliamonte 2020),复杂动态系统理论所涉及的情境和个体因素(Deshors 2020),以及认知社会语言学所关注的影响方言变异<sup>1</sup>的语体、语域等因素(Engel *et al.* 2022)。另一方面,部分理论为厘清交替结构的异同提供了阐释性框架,包括关注交替结构原型特征的原型范畴理论(Gries 2003)、强调表层形式间关系的构式语法(Goldberg 2006)和构式变体理论(Perek 2015)。

综上可知,多样化的理论视角赋予了交替研究丰富的议题,但因侧重不同,单一理论框架的阐释力度或应用受限,例如概率语法等理论构建有助于概括交替结构选择的概率性本质,但无法进一步为认知和社会等因素提供解释;变异社会语言学视角对语料反映出的个体差异关注不够;复杂动态系统理论对非交际互动语料并不适用。认知语言学视域下,交替结构应当被看作两个独立构式还是上位构式下的构式变体尚未有定论。因此,实际研究中的理论融合与互动不可或缺,应根据研究问题、语料参数等构建合适的理论分析框架。此外,独立构式与构式变体之争需要依托更多交替个案分析,充分探讨交替结构意义功能上的联系。

### 2.2 主要话题

传统交替研究未能基于自然语言数据直达交替结构间关系等问题的核心,而当前基于数据的处理与分析深化了对交替结构本体和使用特征的认识。首先,交替结构异同研究仍是焦点,体现在两个方面:一是原型差异,即交替结构的高显著制约因素集的差异(Schäfer 2019);二是语义异同,研究通过语料计算分析近义构式的语义差异(黄莹、任伟 2020),或是验证构式变体共享同一构式义(De Vaere *et al.* 2020)。除此之外,随着对语言使用语境的深入挖掘,制约因素及其理论阐释逐渐成为热点话题,其基本研究思路是:厘清可能影响交替选择的各类因素,通过统计推断确定各因素及其交互效应的影响程度。上述两方面研究均综合了多因素分析和推断统计方法,但一方面差别在于目标取向不同,前者关注交替结构本身,后者强调实际选择;另一方面差别在于理论阐释不同,前者基于一定的认知理论作出假设

并验证,后者基于数据分析得出结论并进行理论探讨。

总体来看,交替研究有的放矢,话题日渐集中,但还存在三个主要问题。一是过度聚焦典型交替现象,有关位序及成分增减的交替现象层见叠出,如与格交替、关系词省略与否等,而结构成分差异较大的交替现象多因素研究,如允让、致使等才初显研究势头。二是从语言单位来看,交替研究涉及语言各个层面,但尤以词汇和句法为甚,语素研究非常少见。三是反应变量多为二项分类,类别有待丰富。可以看出,现有交替研究的广度显得后劲不足。此外,相当数量的国内研究是对国外成果的复制(许家金 2020),创新度亟待提高。鉴于此,后续研究应积极探寻新的研究基点,拓展交替研究范式的应用范围,如关注教学实践,了解容易出现偏误的交替使用行为;熟悉词典、语法教材等现行教学材料中未有定论的交替使用实例,尝试用多因素统计方法解决问题。

### 2.3 研究方法

语料库研究方法以大量真实语言数据见长,较早应用于交替现象的实证研究。首先,因分析手段差异,基于语料库的多因素研究由独立因素研究向统计建模发展。独立因素研究,又称单变量分析,在内省讨论单因素的基础上发展而来,涉及交替行为的多重因素制约,但各因素分析彼此独立,统计数据无法反映因素的影响权重及交互效应,关于这点可参见 McEnery & Xiao (2005) 的早期研究。常用统计方法包括费希尔精确检验、卡方检验等。统计建模弥补了这一缺陷,实现了对制约因素特征的深入挖掘,如因素影响的显著性、层级、方向、交互等。其次,因研究目标差异,不同的统计模型得以应用。一是探究交替结构间关系,如聚类分析、对应分析、多维尺度分析等。二是通过影响因素分析预测结构选择,如判别分析、决策树分析、逻辑回归分析。三是,逻辑回归分析一方面因交替结构的特征及项数,可进一步细化为二元、序列、多类别等模型,另一方面因自变量的固定及随机特点,发展成为混合效应模型(Glynn & Robinson 2014)。可以说,研究者们正尝试以更成熟的统计手段应对数据分析中的常见问题,如离散性、多重共线性、随机效应变量等,具体选择应视研究目标而定。

心理实验法在交替研究中的应用体现在两个方面。一是直接以实验探究某个或多个因素对交替结构的选择制约;二是结合语料库研究,进行交叉验证。实验任务主要包括语法判断、句子填充、可接受性判断等离线任务,或连续性词汇判定、自定步速阅读等在线范式(Klavan & Divjak 2016)。然而,就交替使用的多因素特性而言,单一实验任务考察的解释性变量非常有限。此外,少数研究使用混合效应模型探究多因素制约机制,但被试多局限于中高水平学习者(Xiang & Chang 2022),学习水平与制约因素的敏感度之间的关系仍待进一步考究,最近一项相关研究可参见 Dubois *et al.* (2022)。综观研究方法的发展趋势可知,将心理实验引入交替研究能为语料库研究得出的结论提供额外证据,以探索认知加工特点,克服过于量化的短板。虽然研究者已较早提出交替研究的实验范式(Bresnan & Ford 2010),但现有

交叉验证研究乏善可陈,直觉实验是主要范式,但被试为交替结构的可接受性打分或者选择符合语境的某个结构往往很难做到即时加工,时效性显得不足。交叉验证的应用广度受限,多以复制研究为主,如与格交替的实验研究范式,缺乏创新性。

### 3 交替研究的主要进展

#### 3.1 制约因素的对比研究

语言类型学、对比语言学等视角下的交替研究丰富了对制约因素的认识,也扩展了相关理论框架的应用范围,主要包括以下方面。一是目标语言从英语、荷兰语等主要欧洲语言拓展至非英语语言。研究角度分为两类:关注与英语类似的交替现象,利用英语交替研究范式,观测本族语中结构选择特征(Li *et al.* 2023; Zhang & Xu 2023);聚焦非英语语言中特有的交替现象,揭示结构选择制约机制的跨语言异同(Fang & Liu 2021)。二是英语变体作为言外因素被引入对比研究。一方面,研究广度不断加宽,同一研究中交替类型增多(Szmrecsanyi *et al.* 2016),覆盖的二语变体数量增加(Grafmiller & Szmrecsanyi 2018);另一方面,研究深度不断拓展,主要体现在制约机制的阐释上,如概率知识本土化(Tamaredo *et al.* 2020)。三是学习者语言产出中的交替研究势头仍然强劲,尤其是以本族语者为参照点,观测学习者交替结构选择的特点(Wulff & Gries 2019)。四是历时对比研究探索了制约因素与时间的互动效应(Larsson 2023),为理解预测变量的认知机制提供了重要窗口(Geyleyn 2017)。五是语料来源从仅关注书面语到同时关注口语(Deshors 2020)。

现有研究逐渐引入言外变量,如语言类别、语言变体、母语背景、时间、语体等,这不仅丰富了影响因素集,通过不断探究变量间的交互作用,还可以加深对交替知识概率性本质的认识,反映出交替研究的良好拓展性。然而,当前研究仍呈现重点分布不均的态势,如欧洲语言研究偏多,英语以外语种的交替研究尚处于薄弱状态。鲜有研究区分二语学习者和外语学习者,从学习者语言知识发展水平差异的角度探究交替结构的理解和产出。制约因素历时变化的动因、历时数据的起始时间点以及时间间隔如何确定尚无定论,如何做到各阶段语料库规模相当、参数大体一致也有待商榷。未来研究一方面可以多从语言类型学角度验证交替研究相关理论框架的普适性,另一方面可以将语言输入环境纳入预测变量,进一步深化学习者交替行为研究。此外,历时研究有必要扩大因素集,考量社会、文化及个人因素如何影响使用者对交替结构的预测能力。

#### 3.2 心理现实性研究

虽然统计建模能科学捕捉结构交替使用中的因素制约规律,但模型能在多大程度上反映语言使用的认知过程仍待验证。针对这一问题,已有研究借助交叉验证的



方法,将统计模型的预测结果与实验任务中被试的预测表现进行比较,以管窥一斑。总体来说,比较维度有以下两方面:模型与被试的预测一致性;模型与被试的预测准确率。从一致性来看,部分研究结果趋向相同,被试的可接受性打分高低、反应快慢,亦或判断选择,均与模型的概率高低一致,说明模型与被试的预测行为受相似主效应的影响,且影响方向一致,从而验证了统计模型具有一定的心理现实性(Bresnan & Ford 2010; 张懂 2020)。从准确率来看,研究显示统计模型与被试的预测准确率均高于基线准确度,且相差不大,说明二者所依赖的语境信息大体一致,模型能较好预测使用者的选择行为(Klavan & Divjak 2016)。然而,以下现象值得注意。一是不同交替研究中模型与被试的预测率有所差异,有的高达百分之九十,而有的则临近于基准,其原因主要包括统计方法差异、研究采取的比较方式不同、被试数量的差异、交替结构的分类(二项或多项)及相似程度不同等。二是多数研究表明,模型预测往往比被试预测更为精准(Klavan 2020),主要原因首先在于被试的个体差异性,对语境中因素特征的分布判断结果不同导致预测行为差异。三是与基于大规模数据的统计模型相比,被试所接受的实验刺激太少。四是实验刺激是否均衡取自统计模型的不同预测区间。

可以看出,交替现象的心理现实性研究体现了量化分析与认知探索的并行趋势,在一定程度上证实了统计模型能较好预测实际使用者的交替选择行为,但当前研究在交替现象、统计方法、实验任务、被试数量及特点、实验刺激等方面的差异程度不一,预测结果的可靠性及普适性有待进一步论证。

### 3.3 高阶统计方法的应用及探索

鉴于语料库数据及交替选择制约因素的复杂性,统计分析愈加严谨。其一,由于语料库数据稀疏性和自变量间的多重共线性易造成混合效应回归模型精确度不高,研究者尝试整合使用多种统计方法,以提高统计模型预测的精准性,如通过借优补劣,融合随机森林及多分类逻辑斯蒂回归模型,完善数据分析过程(田笑语、张炜炜 2020)。其二,完善现有统计方法及其结果阐释,如提高树模型的分类准确率与简约性、将交互效应作为预测变量纳入模型、提升可视化效果等(Gries 2020)。其三,新统计方法的开发也引人瞩目。在研究近义词汇选择时,Gries & Deshors (2014)提出并应用了多因素预测与偏差回归分析,该方法适用于具有参照标准的数据分析,如学习者语言,且在Wulff & Gries (2021)研究学习者领属交替行为中的个体差异时得以进一步完善。其四,为提高验证效度,除引入心理实验外,研究者提倡使用更具心理现实性的统计模型,如朴素判别学习模型、基于记忆的深度学习模型等(Klavan & Divjak 2016)。

客观来说,国外研究中常用统计方法的运用已驾轻就熟,高阶统计探索也在不断深入,而国内研究者的统计意识仍需加强。一方面要根据因素类型,如固定因素、随机因素及因素范围,选择合适的统计方法,另一方面要与时俱进,了解多因素统

计方法的最新发展,如为深掘变量间的因果关系,研究者可尝试应用结构方程模型中的路径分析(Larsson *et al.* 2021)。此外,统计模型的选择应与研究目的相契合,根据研究,就预测准确率来说,混合效应逻辑斯蒂回归与贝叶斯网络、基于记忆的学习模型、朴素判别模型等的差异并不大(Theijssen 2013; Klavan 2020)。最后,无论是何种统计模型,其心理现实性仍需借助心理实验获取汇流证据。

## 4 研究展望

在当前交替现象实证研究中,理论和方法日渐融合,拓展研究也逐渐深入,但是我们认为以下几个方面仍有待突破。

第一,拓宽本土交替现象研究的广度。交替研究范式在英语中的应用已日趋成熟,而对于非英语语言,研究者则重在利用变体或学习者语料优势开展对比分析,根本着力点在语料拓展上,跨出典型交替现象圈的研究屈指可数。以汉语为例,经笔者检索与判断,仅少量研究聚焦汉语所特有的交替现象(Fang & Liu 2021)。值得一提的是,汉语中拥有非常丰富的同义手段,虽然已有研究多从其他视角阐释同义结构,如修辞学(王希杰 1996),但部分观点也为立足汉语开展交替研究提供了有益启示。例如,语法同义句中,“把”字句和“被”字句往往可交替使用,但国内研究多以任务型实验或者 ERP 技术开展此类句子的加工研究(马炳军、常辉 2021),那么基于交替现象的多因素研究趋势,可以思考是否能够利用大型语料库数据展开分析,以实现和实验结果的交叉验证。因此,后续研究一方面可以在既有研究基础上创新研究方法,以进一步挖掘语言使用规律,另一方面则需深耕于跨语言,尤其是跨语系语言,聚合更多交替使用实例,进而深化对概率性语法知识的认识。

第二,丰富影响因素群。多因素分析的必要前提是既有研究成果中的因素探讨,在此基础上的因素补漏是重点也是难点。因研究目标和数据类型差异,研究者往往在预测变量的选择上有所取舍,但趋向穷尽策略,以观交替全貌。多因素分析日趋全面,主要原因在于对认知因素及新奇因素的探索。例如,Kang & Xu (2020)将预测变量与认知过程紧密结合,为揭示交替知识的认知基础提供了新的研究路径。Fang & Liu (2021)探究了鲜少作为交替预测因素的启动效应类型—惊异,同时还利用依存语法理论知识将依存距离纳入解释性变量,为如何丰富因素群作出了很好的学术指引。综合来看,言内及言外因素群已逐步完善,但由于语料标注难度高,语篇层面的因素探讨较为局限。后续研究应注重厘清影响结构选择的语篇手段,如结构启动效应、结构启动的累积效应等(Gries 2017)。此外,针对学习者语料库的多因素研究可结合语料库元数据标注的最新发展,专注更多言外因素,如任务、学习者、文本和标注者(Frey *et al.* 2023)。

第三,采用在线实验范式。现有交叉验证研究中,实验方法多以离线任务为主,



尤其是“100分”可接受性判断实验(Bresnan & Ford 2010)广为使用,而在线实验少有涉及。事实上,已有研究尝试通过自定步速阅读测量反应时,观测被试对含有最强预测变量的句子加工时间与模型预测概率之间的关系,如受事代词性(Ford & Bresnan 2013)、动词的曲折变化(Divjak *et al.* 2016)等。然而值得注意的是,交叉验证中的在线实验不仅需要从语料库中提取自然语言产出作为实验刺激,还需要控制预测变量以规范实验设计,如果期望验证的预测变量较多,还需进行多组实验才能完成验证,难度往往较大。目前来看,自定步速阅读实验对形式差异较大的交替结构,如与格、近义词汇等较为适用,而针对差异较小的交替结构,如语法标记省略,能否有效验证统计模型的心理现实性犹未可知。总体而言,研究者有必要掌握并复制现有实验研究范式,而且要结合其优缺点,不断探索符合研究期待的在线实验范式,如自然阅读下的眼动追踪,以增强验证效度。

第四,强化R语言统计的引领作用。在二语加工数据的处理与分析中,运用R语言拟合混合效应模型已屡见不鲜,然而作为多因素分析的重要领域,量化交替研究尚未与R语言建立起紧密联系。诚然,研究者对统计软件各有偏好,但当前交替研究表明,其语料库、多因素/混合因素、高阶统计等属性与R语言具有较高适配度。首先,交替研究中的预测变量十分丰富,涉及结构、语义、语用、社会文化等各个层面,涵盖固定效应和随机效应,其综合分析需借助高阶统计方法,而R语言赋予了研究者较大的统计自主权,可根据研究需要拟合相应模型。其次,制约因素影响程度及交互效应的统计数据需要可视化呈现,而通过R语言绘制的主效应图美观、直观,便于初学者观察和分析;交互效应图也是呈现变量之间关系、解读交互效应非常有效的方法(吴诗玉 2019)。最后,不仅R语言的功能因其开放、免费特征而不断更新、增强,而且在已有交替研究文献中,多数源代码可直接获取,这对统计新手十分友好。值得一提的是,关于R语言与语言学研究结合的中英文专著也日益增多(Levshina 2015; 许家金等 2023)。研究者只需提高阅读代码的能力、掌握相关软件包的使用规律,即可快速、高效开展重复性研究或拓展研究。针对交替研究的跨学科趋势,R语言作为数据处理与统计分析的常用软件,其优势可见一斑。

## 5 总结

综观国内外相关文献,交替现象研究已体现出语言学、心理学和统计学等学科交叉融合的特点,常见语言现象的分析变得生动有趣,理据阐释也更加全面深刻。受基于使用语言观的影响,交替结构的选择不再被视作受制于硬性限制条件,而是动态的、概率性的。大规模语料库及实验研究为厘清制约因素的影响提供了科学而客观的认识基础,在此基础上的对比研究及心理现实性探索也提高了交替研究的深

度与广度。为促进使用者完善交替知识体系,未来研究应依托语料库、综合运用研究方法和统计分析手段,既要关注交替结构的本体特征、两者/多者间的关系,更要聚焦结构选择的制约因素及机制探讨。最终,研究成果应反哺词典、教材编撰以及教学实践,帮助学习者理解并构建概率性语法知识系统,进而在多样化的语言环境中实现恰当产出。

### 注释

- 1 方语变异 (lectal variation) 是认知社会语言学的重要概念,指的是不同语境下的语言变体使用及其意义变异。为区分其与社会语言学的语言变异概念,本文沿用束定芳、张立飞 (2021) 中的译文。

### 参考文献

- BELLIGH T, WILLEMS K. Epistemological challenges in the study of alternating constructions[J]. *Lingua*, 2022, 280.
- BRESNAN J. Is syntactic knowledge probabilistic? Experiments with the English dative alternation[C]//FEATHERSTON S, STERNEFELD W. *Roots: linguistics in search of its evidential base*. Berlin: Mouton de Gruyter, 2007: 75-96.
- BRESNAN J, FORD M. Predicting syntax: processing dative constructions in American and Australian varieties of English[J]. *Language*, 2010, 86(1): 186-213.
- DE VAERE H, DE CUYPERE L, WILLEMS K. Constructional variation with two near-synonymous verbs: the case of *schicken* and *senden* in present-day German[J]. *Language Sciences*, 2020, 83.
- DESHORS S C. English as a lingua franca: a random forests approach to particle placement in multi-speaker interactions[J]. *International Journal of Applied Linguistics*, 2020, 30(2): 214-231.
- DIVJAK D, ANTTI A, BAAYEN H. Does language-as-used fit a self-paced reading paradigm? (The answer may well depend on how you model the data.)[C]//TANJA A, GATTNAR A, CLASMEIER C. *Slavic languages in psycholinguistics: chances and challenges for empirical and experimental research*. Tübingen: Gunter Narr Verlag, 2016: 52-83.
- DUBOIS T, PAQUOT M, SZMRECSANYI B. Alternation phenomena and language proficiency: the genitive alternation in the spoken language of EFL learners[J]. *Corpus Linguistics and Linguistic Theory*, 2022, 19(3): 427-450.
- ENGEL A, GRAFMILLER J, ROSSEEL L, et al. Assessing the complexity of lectal competence: the register-specificity of the dative alternation after *give*[J]. *Cognitive Linguistics*, 2022, 33(4): 727-766.
- FANG Y, LIU H T. Predicting syntactic choice in mandarin Chinese: a corpus-based

- analysis of ba sentences and SVO sentences[J]. *Cognitive Linguistics*, 2021, 32(2): 219-250.
- FORD M, BRESNAN J. “They whispered me the answer” in Australia and the US: a comparative experimental study[C]//KING T H, DE PAIVA V. *From quirky case to representing space: papers in honor of Annie Zaenen*. Stanford: CSLI Publications, 2013: 95-107.
- FREY J-C, KÖNIG A, STEMLE E, et al. A core metadata schema for L2 data[R]. Paper presented at the 32nd Conference of the European Second Language Association (EUROSLA). UK: University of Birmingham, 2023.
- GELEYN T. Syntactic variation and diachrony. The case of the Dutch dative alternation[J]. *Corpus Linguistics and Linguistic Theory*, 2017, 13(1): 65-96.
- GLYNN D, ROBINSON J. *Corpus methods for semantics: quantitative studies in polysemy and synonymy*[M]. Amsterdam: John Benjamins, 2014.
- GOLDBERG A E. *Constructions at work*[M]. Oxford: Oxford University Press, 2006.
- GRAF MILLER J, SZMRECSANYI B. Mapping out particle placement in Englishes around the world: a study in comparative sociolinguistic analysis[J]. *Language Variation and Change*, 2018, 30(3): 385-412.
- GRIES S T. Towards a corpus-based identification of prototypical instances of constructions[J]. *Annual Review of Cognitive Linguistics*, 2003, 1: 1-27.
- GRIES S T. Syntactic alternation research[J]. *Belgian Journal of Linguistics*, 2017, 31: 8-29.
- GRIES S T. On classification trees and random forests in corpus linguistics: some words of caution and suggestions for improvement[J]. *Corpus Linguistics and Linguistic Theory*, 2020, 16(3): 617-647.
- GRIES S T, DESHORS S C. Using regressions to explore deviations between corpus data and a standard/target: two suggestions[J]. *Corpora*, 2014, 9(1): 109-136.
- KANG H, XU J J. A multifactorial analysis of concessive clause positioning[J]. *Journal of Quantitative Linguistics*, 2020, 28(4): 356-380.
- KLAVAN J. Pitting corpus-based classification models against each other: a case study for predicting constructional choice in written Estonian[J]. *Corpus Linguistics and Linguistic Theory*, 2020, 16(2): 363-391.
- KLAVAN J, DIVJAK D. The cognitive plausibility of statistical classification models: comparing textual and behavioral evidence[J]. *Folia Linguistica*, 2016, 50(2): 355-384.
- LARSSON M. Modelling incipient probabilistic grammar change in real time: the grammaticalisation of possessive pronouns in European Spanish locative adverbial constructions[J]. *Corpus Linguistics and Linguistic Theory*, 2023, 19(2): 177-206.

- LARSSON T, PLONSKY L, HANCOCK G R. On the benefits of structural equation modeling for corpus linguists[J]. *Corpus Linguistics and Linguistic Theory*, 2021, 17(3): 683-714.
- LEVSHINA, N. How to do linguistics with R: data exploration and statistical analysis[M]. Amsterdam: John Benjamins, 2015.
- LI Y, SZMRECSANYI B, ZHANG W W. The theme-recipient alternation in Chinese: tracking syntactic variation across seven centuries[J]. *Corpus Linguistics and Linguistic Theory*, 2023, 19(2): 207-235.
- MCENERY A, XIAO Z H. HELP or HELP to: what do corpora have to say?[J]. *English Studies*, 2005, 86(2): 161-187.
- MEDINA P G. Morphosyntactic alternations: functional and cognitive perspectives[M]. London: Equinox, 2011.
- PEREK F. Argument structure in usage based construction grammar[M]. Amsterdam & Philadelphia: John Benjamins, 2015.
- ROTHLISBERGER M, TAGLIAMONTE S A. The social embedding of a syntactic alternation: variable particle placement in Ontario English[J]. *Language Variation and Change*, 2020, 32(3): 317-348.
- SCHÄFER R. Prototype-driven alternations: the case of German weak nouns[J]. *Corpus Linguistics and Linguistic Theory*, 2019, 15(2): 383-417.
- SZMRECSANYI B, GRAFMILLER J, HELLER B, et al. Around the world in three alternations: modeling syntactic variation in varieties of English[J]. *English World-Wide*, 2016, 37(2): 109-137.
- TAMAREDO I, ROTHLISBERGER M, GRAFMILLER J, et al. Probabilistic indigenization effects at the lexis-syntax interface[J]. *English Language and Linguistics*, 2020, 24(2): 413-440.
- THEIJSEN D, BOSCH T, BOVES L, et al. Choosing alternatives: using bayesian networks and memory-based learning to study the dative alternation[J]. *Corpus Linguistics and Linguistic Theory*, 2013, 9(2): 227-262.
- WULFF S, GRIES S T. Particle placement in learner language[J]. *Language Learning*, 2019, 69(4): 873-910.
- WULFF S, GRIES S T. Exploring individual variation in learner corpus research: methodological suggestions[C]//LE BRUYN B, PAQUOT M. *Learner corpus research meets second language acquisition*. Cambridge: Cambridge University Press, 2021: 191-213.
- XIANG K S, CHANG H. Multiple constraints on second language processing of English dative alternation[J]. *Language Learning and Development*, 2022: 1-21.
- ZHANG D, XU J J. Dative alternation in Chinese: a mixed-effects logistic regression

- analysis[J]. International Journal of Corpus Linguistics, 2023, 28(4): 559-585.
- 董晓明, 林正军. 认知视阈下的论元结构实证研究: 回眸·省思·前瞻[J]. 西安外国语大学学报, 2018, 26 ( 3 ) : 12-17.
- 房印杰. 语言学研究中的多因素分析[J]. 语料库语言学, 2016, 3 ( 1 ) : 82-92.
- 黄莹, 任伟. 英语分析型允让构式的致使倾向研究——多分类逻辑斯蒂回归和多重对应分析法[J]. 外语与外语教学, 2020, ( 3 ) : 11-21.
- 马炳军, 常辉. 汉语二语句法加工研究: 理论与实验[J]. 当代外语研究, 2021, ( 3 ) : 61-70.
- 束定芳, 张立飞. 后“经典”认知语言学: 社会转向和实证转向[J]. 现代外语, 2021, ( 3 ) : 420-429.
- 田笑语, 张炜炜. 汉语变体中分析型致使构式变异研究——多分类逻辑斯蒂回归建模[J]. 外语与外语教学, 2020, 312 ( 3 ) : 22-33.
- 王希杰. 修辞学通论[M]. 南京: 南京大学出版社, 1996.
- 吴诗玉. 第二语言加工及 R 语言应用[M]. 北京: 外语教学与研究出版社, 2019.
- 许家金. 多因素语境共选: 语料库语言学新进展[J]. 外语与外语教学, 2020, ( 3 ) : 1-10.
- 许家金. 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, 2023, ( 3 ) : 442-454.
- 许家金等. 语料库研究方法[M]. 北京: 外语教学与研究出版社, 2023.
- 张懂. 英语与格构式变体的多变量分析及其心理现实性研究[J]. 外语教学与研究, 2020, 52 ( 1 ) : 40-52.

通信地址: 361005 福建省厦门市 厦门大学外文学院

# 基于词频分布指标的莎士比亚十四行诗汉译多维标度和聚类分析研究\*

齐鲁工业大学 陈国兴 鲁东大学 徐德宽

**提要：**本文通过对 8 个与词频分布有关的指标进行方差分析，揭示了莎士比亚十四行诗与 5 个汉译本的异同情况，并通过多维标度和聚类分析，发现没有一个译本能够和原文达到密不可分的程度，但仍有 3 个译本和原文距离较近。本文通过客观数据分析表明，诗歌翻译难度较大，但是在某些量化指标上可以尽量接近原文，同时也印证了某些译本的定性评价。

**关键词：**词频分布指标、量化分析、莎士比亚十四行诗、翻译

## 1 引言

1929 年，译者沈默最早翻译并刊发了莎士比亚十四行诗中的第 18 首，20 世纪 30 年代，朱湘在其选译的诗集《番石榴集》中也收录了多首莎翁的十四行诗，之后又有梁宗岱、方平、屠岸、戴镗龄、虞尔昌、梁实秋、杨熙龄、罗义蕴、高健、孙大雨、丰华瞻、辜正坤、顾子欣和曹明伦等译家陆续加入（卢炜 2019）。几代译者跨越百年，一再重译，努力展现原文风采，虽各有千秋，但评价不一。

人们对于诗歌翻译的评价，多以形式与内容为研究对象，采用定性或者质性研究方法。如李正栓、王心（2019）从文化向度对比了屠岸与辜正坤译本，认为二者都提供了质量上乘的译作，体现了作为跨文化交际使者的责任与担当。卢炜（2019）认为评价诗歌翻译水平有两个重要指标，一是看译作能否充分转译原作的韵式和格律，二是看译作能否在有限的文字空间内转达原作的思想，他通过分析莎士比亚十四行诗第 18 首的朱湘、梁宗岱、梁实秋、丰华瞻和屠岸 5 个中译本，指出梁宗岱译本在意象和文字转译上自然平和，节奏上自然流畅；梁实秋译本力求从简原则，

\* 本文系国家社会科学基金学术外译项目“中国现代设计思想”（21WYSB017）阶段性研究成果。陈国兴为本文通信作者。

作者贡献：

陈国兴：选题及构思修改、研究方法、讨论结论、修改润色、字数占比（55%）；

徐德宽：最初构思、研究方法、数据收集及分析、初稿撰写、字数占比（45%）。



译文可读性较强。此外,卢炜(2019)比较认可屠岸译本,认为屠译本在整体风格上较为统一,在语言和节奏上“以顿代步”,具体到意象和词句的翻译,基本上做到了忠实、晓畅、顺达。黄必康(2017:1)认为除梁实秋、朱生豪等翻译莎士比亚全集的大家外,其中“影响最广的,要数梁宗岱先生、屠岸先生和杨熙龄先生的译本。这几位译诗大家都是诗人,他们译诗十分讲究再现原诗的意境、形式和音律,把莎士比亚译得诗意盎然且朗朗上口,字里行间多显出哲思和诗歌的神韵,外形也译得工工整整。”改革开放后,较为有影响的译家有辜正坤、曹明伦、施颖洲等,他们多为莎士比亚学者,在忠实原诗意境的前提下,在韵脚、行数、字数方面努力追求形似。

语料库语言学兴起以后,量化研究方法被引入翻译研究领域,英国学者 Mona Baker(1995,1999,2001)在这方面进行了引领性研究。采用量化研究方法,我国学者刘泽权等(2011)对《红楼梦》4个英译本的风格进行了对比研究。袁艳玲、戈玲玲(2019)对《二马》英译本中言语幽默机制转换进行了比较分析。秦颖、文秋芳(2012)融合语言形式和语义内容特征,开发了英汉互译自动评分系统。王金铨等(2018)把翻译质量评价方法分为质化和量化两类,指出了量化评价对于自动评价系统的重要性,但像诗歌这种文学性极强的文本是否能够采用量化方法,对不同译本进行质量评价,则是一道难题。潘夏星等运用计量语言学,从词汇量、词频及词性分布等角度,对莎士比亚十四行诗中20首原诗及其4个译本进行了比较,是量化评价应用在诗歌翻译领域的一次有益尝试(Pan *et al.* 2018)。罗益民、刘佺(2019:128)基于语料库语言学方法,对梁宗岱、辜正坤两个译本的翻译风格进行了比较分析,认为“均体现了翻译文本词语层面的简化趋势,句法层面的明确化倾向和意合特征、高频实词与英语原作高度重合、高频虚词中显著增译人称代词等。”沿着这一思路,我们尝试探讨基于词频分布指标的诗歌翻译的量化评价方法及其有效性。

## 2 研究方法

### 2.1 文本及其处理

针对随机抽样可能带来的数据偏差问题,我们将莎士比亚154首十四行诗全部纳入分析范围,并收集了曹明伦、辜正坤、梁宗岱、梁实秋和屠岸5个译本。这样,我们的研究对象为6位作/译者的924首十四行诗,分别以ENG、CML、GZK、LZD、LSQ和TA为代表。我们先对文本进行了认真校对<sup>1</sup>,然后用segtag软件<sup>2</sup>对汉语译本进行了自动分词。文本的基本情况见表1。

表 1 文本基本情况描述统计（词数）

作/译者	个案数	平均值	标准差	标准误	最小值	最大值
ENG	154	116.82	6.256	0.504	93	132
CML	154	124.76	8.203	0.661	102	144
GZK	154	132.16	10.586	0.853	107	160
LSQ	154	124.58	9.388	0.757	105	152
LZD	154	122.51	5.718	0.461	95	147
TA	154	127.25	7.702	0.621	100	156
总计	924	124.68	9.36	0.308	93	160

从表 1 可以看出，原文的平均词数最少，为 116.82 个词，译文平均词数均比原文多，其中平均词数最多的为辜正坤译本，为 132.16 个词；平均词数最少的为梁宗岱译本，为 122.51 个词。

2.2 词频分布指标

2.2.1  $h$ 点、指标 $\alpha$ 和指标 $R_1$

$h$ 点原为文献计量学中的一个概念，由 Hirsch 于 2005 年提出，用来衡量学者论文产出的数量和质量。后来由 Popescu 等人将  $h$ 点引入语言学领域（Popescu & Altmann 2006；Popescu 2009）。

$h$ 点的计算是基于词频数据的，词频数据是指按照具体的词语及该词语出现次数降序排列的数据。 $h$ 点是在按照频次 $f(x)$ （具体词语在文本中的出现次数）从大到小排序的词频表上，频次和频数相等的点。

以莎士比亚十四行诗第一首为例，如表 2 所示，共有单词数 112 个（我们把's 后的 s 单独算作一个单词），共使用了 84 个不同的单词。其中，the 出现了 6 次，次数最多，频数为 1，'s、to 和 thy 各出现了 4 次，频数为 2 到 4，world 出现了 3 次，频数为 5。由于该诗频次和频数相等的点为 4，则该诗的 $h$ 值为 4。

表 2 莎士比亚十四行诗第一首词频数据示例

词语 $x$	频次 $f(x)$	频数 $r(x)$	相对频数 $r_{rel}$	累积频次 $F(x)$	频率 $f_{rel}$	累积频率 $F_{rel}$
the	6	1	0.012	6	0.054	0.054
's	4	2	0.024	10	0.036	0.089
to	4	3	0.036	14	0.036	0.125

（待续）

(续表)

词语 $x$	频次 $f(x)$	频序 $r(x)$	相对频序 $r_{\text{rel}}$	累积频次 $F(x)$	频率 $f_{\text{rel}}$	累积频率 $F_{\text{rel}}$
thy	4	4	0.048	18	0.036	0.161
world	3	5	0.012	21	0.027	0.188
.....	.....	.....	.....	.....	.....	.....
due	1	82	0.976	110	0.009	0.982
grave	1	83	0.988	111	0.009	0.991
thee	1	84	1	112	0.009	1

注：总词数（tokens） $N = 112$ ，不同单词个数（types） $V = 84$

$h$ 点可以被认为是文本中的助词或句法语义词和实义词的平衡点（Popescu 2009: 18），排在 $h$ 点之前的词语多为虚词，出现次数较多。排在之后的词语多为实词，出现次数较少。一般来讲，如果某个文本中的 $h$ 值越大，则该文本中虚词的使用数目就越大，实词使用数目就小；反之， $h$ 值越小，则该文本中虚词的使用数目就越小，实词数目就越大。

然而， $h$ 点会受文本长度的影响，有时 $h$ 点之前的词不全是虚词， $h$ 点后的词也不全是实词，为了消除这种影响，计量语言学家对 $h$ 指标进行了修正，提出了 $\alpha$ 指标（Popescu 2009: 19），该指标的计算公式为：

$$\alpha = N/h^2$$

从该公式可以看出， $\alpha$ 指标与 $h$ 的平方成反比例关系，即 $\alpha$ 值越大，则该文本中虚词的使用数目就越小，实词使用数目就越大； $\alpha$ 值越小，虚词使用数目就越大，实词数目越小。根据上述公式，莎士比亚十四行诗第一首的 $\alpha$ 值为：

$$\alpha = 112/4^2 = 7。$$

$R_1$ 为文本中 $h$ 点之后的词语的覆盖率，即实词语在全文中的覆盖率。表1第4列为累积频次，是第2列的累加值；第5列为频率，是第2列除以总词数所得的商；第6列为累积频率，是第4列除以总词数所得的商。 $H$ 点对应的累积频率， $F_{\text{rel}(h)}$ ，为 $h$ 点之前的词（虚词）在全文中所占的比例，则 $1-F_{\text{rel}(h)}$ 就是实词在全文中所占的比例。由于受到文本长度的影响， $h$ 点之前的词并不全是虚词， $h$ 点之后的词也不一定全都是实词，为了消除文本长度的影响，也引入了 $h^2/2N$ 对实词比例进行修正，其计算公式为：

$$R_1 = 1 - (F_{\text{rel}(h)} - h^2/2N)$$

一般来讲， $R_1$ 值越大，文本中实词覆盖面就越大；反之，虚词覆盖面就越大。根据上述公式，sh-001的 $R_1$ 值为：

$$\begin{aligned} R_1 &= 1 - (F_{\text{rel}(h)} - h^2/2N) = 1 - (F_{\text{rel}(h)} - 4^2/(2 \times 112)) \\ &= 1 - (0.161 - 0.714) = 0.911 \end{aligned}$$

2.2.2 *k*点、指标*b*和指标*R*<sub>2</sub>

上一组指标是建立在词频数据上的。我们还可以根据不同词语的出现次数以及出现次数相同的词语的数量进行排列，这种数据被称为“词频谱（word frequency spectrum）”数据。我们以莎士比亚十四行诗第二首为例（见表3）。

表3 莎士比亚十四行诗第二首词频谱数据示例

出现 次数 <i>x</i>	例词	出现次数为 <i>x</i> 的 词语的数量 <i>g(x)</i>	累积数量 <i>G(x)</i>	<i>g(x)</i> 的相 对频率 <i>g<sub>rel</sub></i>	<i>g(x)</i> 的累积 频率 <i>G<sub>rel</sub></i>
1	cold, warm, blood...	70	70	0.814	0.814
2	thou, my, this...	10	80	0.116	0.930
3	<b>when, s, of</b>	<b>3</b>	<b>83</b>	<b>0.035</b>	<b>0.965</b>
4	and, beauty	2	85	0.023	0.988
6	thy	1	86	0.012	1

如表3所示，出现1次的词（如 cold、warm、blood 等）共有70个，出现2次的词（如 thou、my、this 等）共有10个，出现3次的词（when、s、of）共有3个，出现4次的词（and、beauty）共有2个，出现6次的词（thy）只有一个。仿照*h*点的定义，人们把在词频谱数据中词语出现次数*x*与出现该次数的词语的数量*g(x)*相等的点称为*k*点，只是二者确定的角度不同。从表2可以看出，在莎士比亚十四行诗第二首，*k*点为3，即出现次数为3的词语共有3个。

与*h*点一样，*k*点把文本中出现的词汇也分成两部分：*k*点之前的每个词出现次数较少，意义较实在，数量较多；*k*点之后的每个词出现次数较多，意义较虚，数量较少。一般来讲，如果*k*值越大，文本中使用的不同的虚词个数就越多，实词个数就越少；反之，*k*值越小，虚词个数越少，实词个数越多。因此，*k*点可被认为是文本中常用词和非常用词之间的分隔点。

由于*k*点也会受到文本长度的影响，为了消除这种影响，计量语言学家提出了*b*指标（Popescu 2009：44），该指标的计算公式为：

$$b = V/k^2$$

从计算公式可以看出，*b*指标与*k*的平方成反比例关系，因此，其意义与*k*也相反：如果某个文本的*b*值越大，则该文本中虚词的使用数目就越小，实词使用数目就大；反之，*b*值越小，虚词数目就越大，实词数目就越小。

根据 $b$ 值计算公式，莎士比亚十四行诗第二首的 $b$ 值为：

$$b = 86/(3^2) = 86/9 = 9.56$$

$R_2$ 为文本中 $k$ 点之前的词语的占文本中总词汇量的比例，即意义比较实在的词语在总词汇量中的比例。为了消除文本长度的影响，也对其进行了修正，其计算公式为：

$$R_2 = G_{\text{rel}(k)} - k^2/2V$$

一般来讲， $R_2$ 值越大，文本中含义比较实在的非常用词语占总词汇量中所占比例就越大；反之，含义较虚的常用词语所占比例就越大。根据上述公式，莎士比亚十四行诗第二首的 $R_2$ 值为：

$$R_2 = G_{\text{rel}(k)} - k^2/2V = G_{\text{rel}(3)} - 3^2/2 \times 86 = 0.965 - 0.052 = 0.913。$$

### 2.2.3 $m$ 点和指标 $R_3$

$m$ 点由词频表中的数据，根据以下公式确定：

$$D_r = \sqrt{r_{\text{rel}}^2 + (1 - F_{r,\text{rel}})^2}$$

该公式中 $D$ 值的计算方法为：将词频表中每个词语的相对频率的平方，加上 $(1 - \text{累积频率})$ 的平方，然后再开方。所有词语的 $D$ 值计算出来后， $D$ 值最小的单词对应的点，就是 $m$ 点。

根据上述公式，莎士比亚十四行诗第一首的各个词语的 $D$ 值为：

$$D_1 = \sqrt{0.012^2 + (1 - 0.054)^2} = 0.946503$$

... ..

$$D_{28} = 0.600925$$

$$D_{29} = 0.600284$$

$$D_{30} = 0.600011^*$$

$$D_{31} = 0.600107$$

$$D_{32} = 0.600571$$

... ..

从以上数据可知， $D_{30}$ 的值最小，因此 $m$ 值为 30。从表 1 可以看出， $m$ 点以后的词语出现频次很少，一般为 1 次。因此， $m$ 点可以被理解为文本中罕用词或者个性词与普通词之间的分隔点， $m$ 值越大，文本中出现的个性词越多，常用词越少；反之， $m$ 值越小，个性词越少，常用词越多。

$R_3$ 为文本中个性词语的覆盖率，其计算公式为：

$$R_3 = 1 - F_{\text{rel}(m)}$$

根据该公式, sh-001 的 $R_3$ 值为:

$$R_3 = 1 - 0.518 = 0.482$$

需要指出的是, 在确定 $h$ 点和 $k$ 点的时候, 实际文本中经常不存在频次 $f(x)$ 和频序 $r(x)$ 恰好相等、出现次数 $x$ 与出现次数为 $x$ 的词语的数量 $g(x)$ 恰好相等的点, 这时候需要运用插值法 (interpolation) 确定 $h$ 点和 $k$ 点, 具体计算公式如下 (Popescu 2009: 18-35):

$$h = \frac{r_j \times f_{(i)} - r_i \times f_{(j)}}{r_j - r_i + f_{(i)} - f_{(j)}}$$

$$k = \frac{x_j \times g_{(i)} - x_i \times g_{(j)}}{x_j - x_i + g_{(i)} - g_{(j)}}$$

我们用 Python 语言编写了词频分布指标计算程序, 计算出莎士比亚十四行诗语料中的上述指标。

## 2.3 拟研究的问题

- (1) 莎士比亚十四行诗的原/译文之间在词频分布指标上是否相同?
  - (2) 以词汇分布指标为判断依据, 莎士比亚十四行诗原/译文的近似程度如何?
- 以上两个问题的零假设分别是:

- (1): 莎士比亚十四行诗的原/译文之间在词频分布指标上不存在显著差异。
- (2): 以词汇分布指标为判断依据, 莎士比亚十四行诗的原/译文之间近似程度一致。

## 2.4 数据分析方法

对于问题 1, 我们采用方差分析法; 对于问题 2, 采用多维标度分析和聚类分析进行研究 (桂诗春、宁春岩 1997; 许宏晨 2013)。

# 3 研究结果

## 3.1 方差分析结果

### 3.1.1 指标 $h$ 分析结果

指标 $h$ 的描述性统计见表 4, 各个文本的指标 $h$ 平均值图见图 1, 方差分析结果见表 5。从表 5 可看出, 不同作/译者的指标 $h$ 平均值有显著差异。因指标 $h$ 的方差齐性差异检验不显著, 我们采用 Tukey 事后检验。结果表明, 差异来源于两对译本: 曹明伦和辜正坤译本的 $h$ 指标平均值分别显著低于屠岸 (TA) 译本的指标 $h$ 平均值, 而



原文和各个译本之间以及其他译本之间指标*h*平均值不存在显著差异。

表 4 指标*h*描述性统计

作/译者	个案数	平均值	标准差	标准误	最小值	最大值
ENG	154	3.822	0.494	0.040	3.000	5.000
CML	154	3.733	0.524	0.042	2.500	5.000
GZK	154	3.721	0.607	0.049	2.000	5.500
LSQ	154	3.820	0.518	0.042	2.875	5.250
LZD	154	3.800	0.541	0.044	2.500	5.000
TA	154	3.936	0.554	0.045	3.000	5.667
总计	924	3.805	0.544	0.018	2.000	5.667

表 5 指标*h*方差分析结果

	ENG	CML	GZK	LSQ	LZD	TA	<i>F</i> 值 ( 5,918 )	事后检验 ( Tukey )
<i>h</i> 指标 ( 平均值 )	3.822	3.733	3.721	3.820	3.800	3.936	3.141*	GZK、 CML < TA

\**p* < 0.05。

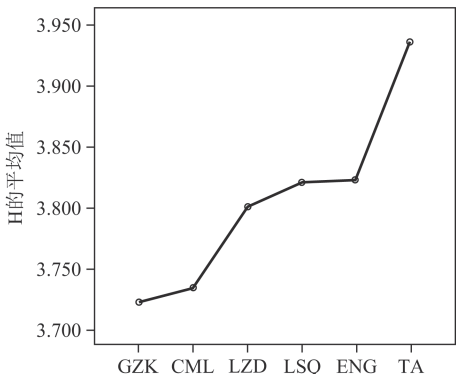


图 1 指标*h*平均值

3.1.2 指标*a*分析结果

指标*a*的描述性统计见表 6，各个文本的指标*a*平均值图见图 2，方差分析结果见表 7。从表 7 可以看出，不同作/译者的指标*a*平均值有显著差异。因指标*a*的方差齐性差异检验显著，我们采用 Tamhane’s 事后检验，结果表明：原文显著低于曹明伦

和辜正坤译本的指标 $a$ 平均值，而屠岸、梁实秋和梁宗岱译本的指标 $a$ 平均值显著低于辜正坤译本的指标 $a$ 平均值。

表 6 指标 $a$ 描述性统计

作/译者	个案数	平均值	标准差	标准误	最小值	最大值
ENG	154	8.375	2.140	.172	4.320	13.444
CML	154	9.432	2.582	.208	4.880	19.520
GZK	154	10.238	3.414	.275	4.529	33.000
LSQ	154	8.955	2.345	.189	4.644	15.778
LZD	154	9.008	2.656	.214	4.760	17.920
TA	154	8.658	2.355	.190	4.711	14.778
总计	924	9.111	2.675	.088	4.320	33.000

表 7 指标 $a$ 方差分析结果

	ENG	CML	GZK	LSQ	LZD	TA	$F$ 值 ( 5,918 )	事后检验 ( Tamhane's )
指标 $a$ ( 平均值 )	8.375	9.432	10.238	8.955	9.008	8.658	9.710*	ENG < CML, GZK TA, LSQ, LZD < GZK

\* $p < 0.05$ 。

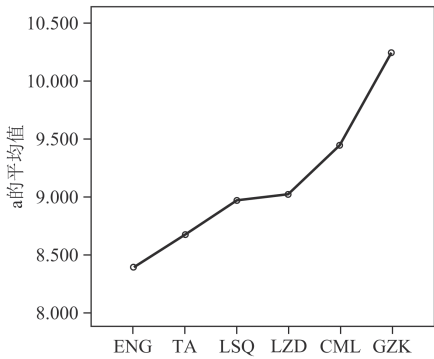


图 2 指标 $a$ 平均值

3.1.3 指标 $R_1$ 分析结果

指标 $R_1$ 的描述性统计见表 8，各个文本的指标 $R_1$ 平均值图见图 3，方差分析结果见表 9。从表 9 可以看出，不同作/译者的指标 $R_1$ 平均值有显著差异。因指标 $R_1$ 的方差齐性差异检验显著，通过 Tamhane's 事后检验，结果表明：5 个译本指标 $R_1$ 平均值

均显著低于原文的 $R_1$ 平均值,此外,有6对译本的 $R_1$ 平均值之间也存在显著差异。梁实秋、屠岸、梁宗岱译本的 $R_1$ 平均值分别显著低于辜正坤、曹明伦译本,而梁实秋、屠岸、梁宗岱译本的 $R_1$ 平均值之间分别没有显著差异,曹明伦和辜正坤译本 $R_1$ 平均值之间也没有显著差异。

表 8 指标 $R_1$ 描述性统计

作/译者	个案数	平均值	标准差	标准误	最小值	最大值
ENG	154	.904	.026	.002	.801	.959
CML	154	.855	.040	.003	.720	.968
GZK	154	.859	.045	.004	.710	.966
LSQ	154	.826	.044	.004	.681	.936
LZD	154	.837	.042	.003	.722	.946
TA	154	.833	.042	.003	.732	.961
总计	924	.852	.048	.002	.681	.968

表 9 指标 $R_1$ 方差分析结果

	ENG	CML	GZK	LSQ	LZD	TA	$F$ 值 (5,918)	事后检验 (Tamhane's)
指标 $R_1$ (平均值)	0.904	0.855	0.859	0.826	0.837	0.833	75.767*	各个汉语译本 < ENG LSQ、TA、LZD < CML、 GZK

\* $p < 0.05$ 。

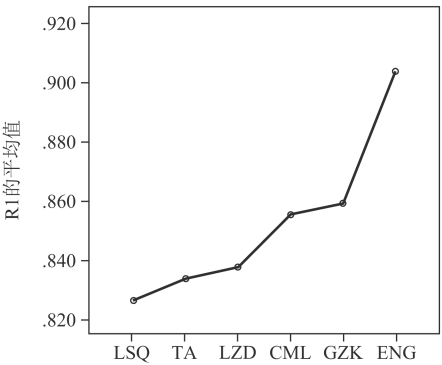


图 3 指标 $R_1$ 平均值

3.1.4 指标 $k$ 分析结果

指标 $k$ 的描述性统计见表 10,各个文本的指标 $a$ 平均值图见图 4,方差分析结果

见表 11。从表 11 可以看出，不同作/译者的指标 $k$ 平均值有显著差异。因指标 $k$ 的方差齐性差异检验不显著，通过 Tukey 事后检验，结果表明：5 个译本指标 $k$ 的平均值均显著低于原文的 $k$ 平均值，而 5 个译本之间的指标 $k$ 的平均值均不存在显著差异。

表 10 指标 $k$ 描述性统计

作/译者	个案数	平均值	标准差	标准误	最小值	最大值
ENG	154	3.415	.520	.042	0.000	4.750
CML	154	3.183	.505	.041	2.000	4.947
GZK	154	3.108	.432	.035	2.333	4.750
LSQ	154	3.197	.489	.039	2.333	5.706
LZD	154	3.149	.429	.035	2.000	5.267
TA	154	3.174	.483	.039	0.000	5.333
总计	924	3.204	.486	.016	0.000	5.706

表 11 指标 $k$ 方差分析结果

	ENG	CML	GZK	LSQ	LZD	TA	$F$ 值 (5,918)	事后检验 (Tukey)
$k$ 指标 (平均值)	3.415	3.183	3.108	3.197	3.149	3.174	7.847*	各个汉语译本 < ENG

\* $p < 0.05$ 。

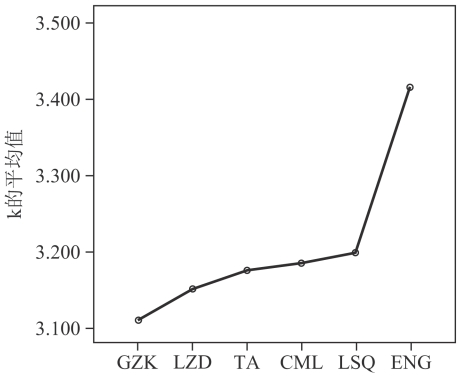


图 4 指标 $k$ 平均值

3.1.5 指标 $b$ 分析结果

指标 $b$ 的描述性统计见表 12，各个文本的指标 $b$ 平均值图见图 5，方差分析结果

见表 13。从表 13 可以看出，不同作/译者的指标*b*平均值有显著差异。因指标*b*的方差齐性差异检验显著，采用 Tamhane’s 事后检验，结果表明：原文的*b*平均值显著低于各个译本指标*b*平均值均，此外，有 3 对译本的*b*平均值之间也存在显著差异：屠岸、梁实秋和梁宗岱译本的*b*平均值分别显著低于辜正坤译本的*b*平均值，其他译本之间*b*值没有显著差异。

表 12 指标*b*描述性统计

作/译者	个案数	平均值	标准差	标准误	最小值	最大值
ENG	154	7.269	2.029	.163	0.000	12.033
CML	154	9.442	2.939	.237	3.705	23.000
GZK	154	10.499	2.649	.213	3.951	17.265
LSQ	154	8.766	2.336	.188	2.703	14.327
LZD	154	9.047	2.387	.192	3.028	21.750
TA	154	8.739	2.284	.184	0.000	16.640
总计	924	8.960	2.630	.087	0.000	23.000

表 13 指标*b*方差分析结果

	ENG	CML	GZK	LSQ	LZD	TA	<i>F</i> 值 (5,918)	事后检验 (Tamhane’s)
指标 <i>b</i> (平均值)	7.269	9.442	10.499	8.766	9.047	8.739	28.387*	ENG < 各个汉语译本 TA、LSQ、LZD < GZK

\**p* < 0.05。

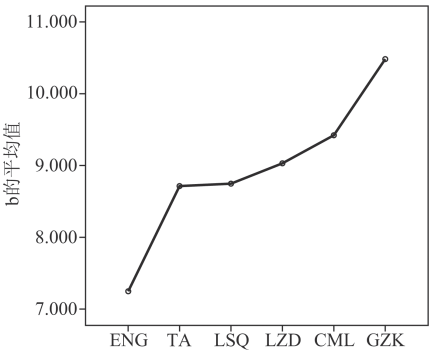


图 5 指标 *b* 平均值

3.1.6 指标*R*<sub>2</sub>分析结果

指标*R*<sub>2</sub>的描述性统计见表 14，各个文本的指标*R*<sub>2</sub>平均值图见图 6，方差分析结

果见表 15。从表 15 可以看出，不同作/译者的指标 $R_2$ 平均值有显著差异。因指标 $R_2$ 的方差齐性差异检验不显著，Tukey 事后检验表明：原文的 $R_2$ 平均值显著低于曹明伦、辜正坤译本的 $R_2$ 平均值，此外，有 3 对译本的 $R_2$ 平均值之间也存在显著差异：屠岸译本的 $R_2$ 平均值分别显著低于曹明伦、辜正坤译本的 $R_2$ 平均值，梁实秋译本的 $R_2$ 平均值显著低于辜正坤译本的 $R_2$ 平均值，其他译本之间的 $R_2$ 平均值没有显著差异。

表 14 指标 $R_2$ 描述性统计

作/译者	个案数	平均值	标准差	标准误	最小值	最大值
ENG	154	.881	.076	.006	0.000	.929
CML	154	.901	.024	.002	.810	.953
GZK	154	.911	.021	.002	.842	.948
LSQ	154	.894	.024	.002	.798	.933
LZD	154	.895	.025	.002	.816	.941
TA	154	.885	.077	.006	0.000	.947
总计	924	.894	.049	.002	0.000	.953

表 15 指标 $R_2$ 方差分析结果

	ENG	CML	GZK	LSQ	LZD	TA	F 值 ( 5,918 )	事后检验 ( Tukey )
指标 $R_2$ ( 平均值 )	0.881	0.901	0.911	0.894	0.895	0.885	7.927*	ENG < CML、GZK TA < CML、GZK LSQ < GZK

\* $p < 0.05$ 。

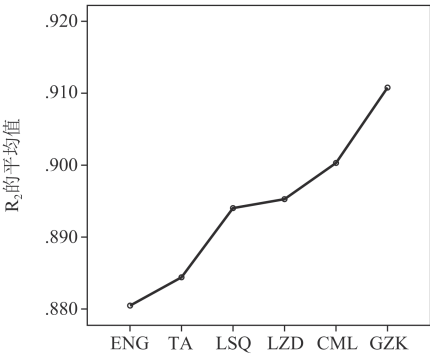


图 6 指标 $R_2$ 平均值

3.1.7 指标 $m$ 分析结果

指标 $m$ 的描述性统计见表 16，各个文本的指标 $m$ 平均值图见图 7，方差分析结果



见表 17。从表 17 可以看出，不同作/译者的指标 $m$ 平均值有显著差异。因指标 $m$ 的方差齐性差异检验显著，Tamhane’s 事后检验表明：6 个原/译本中，辜正坤、曹明伦译本的指标 $m$ 平均值均显著高于原文的 $m$ 平均值，此外，有 7 对译本的 $m$ 平均值之间也存在显著差异：梁实秋、屠岸、梁宗岱译本的 $m$ 平均值分别显著低于辜正坤、曹明伦译本，曹明伦译本 $m$ 平均值显著低于辜正坤译本 $m$ 平均值，而梁实秋、屠岸、梁宗岱译本的 $m$ 平均值之间分别没有显著差异。

表 16 指标 $m$ 描述性统计

作/译者	个案数	平均值	标准差	标准误	最小值	最大值
ENG	154	27.708	3.632	.293	19.000	35.000
CML	154	30.474	4.911	.396	18.000	48.000
GZK	154	33.812	5.847	.471	21.000	52.000
LSQ	154	27.156	4.475	.361	18.000	38.000
LZD	154	28.188	4.610	.371	18.000	41.000
TA	154	27.331	4.756	.383	18.000	43.000
总计	924	29.111	5.299	.174	18.000	52.000

表 17 指标 $m$ 方差分析结果

	ENG	CML	GZK	LSQ	LZD	TA	$F$ 值 (5,918)	事后检验 (Tamhane’s)
指标 $m$ (平均值)	27.708	30.474	33.812	27.156	28.188	27.331	46.085*	ENG < CML、GZK LSQ、TA、 LZD < CML、GZK CML < GZK

\* $p < 0.05$ 。

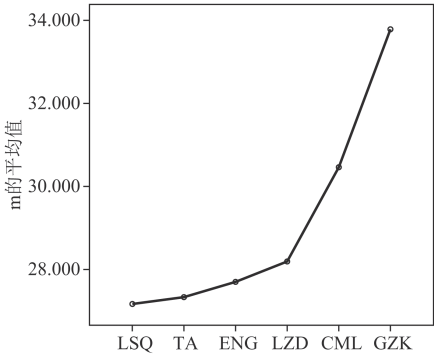


图 7 指标 $m$ 平均值

3.1.8 指标 $R_3$ 分析结果

指标 $R_3$ 的描述性统计见表 18，各个文本的指标 $R_3$ 平均值图见图 8，方差分析结果见表 19。从表 19 可以看出，不同作/译者的指标 $R_3$ 平均值有显著差异。因指标 $R_3$ 的方差齐性差异检验不显著，Tukey 事后检验显示：6 个原/译本中，屠岸译本的指标 $R_3$ 平均值均显著低于英文原文的 $R_3$ 平均值，而辜正坤译本的指标 $R_3$ 平均值均显著高于原文的 $R_3$ 平均值；此外，有 5 对译本的 $R_3$ 平均值之间也存在显著差异：屠岸、梁实秋译本的 $R_3$ 平均值分别显著低于曹明伦和辜正坤译本的 $R_3$ 指标平均值，梁宗岱译本的 $R_3$ 指标平均值显著低于辜正坤译本的 $R_3$ 指标平均值，其他译本之间指标 $R_3$ 平均值不存在显著差异。

表 18 指标 $R_3$ 描述性统计

作/译者	个案数	平均值	标准差	标准误	最小值	最大值
ENG	154	.467	.021	.002	.341	.491
CML	154	.471	.016	.001	.381	.500
GZK	154	.475	.014	.001	.434	.500
LSQ	154	.462	.021	.002	.369	.491
LZD	154	.466	.019	.002	.351	.492
TA	154	.461	.020	.002	.350	.492
总计	924	.467	.019	.001	.341	.500

表 19 指标 $R_3$ 方差分析结果

	ENG	CML	GZK	LSQ	LZD	TA	F 值 (5,918)	事后检验 (Tukey)
指标 $R_3$ (平均值)	0.467	0.471	0.475	0.462	0.466	0.461	12.563	TA < ENG < GZK TA、LSQ < CML、GZK LZD < GZK

\* $p < 0.05$ 。

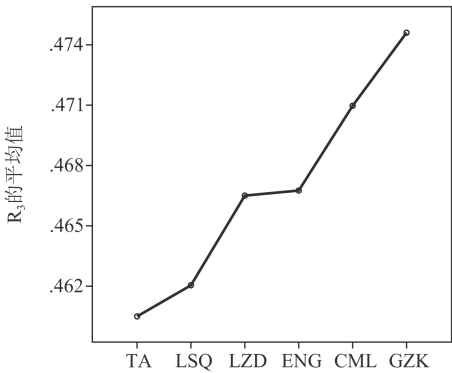


图 8 指标 $R_3$ 平均值

### 3.2 多维标度分析结果

我们使用上面 8 个词频分布指标的平均值作为原始数据，利用 SPSS 的多维标度程序进行分析，得到如图 9 的二维图，其拟合结果为： $S\text{-stress} = 0.036$ ， $RSQ = 0.997$ 。多维标度模型拟合有两个指标：一般说来， $S\text{-stress} \leq 0.05$ ，最好； $0.05 \leq \text{Stress} \leq 0.1$ ，较好； $\text{Stress} > 0.1$ ，较差；RSQ 值越接近 1 越好（桂诗春、宁春岩 1997；许宏晨 2013）。从这两个指标看，图 9 与原始数据之间的一致性关系很高，最终结果可以充分解释原始数据。

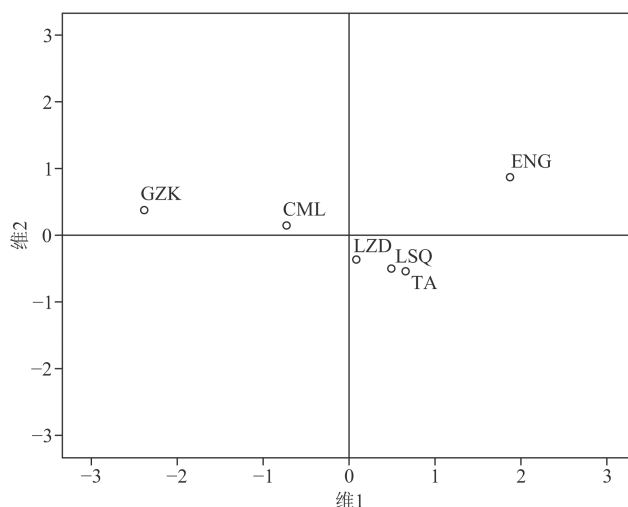


图 9 十四行诗词频分布指标欧式距离多维标度分析结果

从图 9 可以看出，中间坐标为 0 的两个轴将 6 篇原/译文的分布空间分割为 4 个象限，其中原文独居右上的第一象限，显示出一定的独立性；而梁宗岱、梁实秋和屠岸译本位于右下的第四象限，虽与原文处于不同象限，但它们在横轴上的取值均为正值，都位于横轴右侧，且它们之间的空间距离较近，可以认为，它们与原文的相似度较高；辜正坤和曹明伦译本位于左上的第二象限，与原文距离较远，与原文相似度较低。

### 3.3 聚类分析结果

我们使用上面 8 个词频分布指标的平均值作为原始数据，利用 SPSS 的系统聚类程序进行聚类分析，采用余弦相似性计算距离，类别之间使用质心连接，得到如图 10 的聚类分析结果。

从图 10 可以看出，根据词频分布指标的平均值相关性距离，6 篇原/译文被聚成两类，曹明伦和辜正坤译本聚为一类，而梁实秋、屠岸、梁宗岱译本与英文原文聚为另一类。根据聚类分析的结果，梁实秋、屠岸、梁宗岱译本与原文相似度要高，

而曹明伦、辜正坤的译本与原文相似度要低。

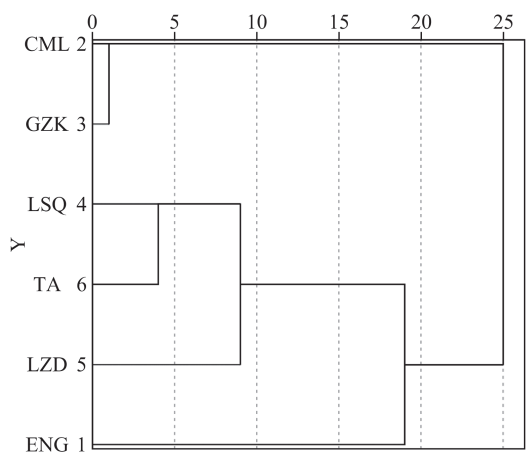


图 10 基于词频分布指标的十四行诗余弦相似性聚类分析结果

#### 4 研究结论

通过上面对词频指标的研究，对本文研究问题的结论如下。

莎士比亚十四行诗的原文与译文之间在词频分布指标存在显著差异。在所有 6 个词汇分布指标上，方差分析的 $F$ 值均超过了 0.05 的显著水平，即在每个指标上，均至少有一个文本与其他文本存在显著差异。总体而言，具体差异因指标不同而有所不同：在指标 $h$ 上，原文与每个译本均不存在显著差异，而仅仅是某些译本之间存在显著差异；在指标 $R_1$ 和 $b$ 上，原文与所有译本之间均存在显著差异，而某些译本之间也存在显著不同；在指标 $k$ 上，原文与所有译本之间存在显著差异，而译本之间均不存在显著差异；在指标 $a$ 、 $R_2$ 、 $m$ 和 $R_3$ 上，某些译本之间差异性较为显著，有的不明显。以上情况表明，各种不同类型的文本之间可译程度高低不一（陈丹 2017），即使是相同的文本，不同的译者仅就词频分布指标而言，也表现出巨大差别。莎士比亚十四行诗原文与译文的近似程度不一多维标度分析表明，原文独处二维平面的一角，其余 5 个译本分布在另外 2 个不同的象限，显示出原文具有一定的独立性，译文与原文有所不同；同时，梁宗岱、屠岸和梁实秋 3 个译本和原文同处于横轴同右侧且距离较近，表明这 3 个译本与原文近似程度更高。

#### 5 结语

通过对莎士比亚十四行诗原文及 5 个译本词频分布指标的方差分析，我们发现，

不同译本在不同指标上会接近原文,原文在某个指标上与所有译本均区别显著,译本之间在某个指标上差异很大,但通过多维标度分析和聚类分析,我们看到,虽然5个译本无一能与原著“密不可分”,但仍有三个译本和原文近似度较高。因此,我们借助基于词频指标的量化分析,得以窥见莎诗汉译的一二“长袖善舞者”,印证了传统定性研究法与基于语料库语言学定量研究方法评价结果的基本一致性。

### 注释

- 1 感谢鲁东大学研究生高菲所做的校对工作。
- 2 segtag 为厦门大学史晓东教授开发的汉语自动分词软件。

### 参考文献

- BAKER M. The role of corpora in investigating the linguistic behaviour of professional translators[J]. International Journal of Corpus Linguistics, 1999, 4(2): 281-298.
- BAKER M. Corpora in translation studies: an overview and some suggestions for future research[J]. Target, 1995, 7(2): 223-243.
- BAKER M. Towards a methodology for investigating the style of a literary translator[J]. Target, 2001, 12(12): 241-266.
- PAN X, CHEN X, LIU H. Harmony in diversity: the language codes in English-Chinese poetry translation[J]. Digital Scholarship in the Humanities, 2018, 33(1): 128-142.
- POPESCU I. Word frequency studies[M]. Berlin: De Gruyter Mouton, 2009.
- POPESCU I, ALTMANN G. Some aspects of word frequencies[J]. Glottometrics, 2006, 13: 23-46.
- 陈丹. 论可译性的不同程度[J]. 上海翻译, 2017 (5): 47-51.
- 桂诗春, 宁春岩. 语言学方法论[M]. 北京: 外语教学与研究出版社, 1997.
- 黄必康. 莎士比亚十四行诗仿词新译[J]. 中国莎士比亚研究通讯, 2017 (1): 1-5.
- 李正栓, 王心. 莎士比亚诗歌翻译中的文化取向——屠岸和辜正坤比较研究[J]. 外国语文研究, 2019 (2): 95-103.
- 刘泽权, 刘超朋, 朱虹. 《红楼梦》四个英译本的译者风格初探——基于语料库的统计与分析[J]. 中国翻译, 2011 (1): 60-64.
- 卢炜. 评《莎士比亚十四行诗第18首》的五个中译本[J]. 东方翻译, 2019 (5): 29-34.
- 罗益民, 刘佺. 基于语料库的莎士比亚十四行诗两种汉译本风格研究[J]. 英语研究, 2019 (2): 128-138.
- 秦颖, 文秋芳. 大规模考试英汉互译自动评分系统的研发与应用[M]. 北京: 高等教育出版社, 2012.
- 王金铨, 万昕, 董子云. 翻译质量评价方法及其在计算机翻译评价系统中的应用[J]. 中国翻译, 2018 (4): 73-78.

- 许宏晨. 第二语言研究中的统计案例分析[M]. 北京: 外语教学与研究出版社, 2013.
- 袁艳玲, 戈玲玲. 汉语小说中言语幽默生成机制与翻译转换: 基于《二马》汉英双语平行语料库的考察[J]. 上海翻译, 2019 (6): 75-81.
- 朱湘. 番石榴集[M]. 上海: 商务印书馆, 1936.

通信地址: 250353 山东省济南市 齐鲁工业大学外国语学院 (陈国兴)  
264025 山东省烟台市 鲁东大学文学院 (徐德宽)



# 评价理论视阈下黄河流域媒体形象的构建与传播对策研究\*

山东建筑大学 王海峰 李文宇

**提要:** 作为国家形象构建的重要一环——黄河流域对外话语构建与形象传播效度成为学界研究的重要课题之一。本研究以 China Daily 官方网站发布的有关黄河流域的英文报道为语料,以评价理论为理据框架,从积极话语分析角度出发,采用语料库语言学量化分析范式,探究黄河流域新闻语篇中态度资源对外传播的话语意义。研究表明:(1)目前“环境治理”“生态保护”“高质量发展”已取代“旱涝灾害”“洪水”“环境污染”成为黄河流域媒体形象新闻语篇报道的三大关键词;(2)正面情感、判断、鉴赏的态度评价词构建了黄河流域新闻语篇的主体、积极、客观的语言环境,突显出黄河流域形象对外传播的积极话语空间构建,展现出高质量、优生态的黄河流域媒体形象;(3)针对黄河流域媒体形象的现状,本研究进一步提出了“自塑”与“他塑”相结合,显性与隐性传播相融合,分众与多模态相契合的对外传播策略,以期为黄河流域对外话语传播提供参考与借鉴。

**关键词:** 黄河流域媒体形象、评价理论、语料库研究范式、传播对策

## 1 引言

黄河流域人口密集,经济发展迅速,是我国北方地区重要的发展地区,同时也是中华文明的重要象征,在我国经济社会发展和生态文明方面处于重要地位。2019年9月,习近平总书记在郑州提出“黄河流域生态保护和高质量发展是重大国家战略”,2022年党的二十大报告再次强调要“推动黄河流域生态保护和高质量发展”。这为黄河流域生态保护和经济发展做出指引。在此背景下,研究中国主流媒体如何对黄河流域进行对外传播,对构建积极话语空间和大国形象具有直接的现实意义。

由中国知网的计量可视化分析结果可知,虽然国内学者很早便开始了对黄河流域的研究,但研究总量有限,研究范围还有待深入和扩展。近年来,国内学界对黄

\* 本文系山东省教学改革重点项目“基于《习近平谈治国理政》中英平行语料库的外语教学体系构建研究”(Z2022124)、山东省重点研发计划(软科学)项目“基于语料库的黄河流域媒体形象构建研究”(2023RKY07017)阶段性研究成果。王海峰为本文通信作者。

作者贡献:

王海峰:选题构思、研究方法、讨论结论、修改润色、字数占比(50%);

李文宇:数据收集、数据分析、讨论结论、初稿撰写、字数占比(50%)。

河流域的研究主要在水利水电工程、环境科学与资源利用、经济体制改革、地质学、农学等视角下开展，而从语料库语言学角度，尤其是在积极话语分析视域下的区域以及国家媒体形象构建的研究较少。因此，本研究以语料库方法为研究范式，以积极话语分析理论为结构框架，以中国对外官方媒体（以 China Daily 为例）的新闻语篇为语料来源，通过情感、判断、鉴赏三个维度对黄河流域形象对外构建进行考察，探索黄河流域形象在新闻语篇中的构建方式以及嬗变过程。

## 2 研究现状

通过搜索中国知网文献库发现，同时将黄河流域、话语分析和传播研究拓展到国家形象构建的相关研究较少，以“黄河”“形象构建”为标题或关键词的期刊发文量不足 5 篇，且未搜索到与“黄河”“话语分析”相关的文章。其中，吴翔、付邦道（2003）在突出“黄河文化”的理念下，以古文化遗产为依托，对开封旅游形象策划与构建进行了研究。王佳、秦娜（2019）再进一步提出，在新媒体环境下要深入挖掘城市文化内涵，黄河则是兰州市具有代表性的文化符号。随后，曹德春（2020）在跨文化视角下，将黄河生态文明建设突破性地上升到了国家形象构建的高度，提出要改善长期以来西方媒体眼里中国环境污染、生态破坏严重的国家形象，并认为国际社会宣传黄河生态文明成就是改善我国国际形象的良好途径。尤元学（2021）则以一部济南形象宣传片为基础，辩证地呈现出这座黄河流域中心城市所负载的多元文化，对黄河流域中心城市历史文化形象的品牌建构与传播，乃至黄河文化蕴含的时代价值的挖掘、黄河故事的讲述起到了示范作用。

总体来看，当前研究已提及“黄河流域形象构建”的概念，但数量有限，且均从历史文化角度出发，以宣传城市为目的进行研究，而很少对新闻报道中的语料进行积极话语分析。目前，有关黄河流域的对外话语研究十分鲜见，主题中包含“黄河”与“话语”的文献与期刊不足 10 篇，且“黄河”多以“黄河文化”为代表出现在文章中，视域较为具体。其中，邢祥、邢军（2020）以黄河文化的传播路径为切入点，从新时代黄河文化传播的时代价值、现实制约和路径创新三方面探讨新时代如何“讲好黄河文化，传播好黄河声音”；焦丹、苏铭（2022）从传播学、翻译学等视角，借鉴翻译传播话语体系构建和中国武术文化国际传播话语体系构建理论依据，探索黄河文化国际传播话语体系构建实践路径及策略建议，同时提出话语只有符合受众的语境、习惯、文化，才能发挥其传播作用。

传播学领域对话语相当倚赖。无论报刊、广播电视，还是网络和社交媒体，无不通过话语传递信息，宣扬立场（许家金 2023）。鉴于黄河流域形象构建的重要现实意义和积极话语分析理论的应用，本研究依托语料库范式，从新闻语篇出发，结合评价理论对“黄河流域”主题的英文报道进行话语分析，旨在研究《中国日报》对“黄河流域”相关报道中的态度资源，分析其对国家战略、区域发展以及国家形

象的构建作用,并探索语料库辅助积极话语分析的研究范式。

### 3 理论基础

#### 3.1 积极话语分析

Zellig Harris 在 1952 年提出了话语分析( discourse analysis )这一术语, Fairclough 在《语言与权利》一书中首次使用“批评话语分析”( critical discourse analysis, 简称 CDA )的概念。批评话语分析以“批评”为核心,揭示话语构建背后的权利以及意识形态,旨在消除不平等的现象,进而改革社会体制(杨林秀、赵浩 2018)。由此可见,批评话语分析多以揭露弊端为主,而非以积极的态度进行话语分析。

在以和平为主题的时代背景下,积极话语分析应运而生,从某种意义上进一步完善了批判话语分析。这一术语由 J. R. Martin 于 1999 年在国际会议上提出,正式应用于 2006 年发表的 Positive Discourse Analysis: Solidarity and Change 一文中,主张采取积极友好的态度来分析社会事实。不同于批评话语以“坏新闻”为对象进行解构式的分析,积极话语以“好新闻”为语言分析对象,倡导采取积极友好的态度,建构一个“宽松、和解、共处”的人类社会( Martin 2004 )。

同年,国内学者也开始了相关研究。朱永生(2006)认为,批评话语分析对所分析的社会语料一般采取揭露和批评的态度,相反,积极话语分析研究权力因素对话语的影响,且更注重和谐因素。因此,积极话语分析在动机、语料选择方法和语料分析方法上对批评话语分析起到了反拨作用。胡壮麟(2012)认为,虽然批评话语分析是解构的,积极话语分析是建构的,但两者并未割裂,且具有互补性。积极话语分析力图营造一种气氛,以共同努力,实现一个既定目标,最后建造一个平等和谐的社会。

#### 3.2 评价理论

评价理论( appraisal theory )是积极话语分析的主要理论基础,是建构人际意义的语篇语义资源。Martin & Rose (2003)将其定义为:“评价理论是关于评价的,即语篇中协商的各种态度、涉及的情感的强度以及表明价值和联盟读者的各种方式”。因此,评价理论从说话者视角出发,利用语言对形象、人际关系构建展开研究,并探索说话者表明主观态度、观点和情感。

评价理论构建了一个庞大而精密的体系,用于研究语言使用者的立场,即态度系统、介入系统和级差系统,是一种目前较为完善的分析系统(陈梅、文军 2013)。如图 1 所示, Martin 将态度子系统分为情感、判断和鉴赏 3 个系统;级差子系统分为语势和聚焦,指态度的强弱程度,所有评价的价值可以根据强度分成不同的等级(王振华 2001)。

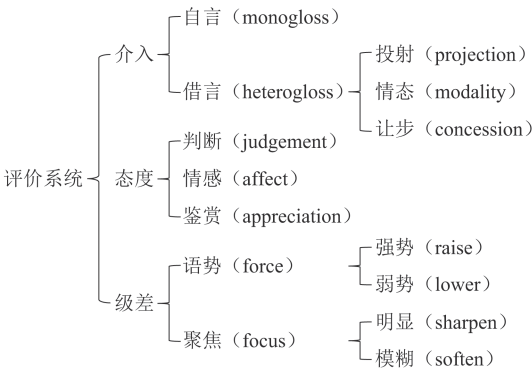


图 1 评价理论体系 (Martin & Rose 2003)

在评价系统的 3 个子系统中，态度系统是重要核心，因此，本研究依据评价理论，从情感、判断、鉴赏三个角度考察了黄河流域新闻语篇中态度资源的运用及其对黄河流域形象与国家形象的构建作用。

4 研究设计

4.1 研究问题

本研究尝试回答以下问题：

- (1) 国内外媒体有关黄河流域的发稿量与主题词呈现出怎样的特点？有何变化？
- (2) 黄河流域文化形象从情感、鉴赏、判断三个维度各自呈现出怎样的特征？
- (3) 基于目前黄河流域文化形象构建与对外传播现状，有何建议和对策？

4.2 语料来源

China Daily 是我国综合性新闻媒体网站，也是中国最大的英文资讯门户。作为沟通中国与世界的网上桥梁，China Daily 已经成为海外读者了解中国的重要窗口。自 2013 年以来，China Daily 中关于黄河流域地区的报道数量不断上升，引发国内外对于这一中国发展区域的关注。因此，本文选取 China Daily 官方网站上有关黄河流域的新闻报道为语料，自建语料库，其中共包含 110 篇新闻报道，共计 32,745 个形符、6589 个类符。所选语料时间跨度设定为 2003 年 1 月至 2022 年 6 月。

4.3 研究过程

本研究从积极话语分析的视角出发，以态度资源为框架，结合语料库研究范式，研究词汇在语篇中的积极意义。语料库语言学的本质是通过定量的方法研究语言的

使用。语言的使用离不开上下语境，Firth（1957）在研究词语的搭配现象时也非常强调语境的重要性，认为词语的意义是语境（搭配词）选择的结果。语言具有丰富的个体特色和语境依存性，语境在语言使用行为上，体现为共选关系，这是一种依存式的关联（许家金 2020）。本研究将通过观察有关黄河流域报道中主题词表及其索引行，分析新闻语篇中态度资源的运用及其对国家形象构建的作用。

研究过程包括以下 3 个步骤：

首先，以 China Daily 2003 年 1 月至 2022 年 6 月关于黄河流域的 110 篇英文专题报道为语料，自建语料库。研究采用的工具为文本整理器 3.0、语料库软件 AntConc 3.5.7 与 BFSU Collocator 1.0。分别利用 AntConc 3.5.7 与 BFSU Collocator 1.0 生成主题词表和强搭配词词表。

其次，结合态度资源分析关键词及其索引行，抽取出搭配词的索引行，并通过上下文信息对索引行所处语境进行分析，从而得到能够描述该主题词的态度资源词。结合黄河流域对外传播新闻语篇中态度资源建构的中国形象，对中国主流媒体黄河流域的对外新闻语篇展开分析。

最后，总结新闻语篇特点，从国家形象“自塑”与“他塑”以及多模态形象构建等多重角度推动时事文化相融合，传播黄河流域文化，同时促进传播渠道融合，扩大新闻覆盖面，为黄河流域的发展和国家形象的构建建言献策，从而扩大国际传播影响力、提高中华文化感召力、提升中国形象亲和力、增强中国话语说服力、壮大国际舆论引导力（徐国亮 2021）。

## 5 研究结果与讨论

### 5.1 国内外媒体新闻语篇数量对比研究

本研究统计了 2004 年至 2022 年 10 月国内外主流媒体有关黄河流域新闻报道的数量（见图 2）。其中，国内主流媒体即 China Daily，国外媒体以 Time、The Guardian、BBC、CNN、REUTERS 为主。

由图 2 可见，国内新闻报道呈逐渐增长趋势，在 2016 年突破性地增长到 2683 篇，其中多为习近平总书记对黄河流域进行考察、提出相关区域治理理念的主题报道。自 2020 年以来，报道黄河流域生态保护和高质量发展成效的新闻数量比重达到了 46%。同时，国外媒体在 2016 年也达到了报道数量的最高峰，说明国家形象自塑具有传播力和成效性。而国外有关黄河流域的新闻报道均在 750 篇以下，表明国外对黄河流域的生态保护和高质量发展的报道还未形成广泛传播，篇数虽远低于国内报道数量，但呈现上升趋势，因此，黄河流域生态保护和高质量发展在一定程度上也引起了相当多国外媒体的关注。



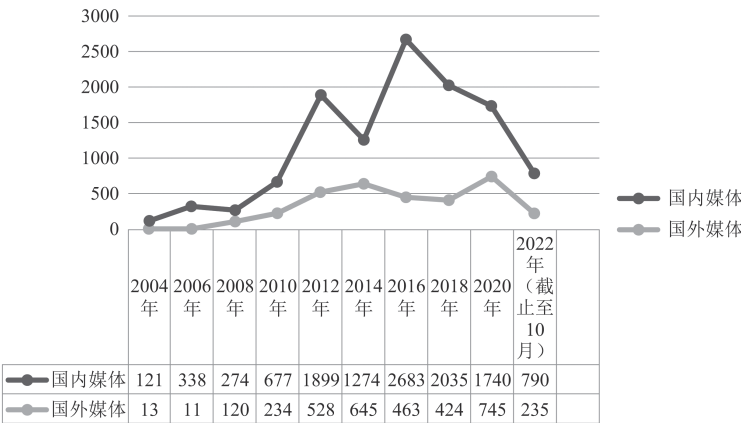


图 2 国内外主流媒体黄河流域新闻语篇数量

5.2 主题词的统计、分析与嬗变

在语料库研究范式中，研究者可以通过词频统计分析语篇的显著特征。通过观察语料库软件 AntConc 3.5.7 生成的主题词表，可以看到黄河流域相关报道的主题。表 1 是以 2012 年为时间分节点，用 AntConc 3.5.7 检索功能得到本语料库的关键词，排除虚词后得到本文使用频率最高的 15 个实词，左侧为 2003—2011 年的语料，右侧为 2012—2022 年语料。

表 1 自建语料库中前 20 个高频实词对比

Rank	Word	Frequency	Rank	Word	Frequency
1	river	168	1	river	415
2	water	101	2	yellow	259
3	yellow	89	3	China	219
4	China	46	4	Xi	125
5	flood	34	5	province	112
6	province	25	6	ecological	101
7	ice	19	7	national	99
8	people	18	8	people	99
9	control	17	9	development	98
10	cubic	15	10	water	97
11	flooding	14	11	protection	95
12	pollution	14	12	central	63
13	level	13	13	quality	63

( 待续 )



(续表)

Rank	Word	Frequency	Rank	Word	Frequency
14	reservoir	13	14	high	56
15	river	11	15	environment	54

从表 1 可以得出,中国媒体有关“黄河流域”报道的高频词,去掉虚词之后,有 flood (34 次)、ice (19 次)、control (17 次)、flooding (14 次)、pollution (14 次) 等实词,这说明在 2010 年前,国内媒体多以报道黄河流域的旱涝灾害、洪水、环境污染等负面新闻为主,例如原文中的“China’s Polluted Rivers”“It’s very tragic”“While that disaster helped spark new public awareness of the extent of the nation’s water pollution”“In China, Water You Wouldn’t Dare Swim in, Let Alone Drink”等,多是对黄河流域 21 世纪初的暴雨、洪水等自然灾害以及土壤破坏、水土流失、物种灭绝等生态问题的夸大描述。

而 2011 年后至今的报道中,排名靠前的高频实词有 ecological (101 次)、people (99 次)、development (98 次)、protection (95 次)、quality (63 次)、environment (54 次) 等,说明以“防汛”“治理”“生态保护”“高质量发展”等为主题的正面报道数量逐渐增多,这一现象出现是因为党的十八大以来“绿水青山就是金山银山”的理念已逐渐成为全社会的共识,这既是国内政策和社会认知的共同作用,也受到国际形象塑造和社会需求的推动。这种话语的变化反映了社会对于生态文明、高质量发展等议题的持续关切,也表达了中国在这些领域取得的一系列成就。黄河流域的形象由灾害频发的“忧患河”逐步向人民的“幸福河”转变。

本研究利用 BFSU Collocator 1.0 搭配强度计算功能,得到主题词的强搭配词。搭配强度的计算方法包括互信息值 (MI 和 MI3)、Z 值 (Z-score)、T 值 (T-score)、Log-log 值、卡方值 (C2)、对数似然比等。由于 T 值计算出的搭配词以高频词为主,而学界一般认为 T 值搭配计算有偏重高频词的问题。T 值的临界值通常是 2 (梁茂成等 2010)。为避免搭配计算中偏重高频词 (高频词以功能词居多) 的问题,英国学者 Adam Kilgarriiff 提出了 Log-log 搭配计算公式,将高频词的权重降低,从而使其他实词搭配词的权重得以提高。因此,本研究在参考 T 值 (T 值 > 2) 与 Log-log 值的基础上,剔除了不符合研究需要的主题词及其搭配词。结果见表 2。

表 2 主题词及其强搭配词

主题词	搭配词	频次	T 值	Log-log 值
Yellow river	development	98	4.3159	2.7319
	high-quality	47	4.8309	5.9997

(待续)

(续表)

主题词	搭配词	频次	T值	Log-log 值
Yellow river	reserve	42	3.327	3.3113
	nature	33	4.4138	3.9299
	delta	32	4.5835	6.5491
	wetland	22	2.152	4.081
national	reserve	42	3.4423	6.5010
	nature	33	2.5632	4.6757
	delta	32	2.5764	4.7481
	Congress	24	4.8094	10.3870
	deputies	16	2.7882	6.1344
development	ecological	101	6.0979	9.8306
	high-quality	48	6.5718	12.2191
	conservation	41	4.4612	4.6810
	prospect	27	3.6710	1.7319
	balanced	13	2.6618	3.5017
ecological	development	98	6.0979	9.8306
	protection	95	5.0537	8.1289
	civilization	19	2.5330	5.1135
	emphasis	14	2.3472	5.6758
	national-level	10	2.1632	6.2721
protection	ecological	101	5.0537	8.1289
	development	98	4.9761	8.0552
	high-quality	48	4.7291	9.0261
	first-grade	16	2.3532	5.7470
	holistic	11	2.1286	4.0122

由表 2 可看出, China Daily 对黄河流域的报道围绕黄河流域三角区 ( the Yellow River Delta ) 战略 ( strategy ) 意义、生态文明 ( ecological civilization ) 现状和发展前景 ( development prospect ) 进行描述。在习近平总书记 ( President Xi Jinping ) 的战略部署 ( strategic disposition ) 和政府各部门的实践下, 以转变经济发展方式 ( the pattern of economic development ) 为根基, 建立自然保护区 ( natural reserve ), 保护水域 ( waterways、wetlands ) 和生物多样性 ( biological diversity、birds ), 最终实现

可持续、高质量发展 ( sustainable、high-quality ), 达到经济与生态平衡 ( balanced )。此外, 在追求区域高质量发展的同时, 应积极探索具有区域特色的新途径。

5.3 主题词的评价系统中态度资源分析

通过 BFSU Collocator 1.0 软件, 可以得到主题词的强搭配词。结合互信息值和对数似然比, 设置左右 Span = 5, 排除虚词后提取主题词的强搭配词。由于本研究中新闻语篇的关键词与黄河流域高质量发展密切相关, 因此, 选取主题词 high-quality 为节点词, 结果见表 3。检索与 high-quality 的高强搭配词, 对提取的搭配词所在索引行中的态度资源进行分类, 以分析黄河流域新闻语篇中塑造的中国形象。

表 3 自建语料库中 high-quality 强搭配词信息

排序	搭配词	互信息值	对数似然比	排序	搭配词	互信息值	对数似然比
1	development	17.2866	912.4604	11	achievements	7.4403	72.1748
2	ecological	16.3554	670.0132	12	ensure	7.6698	67.7689
3	Yellow	14.7496	503.6872	13	fuel	7.2196	53.2846
4	River	14.0042	462.0016	14	pursuing	6.1425	50.4207
5	protection	15.1551	440.2677	15	vital	6.1127	44.5722
6	new	13.0113	197.7767	16	ecological	6.1018	39.1024
7	promote	10.7277	83.8657	17	national	5.9281	37.7527
8	central	9.3883	87.7338	18	people	5.9233	35.6870
9	advance	8.8808	73.3896	19	adopt	4.3256	32.9342
10	enhance	8.0253	77.3869	20	further	3.5441	28.5621

主题词表中的词汇除表达信息之外, 有一些还具有评价意义。例如形容词 vital ( MI 值 = 6.1127 )、central ( MI 值 = 9.3883 ); 名词 development ( MI 值 = 17.2866 )、difficulty ( MI 值 = 2.1325 ); 动词 fuel ( MI 值 = 7.2169 )、ensure ( MI 值 = 7.4403 )等。

Martin 指出, 在分析边缘性的评价表达时不能局限于具体内容, 要考虑其所处的语境 ( Martin & Rose 2003 ), 而语料库工具中的索引行功能可以为态度资源的分析提供上下文信息。本研究以主题词作为节点词, 分别从语料库中抽取与其强搭配词的具体搭配和共现关系, 并从中提取描述主题词且具有态度含义的词语, 最后依据 Martin 评价理论在情感、判断和鉴赏下的子系统判断标准进行分类。以表 3 中的 development ( 发展 ) 为例, 提取出描绘主题词 high-quality 并具有评价意义的词汇, 即 ecological、protection、promote、coordinate、harmonious、firmly、sustainable、boost、

balanced、pursuing、demand、effectively 和 adopt。依据 Martin 的评价理论框架分类体系，子系统的分类情况如表 4 所示。

表 4 development 态度资源的分布情况

态度资源	子资源	示例
情感	意愿性	expect, demand
	安全性	firmly
判断	能力性	promote, boost, protection, pursue, adopt, boost
	可靠性	sustainable
鉴赏	反应性	further
	构成性	coordinate
	评估性	balanced, harmonious, ecological

根据 Martin & White (2003) 的评价词汇表，以及表 3 关键词的搭配情况，可将表 4 中具有评价意义的主题词进行分类（见表 5），结合索引行分析新闻中的态度资源。从态度资源种类来看，China Daily 中有关黄河流域的报道大部分使用了鉴赏资源和判断资源，且均为正面词汇。

此外，同样检索其他搭配词与主题词 development 的共现行，挑选出描绘主题词的评价词，再进行子系统分类。本研究选取文中两个典型的主题词为节点词，即 high-quality 和 development，按上述方法提取具有评价意义的词语，最终得到以下结果，如表 5 所示。

表 5 主题词的评价词语分类

态度资源	子资源	示例	百分比/%
情感	意愿性	expect, demand, need	13.2%
	安全性	faithfully, firmly, ready	
	满意性	achieved, highlighting	
判断	能力性	promote, boost, pursuing, adopted, boosting, successfully, advance, improving, enhance, expanding, create, emphasizing, enhance, stimulate, change, deepen, facilitating	38.2%
	可靠性	sustainable, support, stability	
	真实性	truly, credit	
	常规性	new, traditional, national	
	恰当性	integrated	

（待续）

(续表)

态度资源	子资源	示例	百分比/%
鉴赏	反应性 构成性 评估性	further, better, upgrade, crucial coordinate, jointly, maintained, connectivity, concerted, rejuvenation balanced, harmonious, ecological, comprehensive, benefits, effective, profound, strategy, unique, higher level, pleasant, significant, green, potential, growth, major, robust, important, largest, opening, systematically, protection	48.6%

总体来看, China Daily 在对黄河流域进行报道时, 情感类词汇使用较少, 仅占 13.2%。在评价系统中, 情感资源最能直接表明自身态度和立场, 较少使用情感资源则表明个人情感态度在新闻语篇中占比低, 即论述新闻的角度较为客观, 具有一定的客观性。对客观性的遵循, 是一种理念、一种职业原则、一种传播态度, 同时也是英语新闻写作中的重要方式(王海峰 2016)。判断类词汇占 38.2%, 且都是积极、正面词汇。鉴赏类词汇最多(48.6%), 肯定了十八大以来黄河流域发展迅速的积极意义。China Daily 代表客观真实的态度, 打造联通中外的窗口, 建构正面的国家形象, 向外界报道黄河流域现阶段的发展情况, 促进黄河流域的形象转变, 从语言层面倡导构建积极的话语环境。

5.3.1 情感系统分析

情感系统是三个子系统的核心, 因为它是我们与生俱来的表达资源(Painter 2003)。具体来说, 情感系统用来解释语言使用者对行为、文本(或过程)、现象做出的情感反应。通过分析语料发现, 情感多由对社会事实的正面报道出发。下面通过例(1)至例(3)对情感资源进行分析。

(1) While highlighting progress made in past years of protecting the Yellow River, Xi also pointed out difficulties and problems that still remain, including a fragile ecological environment, severe water preservation situation and development quality that needs to be improved.

(2) Xi called for focusing on the problems, faithfully implementing the national strategy of promoting ecological protection and high-quality development of the Yellow River Basin, and firmly pursuing the green development path.

(3) In 2020, China’s environmental protection tax receipts totaled 20.7 billion yuan, he said, while tax reductions and exemptions for emitters that achieved low emissions reached 10 billion yuan.

例(1)来自2019年黄河流域生态保护和高质量发展座谈会的报道。highlighting一词属于情感子资源中的满意性资源,表达出中国开展了大规模的黄河治理保护工作,并取得了举世瞩目的成就。即使如此,中国坚持实事求是,积极面对目前黄河流域仍存在的一些突出困难和问题。本句中needs一方面引出了黄河流域生态发展的重要性;另一方面,充分展示了当前任务将持续进行,尽管问题存在,但并非不可解决的永久性难题。因此,情感评价词具有意义传递功能,能够将不同的语境引领到积极的话语空间中,营造出积极的话语氛围。

例(2)来自2021年习近平总书记在山东省济南市主持召开深入推动黄河流域生态保护和高质量发展座谈会的报道。本句中faithfully和firmly分别译为“深入地”和“坚定不移地”,均从积极情感角度表达出中国推进黄河流域高质量发展的决心,锚定了深入贯彻落实国家生态保护和黄河流域高质量发展战略和绿色发展道路的战略方向,体现出黄河流域高质量发展已提升到国家战略的高度,向受众传递黄河流域的重要性和黄河地区发展的必要性。

例(3)来自对2022年生态环境部举行2月例行新闻发布会的报道。词语achieved意为“达到,取得”,带有明显情感色彩,是主体对客体做出的正面评价。2020年,中国环境保护税收收入总计207亿元,而对低排放企业的税收减免达到100亿元。在上述语境中体现出国家对有利于环境保护的企业降低税收减免的行为予以高度评价,表达了该项措施对环境保护具有实际意义。这一方面说明了黄河流域政策得到了积极落实,并取得了相应成果;另一方面也向受众传递了乐观、向好的情感态度,为黄河流域形象对外传播构建了客观、积极的话语空间,塑造了中国积极践行高质量发展的形象。

通过分析情感资源,不难看出,在《中国日报》对黄河流域的相关对外报道中,以积极情感为侧重点,为受众创造出显化的情感语境,传递了正面、客观、积极的信息。首先,中国在促进黄河流域环境保护工作中取得了很大进展,取得了举世瞩目的成就,黄河流域经济社会发展和百姓生活发生了很大的变化。其次,在取得成就的同时,黄河流域也依然存在生态环境脆弱、水资源保护形势严峻的问题,发展质量有待提高。最后,我们拥有坚定的信心,能够保持历史耐心和战略定力,努力让黄河造福人民。

### 5.3.2 判断系统分析

判断是根据一定的评价标准对被评价者做出的价值评价,是根据伦理道德的标准来评价语言使用者的行为。判断涉及对行为的态度,我们赞赏或批评,赞扬或谴责(Martin 2004)。判断指语言使用者依据特定的社会规范对某些行为方式做出自己的评价,包括社会评判和社会约束两方面(王振华 2001)。通过数据分析,判断资源中的能力性词汇占比最大,如promote、boost、enhance、deepen、create、promote、successfully、improving和stimulate等,多表达“深化”“加强”“促进”“提高”“成功”“创造”等积极意义。下面通过例(4)、例(5)对判断资源进行分析。



(4) He called for concerted and coordinated efforts to strengthen ecological conservation, promote better use of water resources, ensure safety from flooding along the riverbanks and advance high-quality development of the basin.

(5) “In Jincheng, coal-bed methane exploration goes along with the intelligent operation of coal mines and mining upgrades for better environmental protection,” Dong said. “So its practice in this sector is of great significance to boosting production safety and efficiency, as well as improving the environment.”

例(4)来自2019年黄河流域生态保护和高质量发展座谈会的报道。其中strengthen、promote、ensure和advance 4个词语都属于判断系统中的能力性子资源,在本语境中分别译为“加强”“提升”“保障”“推动”,具有正面评价意义。一方面它体现了中国在推动黄河流域高质量发展过程中从多方面入手,脚踏实地,统筹发展,对维护社会稳定、促进民族团结具有重要意义;另一方面,中国为维护生态环境采取的措施为全球环境治理做出了重要贡献,体现中国致力于绿色发展、构建人类命运共同体的大国担当,展示出大国应有责任意识。

例(5)来自2022年十三届全国人大五次会议山西代表团举行全体会议的报道。其中,upgrades、boosting和improving分别借助environmental protection、production safety and efficiency以及the environment产生积极话语意义。在判断系统中,新闻媒体语言的研究把评价系统分为社会评判和社会约束两部分(Martin 1998),社会评判分正面含义和负面含义(Martin 1998)。正面含义理应获得人们赞扬,负面含义应当受到社会谴责。社会约束也分正面和负面。正面含义是表扬性的,负面含义是谴责性的。例句中upgrades、boosting、improving判断子资源的利用展现了中国主动进行经济转型升级的积极性,具有正向话语意义。

通过分析判断系统资源,不难发现China Daily在对黄河流域的相关报道中传递出中国在进行治理黄河流域的过程中,并非仅局限于改善生态环境,而是从多方面入手。例如,黄河流经区域努力实现采掘智能化,推动现代科技与煤炭产业深度融合,实现环境保护的同时提升效能,推动煤炭行业走绿色智能发展之路,直接与间接举措并行,多管齐下,为黄河流域形象构建主动积极的一面,在对外传播中彰显了中国致力于推动全球环境治理改革的坚定决心和正面形象。

### 5.3.3 鉴赏系统分析

鉴赏指语言表达者依据美学标准对事物或一些行为进行评价(Martin & Rose 2003)。该系统围绕三方面展开,即反应性、构成性和评估性(Martin & White 2005)。在鉴赏资源中,表示评估性的词汇出现次数最多。评估指用社会标准来看文本或过程时,判断文本或过程是否重要、是否有价值。其中,反应性包括影响与质量,如further、better、upgrade和crucial,描绘出事物所带来的积极影响。构成性是指组成

成分是否平衡和复杂,如 coordinate、jointly、maintained、connectivity 和 concerted,能够体现出事物在推动过程中的平衡性。价值指用社会标准判断文本或过程是否重要,如 harmonious、comprehensive、benefits、effective、unique、significant、major、robust 和 important,能够描述出事物的重要性和积极性。下面通过例(6)、例(7)对鉴赏资源系统进行分析。

(6) China will tap the potential of financial instruments in enhancing environmental protection during the 14th Five-Year Plan (2021-25) period, with climate mitigation as one of the priorities, according to the Ministry of Ecology and Environment.

(7) He called for concerted and coordinated efforts to strengthen ecological conservation, promote better use of water resources, ensure safety from flooding along the riverbanks and advance high-quality development of the basin.

例(6)来自 2022 年对生态环境部举行 2 月例行新闻发布会的报道。其中, potential 译为“潜在的”,仅从意义层面无法直接判断其类别,因为词汇个体在评价研究中的作用有限,必须把词汇研究和语法研究以及语境研究紧密结合起来,才能对评价意义的解读获得比较全面的认识(朱莉莉、朱永生 2019)。句中 potential 与 enhancing environmental protection、priorities 等共同出现,用于搭配“挖掘运用金融工具促进环境保护的潜力”,含有对这一现象的主体评价,故属于“态度—鉴赏—评估”,可以有效提高受众的评估意识。这也从鉴赏的角度正面评价了中国对于环境保护采取了多重措施,并且发挥了重要作用。

例(7)来自 2019 年黄河流域生态保护和高质量发展座谈会的报道。concerted 和 coordinated 在该语境中均可译为“协调一致的”“统筹的”,虽是描述对黄河流域发展的讨论,但通过上下文不难感知到话语主体构建了和谐的话语空间,体现出黄河流域治理需坚持多生态类型统筹的理念,同时要协同区域发展,形成合力。句中 concerted、coordinated、strengthen、promote、ensure 和 advance 从鉴赏的角度正面评价了黄河流域高质量发展对保护环境的影响。

上述分析表明,在黄河流域发展的过程中,中国政府采取多层次措施,全面推动黄河流域高质量发展,并呼吁区域间统筹发展,形成合力。中国媒体向世界展示出与全球同呼吸共命运的东方大国形象,中国在推动自身经济发展的同时,也在关心全球环境问题,致力于建设和谐的社会。

#### 5.4 黄河流域文化形象的对外传播对策

国家形象是一个国家对自己的认知与国际体系中其他行为体对它的认知的结

合,是一系列信息输入和输出产生的结果,是一个“结构十分明确的信息资本”(Boulding 1956)。国内学者也指出,国家形象是指国家的客观状态在公众舆论中的投影,也就是社会公众对国家的印象、看法、态度、评价的综合反映,是公众对国家所具有的情感和意志的总和(刘小燕 2002)。树立良好的国家形象在复杂的国际合作与竞争中具有重要的战略意义。构建国家形象主要有两种方式,即内部的“自塑”和外部的“他塑”,而在这一过程中,多模态传播扮演着重要的角色。在调查黄河流域在全球范围内树立形象的过程中,笔者发现,现有推动对黄河流域形象对外话语构建的措施还存在一定不足。有鉴于此,针对黄河流域形象对外构建提出以下分析与对策。

#### 5.4.1 “自塑”: 时事文化相融合, 传播中华优秀传统文化

一个国家的政府和各个阶层通过具体的行为来塑造国家形象,就是国家形象的“自塑”(刘小燕 2002)。在“自塑”过程中,国内媒体扮演着重要的角色,需要通过新闻报道向国内国际传递出国家形象。国内对黄河流域的新闻报道还存在一些局限性,如主题丰富性较弱、传播范围有限、普及度较小等问题。因此,应该系统规划国家黄河话语对外传播内容体系,丰富黄河流域报道主题,促进新闻语篇的多元性,打造垂直、立体的黄河流域外传播内容体系,在满足多元受众的个性化需求的同时,努力推进黄河流域文化的海外传播,助力中国文化“走出去”。此外,还须充分了解受众读者背景,提高新闻语篇可读性。在国内新闻向外界传播的过程中,难免会造成国外读者的“误读”与“歪解”。因此,国内媒体在编译对外传播新闻语篇的同时,要充分了解国外受众的文化背景、历史情况、阅读偏好和表达习惯,进而有机融入目的语国家的话语体系,以贴近读者习惯的方式更“地道”地进行叙事,释放原文本活力。

#### 5.4.2 “他塑”: 促进传播渠道融合, 扩大新闻覆盖面

树立国家形象的过程中,除了“自塑”,“他塑”也十分具有必要性。中国有学者认为,外国人在认识、判断中国之前有预设的标准,他们会按照自己的价值观、世界观、利益取舍和善恶感等来塑造自己心目中的中国形象,而这一国家形象的形成过程可以称之为“他塑”(赵雪波 2006)。因此,“他塑”在一定程度上更具间接性和困难性。

由于新闻具有多重性,中国不断被西方媒体无端抹黑,西方媒体、政客试图将中国“他塑”成“崩溃、落后”的刻板印象。通过语料库软件 AntConc 统计 2002 年至 2022 年的时代周刊、每日邮报、路透社、美国有线电视新闻网等国外媒体新闻语料的高频词,其中不同形式的 flood 共计出现 300 多次, pollution、dangerous 等负面词汇也使用超过 50 次,可见国外依然在树立黄河流域“灾河”“害河”的负面形象,与当前推进黄河绿色治理、建设人民“幸福河”截然相反。因此,构建完善的对外传播路径十分必要。

首先,应加强与国外媒体的沟通交流和合作,提升黄河流域国际新闻报道的广度。截止到2023年2月9日,China Daily共计3,021,007篇报道,其中标题包含the Yellow River的报道仅有958篇,占比0.03%。受搜索结果显示限制,新华网未展示出新闻总数,但新闻总数已超过10,000篇,标题包含the Yellow River的报道仅有237篇。关于中国对外的黄河流域的国际新闻报道路径与宣传力度十分有限。在此背景下,各部门应积极主动地采取措施,加强与国外媒体的沟通与合作,减少新闻“空白地带”,打通新闻壁垒。

其次,应创新传播思路,挖掘黄河流域国际新闻报道的深度。在进行黄河流域的国际新闻报道时,思路通常较为生硬,以“硬”输出为主,缺乏时代性和“软”融合。直接或间接传播黄河流域新形象,提升话语传播对外接受度,打破传统界限,探讨新型传播模式迫在眉睫。

#### 5.4.3 黄河流域对外媒体形象的显性传播与隐性传播

黄河流域的对外形象传播既是宣传中国传统文化、历史和文明的重要途径,也是推动中国与世界各国交流合作、增进相互了解和友谊的重要手段。

黄河流域的显性传播形式主要包括政治、经济、文化等领域。其中,黄河文化是具代表性的文化传播形式。黄河流域作为中国古代文明的重要发祥地之一,具有丰富的文化遗产和历史文化符号,是中国向外传播文化的重要载体。此外,黄河流域生态文明及高质量发展是构成国家形象的重要组成部分,成为黄河流域向外传播的重要内容。例如,黄河流域的灌溉技术和水利工程在古代就已经达到了一定水平,在推动中国古代农业发展、维护国家安全和社会稳定方面发挥了重要作用。显性传播的内容能够直接被外界所认知,从而增强黄河流域的国际知名度和影响力。

黄河流域的隐性传播方式主要表现为文化符号、态度价值观等隐含信息的传递。例如,黄河流域的传统文化中蕴含着尊重自然、敬畏历史、勤劳朴实等价值观,这些价值观不仅是黄河流域地区的文化特征,也是对外话语传播中的隐性信息。此外,黄河流域人们的生活方式、饮食习惯、服饰风格等也能够向外界传递文化信息。这种隐性传播的方式并非简单的信息传递,更多的是对文化、历史、价值观的传递。这些隐性传播的内容,能够传递出黄河流域的文化特色和精神内涵,推动文化多元化和交流互鉴,促进文明进步。

黄河流域的显性传播和隐性传播相互融合,形成复杂的传播网络。这种融合不仅能够丰富对外话语传播的形式和内容,也能够增强其传播的深度和广度。例如,黄河流域的文化、历史和科技成就不仅在经济、文化领域得到了显性传播,也通过文化符号、行为方式、态度价值观等隐性传播方式被传递出去。这些内容相互交织、相互渗透,形成了一个复杂的传播网络。这种多元化的传播形式和内容丰富了中国文化的内涵和外延,使其更具吸引力和影响力。此外,黄河流域对外话语传播的多样性和复杂性也反映了中国在全球化进程中的发展和变化,这也为构建新时代中国



形象起到了借鉴作用。因此,研究黄河流域对外话语传播的显性传播、隐性传播和相互融合,对于深入了解中国传统文化和推动文化交流互鉴具有重要意义。

#### 5.4.4 推动传播与技术融合,促进黄河流域多模态形象构建

区域形象的话语构建已不再停留于单一的语言层面,特别是随着新媒体的发展,信息传播渠道和人类交际方式很大程度上得到改变,多模态表征方式助力话语传播更加多样化、纵深化(李俊义、苏炎奎 2022)。在此背景下可以根据不同国家受众的不同特点选择最恰当的媒体形式,从而实现黄河流域文化与现代信息的有机融合,实现对受众的全面覆盖及最佳传播效果,而传播媒介由于是信息的传输工具,承担着向大众传递外部环境信息和内外信息交换的主要功能(刘小燕 2002),媒体能够在一定程度上影响着大众对于外界的看法,因而全媒体时代可能会对外部产生所谓我国的“歪曲”形象,即“实际”国家形象和“媒体”形象的错位。因此信息技术发展迅速的背景下,黄河流域文化的对外媒体形象传播可以采取以下策略。

突破单一的语言模态,实现多渠道、多模态的国际形象传播。随着信息技术发展,话语实践呈现出新的特点,话语意义不再局限于单一的语言层面,而是同时通过多种语言模态来表达,例如,在北京冬奥会开幕式中,一滴蓝色水墨从天而降,化为黄河之水,倾泻而下,用震撼的视觉效果让国内外观众直观地感受到了“黄河之水天上来”的磅礴气势。同时,又能以“低碳、环保”的方式诠释出中国人对水与江河的崇敬。在对外传播过程中利用多模态形式,让黄河流域形象更加生动立体,让“黄河故事”在世界范围内传播。

大力推动黄河流域区域相关的影视作品发展。黄立鹤、吴贇(2021)指出,情境化的影视片借助画面、声音等综合视听的形象符号体系来塑造形象,具有直观在场性、动态瞬时性、双重文化性等鲜明特质。推动打造黄河流域相关的影视作品或记录宣传片能够利用视听冲击与艺术熏陶,向全球观众展示出黄河流域的形象变化和未来发展,同时向全世界传达了新时代黄河流域特有的区域魅力,构建新时代的大国形象和话语权。

## 6 结语

新闻语篇中深层次的评价意义能够间接影响读者的观点,因此新闻语篇具有隐含性。语料库研究方法可以进行大规模定量研究,一定程度上突破了传统研究方法的局限性。本研究发现在黄河流域的对外新闻语篇中,态度资源主要以判断资源和鉴赏资源为主,并通过评价词汇为受众提供积极的语境和心理因素,发挥了语言的社会功能。研究表明,中国主流媒体《中国日报》对黄河流域的报道多倾向于使用鉴赏资源、判断资源,而较少地使用情感资源,体现出新闻报道的客观性,同时又通过正向情感词汇表达出中国在推动黄河流域高质量发展过程中始终保持实事求是的态度。此外,判断资源中能力性词汇使用最多,表明中国在实现将黄河从“忧

患河”(filthy waterways/poisonous river)逐步转变为“利民安澜河”(attractive river); 鉴赏资源中评估性词汇使用最多,肯定了黄河流域高质量发展意义的同时,也展示了现阶段黄河流域治理和发展的成果。

中国主流媒体以黄河流域为话语中介,以环境保护、绿色可持续发展为载体,为黄河流域新闻语篇创造出积极语言环境,同时传递出中国为维护生态环境采取的措施对全球环境治理的重要贡献,以及致力于绿色发展、构建人类命运共同体的坚定信念,建构了“有责任、有担当、充满正能量”的大国形象,有利于对外传播中国致力于推动全球环境治理改革的正面形象。

### 参考文献

- BOULDING K. The image: knowledge in life and society[M]. Ann Arbo: University of Michigan Press, 1956.
- PAINTER C. Developing attitude: an ontogenetic perspective on appraisal[J]. Text & Talk, 2003, 23(2): 183-209.
- FIRTH J. Papers in linguistics 1934-1951[M]. London: Oxford University Press, 1957.
- MARTIN J. Lectures of academic register and genre[D]. Sydney: University of Sydney, 1998.
- MARTIN J. Positive discourse analysis: solidarity and change[J]. Revista Canaria de Estudios Ingleses, 2004, 49: 179-200.
- MARTIN J, ROSE D. Working with discourse: meaning beyond the clause[M]. London & New York: Continuum, 2003.
- MARTIN J, WHITE P. The language of evaluation: appraisal in English[M]. New York: Palgrave Macmillan, 2005.
- 曹德春. 跨文化视角下黄河生态文明建设与国际传播策略[J]. 新闻爱好者, 2020 (3): 32-34.
- 陈梅, 文军. 评价理论态度系统视阈下的白居易诗歌英译研究[J]. 外语教学, 2013 (4): 99-104.
- 胡壮麟. 积极话语分析和批评话语分析的互补性[J]. 当代外语研究, 2012 (7): 3-8.
- 黄立鹤, 吴贇. 基于贴真体验与建模的多模态口译教学语料库构建及应用[J]. 外语教学理论与实践, 2021 (4): 127-136.
- 焦丹, 苏铭. 黄河文化国际传播话语体系构建与实践路径探索[J]. 新闻爱好者, 2022 (1): 42-44.
- 李俊义, 苏炎奎. 多模态话语互动路径下的区域形象构建研究——以福建形象宣传片为例[J]. 外国语言文学, 2022 (4): 41-52.
- 梁茂成, 李文中, 许家金. 语料库应用教程[M]. 北京: 外语教学与研究出版社, 2010.
- 刘小燕. 关于传媒塑造国家形象的思考[J]. 国际新闻界, 2002 (2): 61-66.



- 王海峰. 英语新闻写作中的语义客规性[J]. 青年记者, 2016 ( 6 ): 60-61.
- 王佳, 秦娜. 新媒体时代城市形象构建与文化表达——基于兰州市受众调查分析[J]. 中国报业, 2019 ( 24 ): 12-13.
- 王振华. 评价系统及其运作——系统功能语言学的新发展[J]. 外国语, 2001 ( 6 ): 13-20.
- 吴翔, 付邦道. 开封旅游形象策划与构建[J]. 地域研究与开发, 2003 ( 5 ): 66-69.
- 邢祥, 邢军. 新时代黄河文化传播创新路径研究[J]. 新闻爱好者, 2020 ( 3 ): 29-32.
- 徐国亮. 提升中国国际话语权的深刻意蕴与重要意义[J]. 人民论坛, 2021 ( 31 ): 12-16.
- 许家金. 多因素语境共选: 语料库语言学新进展[J]. 外语与外语教学, 2020( 3 ): 1-10.
- 许家金. 人文社会科学研究语料库语言学路径[J]. 中国外语, 2023 ( 1 ): 12-19.
- 尤元学. 黄河流域中心城市历史文化形象的传播构建——以《大河之畔》为例[J]. 山东青年政治学院学报, 2021 ( 5 ): 112-116.
- 赵雪波. 关于国家形象等概念的理解[J]. 现代传播, 2006 ( 5 ): 63-65.
- 中共中央国务院印发黄河流域生态保护和高质量发展规划纲要[J]. 中国水利, 2021 ( 21 ): 3-16.
- 朱莉莉, 朱永生. 系统功能语言学对语言使用者的研究: 从 Halliday 到 Martin[J]. 外语教学, 2019 ( 6 ): 54-56.
- 朱永生. 积极话语分析: 对批评话语分析的反拨与补充[J]. 英语研究, 2006 ( 4 ): 36-42.
- 杨林秀, 赵浩. 希望的话语: 国内积极话语分析研究述评[J]. 北京科技大学学报, 2018 ( 1 ): 35-41.

**通信地址:** 250101 山东省济南市 山东建筑大学外国语学院

# 基于语义韵强度算法的中国形象建构研究<sup>\*</sup>

苏州大学 汤慧桃 王 军

**提要:** 本研究使用情感分析技术,基于扩展意义单位模型,对 Coronavirus 语料库中 500 个以 China 为节点词的索引行进行了语义韵强度计算,通过基于数据的定量分析方法来阐述中国形象。研究发现,总体而言,外媒眼中的中国呈现一个积极正面的国家形象,但对中国的消极评价力度则远远高于积极评价。

**关键词:** 语义韵强度、中国形象、扩展意义单位、情感分析

## 1 引言

中国形象建构研究是当今学界极为关注的研究课题,通过文献检索发现,采用扩展意义单位理论以及基于数据的定量分析方法可以用来阐述中国形象建构问题。因此,本研究基于 Coronavirus 语料库,选取 500 个以 China 为节点词的索引行,利用情感分析技术进行语义韵强度计算,从而揭示不同国外主流媒体眼中中国在抗疫过程中所形成的国家形象,这对于更加准确全面地把握中国形象的内涵,不断扩大中国的国际影响力,具有非常重要的现实意义。

## 2 文献综述

### 2.1 扩展意义单位与语义韵

本研究聚焦中国形象建构,并在扩展意义单位的理论框架下展开国家形象研究。Sinclair 于 1996 年首次提出了扩展意义单位的概念,并建立了扩展意义单位模型 (Sinclair 1996),在语料库语言学领域被广泛应用。该模型将扩展意义单位分成节点词、搭配、类联接、语义倾向和语义韵 5 个成分,其中“节点词”指词项作为整体出现的证据,“搭配”是节点词周边所形成的词汇型式 (Sinclair 2004: 141),“类联

<sup>\*</sup> 本文系苏州大学第二十五批大学生课外学术科研基金重大项目“基于语义韵强度算法的中国形象建构研究”(KY2023006Z)阶段性研究成果。汤慧桃为本文通信作者。

作者贡献:

汤慧桃:选题构思、研究方法、数据收集、数据分析、初稿撰写、讨论结论、字数占比(90%);王军:修改润色、字数占比(10%)。

接”指“在句法结构中语法范畴间的相互关系”(Firth 1968: 183),“语义倾向”是节点词周边所形成的语义集(Sinclair 1996)。

在扩展意义单位中,语义韵是确立扩展意义单位边界的关键性因素(李晓红、卫乃兴 2012),亦是语料库语言学迄今发现的重要语言运作机制之一。语义韵是指由节点词的互文单词之间的语义关联性所建立起来的一种和谐意义氛围,用于表达说话者对某种语用情境的态度(Louw 2000: 57)。Sinclair (1996: 87)认为,语义韵兼具语义和语用功能,既可揭示词项结伴行为的语义特征,又可维系意义氛围与语境态度的和谐。

基于频数的极性比值法是现阶段语义韵强度计算的最常见方法,分为两大类。一种基于搭配词频数(Ellis *et al.* 2009),使用三个计量指标来测量某一动词的韵律配价与强度:积极搭配词的形符数( $T_{\text{positive}}$ ),积极搭配词形符数与积极、中性、消极三类搭配词总形符数的比值( $T_{\text{positive}}/T_{\text{positive}} + T_{\text{neutral}} + T_{\text{negative}}$ ),以及积极搭配词形符数与消极搭配词形符数的比值( $T_{\text{positive}}/T_{\text{positive}} + T_{\text{negative}}$ );另一种基于索引行个数(Wei & Li 2014),计算节点词表达积极或消极态度意义时的索引频数与节点词总频数的比值( $F_{\text{concordance-positive(or negative)}}/F_{\text{index-all}}$ )。这两种基于频数的极性比值法得到了一些学者的响应,如陶源、胡谷明(2015)基于俄汉平行语料库进行致使动词语义韵翻译研究;孙莉(2017)通过对比 affect 和 influence 两词探讨中国英语学习者近义词使用情况的语义韵新解;胡雪婵、吴长安(2018)探讨汉语成语语义韵的冲突及其消解问题;高歌、卫乃兴(2019)计算《红楼梦》英译本语料库中的语义韵;李华勇(2020)比较汉译口语与原创汉语口语中“被”字句的语义韵强度等。

以上研究赋予了语义韵更为细致的态度意义,但未将语义韵与扩展意义单位相结合,且上述两种基于频数的极性占比的强度计算方法存在明显缺陷:默认同一极性范畴的搭配词具有等值的情感强度,即在计算极性占比时,将所有表达某一极性态度)的搭配词均赋予同等值,却未对每个搭配词所持态度的具体程度进行认定和区分。例如, fuel 的搭配名词包括 speculation、dissatisfaction、anger、resentment 和 hate 等(李晓红、卫乃兴 2012: 25),虽然这些词整体呈现“消极心理或情绪”的语义特征,但不同语境的交际意图不同,说话者的态度亦不同,即便属于同一极性(即消极)范畴,其态度等级也是有差异的,即它们所反映的消极程度是不同的。因此在测量语义韵强度时应考虑搭配词的态度等级,即使被划分到同一极性类别的搭配词或索引行,其程度也会不同,不适合做等值处理。因此可以借助自然语言处理领域的情感分析技术,通过情感词库,进一步对每一个情感词进行赋值,从而对语义韵强度进行更为精细的测量。

## 2.2 情感分析

情感分析(Sentiment Analysis)是指利用自然语言处理的文本挖掘技术,对带有

情感色彩的文本进行抽取、分析和处理,从而发现潜在的问题以用于预测或改进 (Yang *et al.* 2021)。目前借助情感词库是情感分析的主要方法之一 (王婷、杨文忠 2021),该方法的基本流程为:首先创建或采用现有的适合某一特定语域的情感词库,该词库会收录文本中可能出现的态度词和情感词,并给每个词分配一个先验值;然后在所研究文本中搜寻表达态度意义的词,并基于词典给予情感值;最后计算得出文本语料的整体情感强度。

情感词库 *syuzhet* 在 Jockers 指导下由内布拉斯加文学实验室 (the Nebraska Literary Lab) 开发。该词库是一个  $10,748 \times 2$  的数据框,包括 10,748 个情感词,有 *word* (词) 和 *value* (情感值) 两列,其中负面情感词为 7,161 个,正面情感词为 3,587 个。在 R 中调用情感词库 *syuzhet* 后所得结果见图 1。

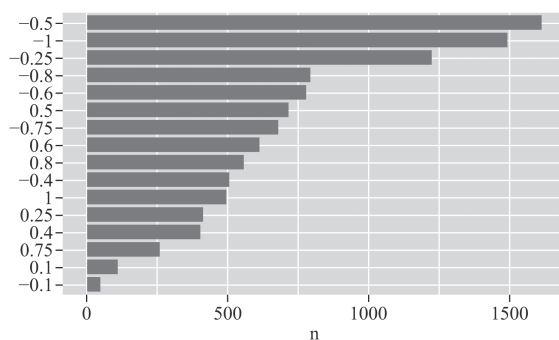


图 1 情感词库 *syuzhet* 中各类情感词情感值

*syuzhet* 库的情感值有 16 类:  $-1.00, -0.80, -0.75, -0.60, -0.50, -0.40, -0.25, -0.10, 0.10, 0.25, 0.40, 0.50, 0.60, 0.75, 0.80, 1.00$ , 因而 *syuzhet* 库的尺度范围很大。在各个情感值的分布中,包括情感词数量最多的三个类别是负面情感类,情感值在  $-0.5$ 、 $-1$  和  $-0.25$  上的情感词数依次为 1616、1493 和 1229。包括正面情感词数最多的三个类别是  $0.5$  (719 个)、 $0.6$  (614 个) 和  $0.8$  (561 个)。包括情感词数最少的三个类别是  $0.75$  (260 个)、 $0.1$  (113 个) 和  $-0.1$  (52 个)。鉴于 *syuzhet* 库的情感值尺度范围很大,具有一定的区分度,且情感词涵盖个数众多,可以作为本研究的情感词库。

中国知网搜索显示,目前使用情感分析技术进行语义韵强度计算的研究并不多见,仅有李晶洁等 (2022) 以网购评论为研究语料,使用情感分析技术计算语义韵强度,得到“衣服”一词的语义韵分布和语义韵强度,但该研究语料语言仅限于中文,文本类型仅限于网购评论,语料涉及领域仅限于服装领域。文中也提出,在后续研究中,需要把情感分析技术应用于不同领域、不同文本类型的语义韵强度计算,进一步检验辅以情感分析技术的语义韵强度测量的适用性和普及性。

基于以上不足,本研究将语言语料、文本类型、语料涉及领域等进行延伸,聚

焦英语语料库 Coronavirus, 将文本类型延伸到新闻报道, 以“China”为节点词, 在扩展意义单位的理论框架下, 以包含 China 作为节点词的多条语料为研究对象, 借助情感分析技术, 通过调用情感词库 syuzhet, 计算节点词 China 在多个扩展意义单位中的语义韵强度, 探寻不同主流外媒眼中的中国抗疫形象, 借助“他者之语”呈现疫情背景下中国在政治、军事、外交、经济、疫情防控等领域表现如何, 并对中国形象的对外传播提供一定启示。

### 3 研究设计

#### 3.1 研究问题

本文的研究问题如下:

(1) 在该语料库中, China 一词总体语义韵强度如何? 不同语义倾向与类联接下, China 一词的语义韵强度如何?

(2) China 一词总体语义韵强度以及不同语义倾向与类联接下 China 一词的语义韵强度折射了怎样的中国形象?

#### 3.2 语料来源

本文的语料来源是 Mark Davies 教授设计开发的在线语料库 Coronavirus, 该语料库收集了自 2020 年 1 月 1 日至 2022 年 12 月 31 日新冠肺炎疫情暴发以来的 20 个英语国家在线报刊文章, 内容涵盖与新冠肺炎疫情相关的社会、文化、经济、卫生健康和政治报道, 是迄今为止最为全面的新冠肺炎新闻语料库(刘鼎甲 2021), 并且每天以三四百万形符的规模持续增长, 目前已于 2022 年 12 月 31 日停止更新。

#### 3.3 语料收集

为保证语料的时效性, 研究者从在线语料库 Coronavirus 中按日期检索, 提取 2022 年 12 月 1 日至 2022 年 12 月 31 日的语料, 以 China 为节点词展开检索, 并将自动选择索引行个数设定为 500, 对 500 个索引行(含所处语境)进行人工预处理, 形成自建语料库 subcorpus, 共计 1,543,500 形符。

#### 3.4 语料分析

本研究的语料分析步骤如下:

(1) 将 subcorpus 导入 R 软件, 方便后期调用情感词库 syuzhet 进行处理。

(2) 根据搭配词的语义特征, 人工分析 500 个以 China 为节点词的索引行(含所处语境)的语义倾向, 并根据每一类语义倾向, 概括每类语义倾向向下最频繁出现

的搭配型式，抽象为类联接。

(3) 对每类语义倾向和每类语义倾向下的类联接进行提取，同时在 R 软件中调用情感词库 *syuzhet*，并对每类语义倾向和每类语义倾向下的类联接进行处理，完成情感词识别。

(4) 用以下三个公式（李晶洁等 2022）计算  $P_{\text{whole}}$ 、 $P_{\text{positive}}$ 、 $P_{\text{negative}}$ ：

$$\begin{cases} P_{\text{whole}} = \frac{\sum_{w \in W} F_w \cdot S_w}{\sum_{w \in W} F_w} \\ P_{\text{positive}} = \frac{\sum_{a \in A} F_a \cdot S_a}{\sum_{a \in A} F_a} \\ P_{\text{negative}} = \frac{\sum_{b \in B} F_b \cdot S_b}{\sum_{b \in B} F_b} \end{cases}$$

(5) 最后基于扩展意义单位框架下 *China* 一词的总体语义韵强度，不同类别的语义倾向和类联接的语义韵强度分析归纳外媒眼中的中国大国形象建构，具体包括他国对中国疫情防控政策巨大调整的整体看法，了解新时期背景下他国眼中的中国“抗疫大国形象”如何，疫情背景下中国在政治、军事、外交、经济、疫情防控等折射综合国力的领域表现如何等。

4 研究结果

4.1 *China* 一词总体语义韵强度、语义倾向、类联接下语义韵强度

通过计算得到的 *China* 一词总体语义韵强度、语义倾向、类联接下语义韵强度如表 1 所示。

表 1 *China* 一词的语义韵分布数据

节点词及其 总体语义韵	语义倾向及其 语义韵	类联接（搭配型式）及其语义韵
<i>China</i> ( $f = 542$ ) $P_{\text{whole}} = +0.1568$ ( 100% ) $P_{\text{positive}} = +0.5482$ ( 65.29% ) $P_{\text{negative}} = -0.5796$ ( 34.71% )	放开政策 ( $f = 147$ ) $P_{\text{whole}} = +0.1417$ ( 100% ) $P_{\text{positive}} = +0.5883$ ( 63.13% ) $P_{\text{negative}} = -0.6229$ ( 36.87% )	$N/NP + P/PP + \text{China} + (P/PP)$ ( $f = 50$ ) $P_{\text{whole}} = +0.1793$ ( 100% ); $P_{\text{positive}} = +0.6868$ ( 65.52% ); $P_{\text{negative}} = -0.7850$ ( 34.48% )

（待续）



(续表)

节点词及其 总体语义韵	语义倾向及其 语义韵	类联接 (搭配型式) 及其语义韵
		China + V/VP + O ( $f = 48$ ) $P_{\text{whole}} = +0.1735$ ( 100% ); $P_{\text{positive}} = +0.6417$ ( 63.64% ); $P_{\text{negative}} = -0.6458$ ( 36.36% )
	放开政策 ( $f = 147$ ) $P_{\text{whole}} = +0.1417$ ( 100% ) $P_{\text{positive}} = +0.5883$ ( 63.13% ) $P_{\text{negative}} = -0.6229$ ( 36.87% )	(P/PP)/NP + China's + N/NP + (P/PP) ( $f = 28$ ) $P_{\text{whole}} = +0.1379$ ( 100% ); $P_{\text{positive}} = +0.6575$ ( 60.61% ); $P_{\text{negative}} = -0.6615$ ( 39.39% ) V/VP + (P/PP) + China/China's + (P/PP)/NP ( $f = 19$ ) $P_{\text{whole}} = +0.0571$ ( 100% ); $P_{\text{positive}} = +0.5250$ ( 57.14% ); $P_{\text{negative}} = -0.5667$ ( 42.86% )
China ( $f = 542$ ) $P_{\text{whole}} = +0.1568$ ( 100% ) $P_{\text{positive}} = +0.5482$ ( 65.29% ) $P_{\text{negative}} = -0.5796$ ( 34.71% )		V/VP + China's + N/NP + (complement) ( $f = 2$ ) $P_{\text{whole}} = -0.4167$ ( 100% ); $P_{\text{positive}} = +0.2500$ ( 33.33% ); $P_{\text{negative}} = -0.7500$ ( 66.67% )
		(P/PP)/NP + China's + N/NP + (P/PP) ( $f = 47$ ) $P_{\text{whole}} = +0.0881$ ( 100% ); $P_{\text{positive}} = +0.6240$ ( 59.52% ); $P_{\text{negative}} = -0.7000$ ( 40.48% )
	动态清零政策 ( $f = 144$ ) $P_{\text{whole}} = +0.1236$ ( 100% ) $P_{\text{positive}} = +0.5547$ ( 61.81% ) $P_{\text{negative}} = -0.5311$ ( 38.19% )	China + V/VP + O ( $f = 39$ ) $P_{\text{whole}} = +0.1202$ ( 100% ); $P_{\text{positive}} = +0.5814$ ( 61.40% ); $P_{\text{negative}} = -0.6136$ ( 38.60% ) N/NP + P/PP + China + (P/PP) ( $f = 34$ ) $P_{\text{whole}} = +0.1412$ ( 100% ); $P_{\text{positive}} = +0.5738$ ( 61.76% ); $P_{\text{negative}} = -0.5577$ ( 38.24% )
		V/VP + (P/PP) + China/China's + (P/PP)/NP ( $f = 17$ ) $P_{\text{whole}} = +0.1152$ ( 100% ); $P_{\text{positive}} = +0.5000$ ( 65.22% ); $P_{\text{negative}} = -0.6063$ ( 34.78% )

(待续)

(续表)

节点词及其 总体语义韵	语义倾向及其 语义韵	类联接 (搭配型式) 及其语义韵
	动态清零政策 ( $f = 144$ ) $P_{\text{whole}} = +0.1236$ (100%) $P_{\text{positive}} = +0.5547$ (61.81%) $P_{\text{negative}} = -0.5311$ (38.19%)	<p>predictive + P/PP + China (<math>f = 4</math>)  <math>P_{\text{whole}} = -0.4000</math> (100%);  <math>P_{\text{positive}} = 0</math> (0%);  <math>P_{\text{negative}} = -0.4000</math> (100%)</p> <p>V/VP + China's + N/NP + (complement) (<math>f = 3</math>)  <math>P_{\text{whole}} = -0.2500</math> (100%);  <math>P_{\text{positive}} = 0</math> (0%);  <math>P_{\text{negative}} = -0.2500</math> (100%)</p>
		<p>(P/PP)/NP + China's + N/NP + (P/PP) (<math>f = 43</math>)  <math>P_{\text{whole}} = +0.2012</math> (100%);  <math>P_{\text{positive}} = +0.4468</math> (73.81%);  <math>P_{\text{negative}} = -0.4909</math> (26.19%)</p>
China ( $f = 542$ ) $P_{\text{whole}} = +0.1568$ (100%) $P_{\text{positive}} = +0.5482$ (65.29%) $P_{\text{negative}} = -0.5796$ (34.71%)	对自身经济影响、 国内经济建设 ( $f = 94$ ) $P_{\text{whole}} = +0.2637$ (100%) $P_{\text{positive}} = +0.5419$ (69.05%) $P_{\text{negative}} = -0.5850$ (30.95%)	<p>N/NP + P/PP + China + (P/PP) (<math>f = 24</math>)  <math>P_{\text{whole}} = +0.1500</math> (100%);  <math>P_{\text{positive}} = +0.5850</math> (62.50%);  <math>P_{\text{negative}} = -0.5750</math> (37.50%)</p> <p>China + V/VP + O (<math>f = 18</math>)  <math>P_{\text{whole}} = +0.2355</math> (100%);  <math>P_{\text{positive}} = +0.4906</math> (80.00%);  <math>P_{\text{negative}} = -0.6250</math> (20.00%)</p> <p>V/VP + (P/PP) + China/China's + (P/PP)/NP  (<math>f = 7</math>)  <math>P_{\text{whole}} = -0.0667</math> (100%);  <math>P_{\text{positive}} = +0.6000</math> (44.44%);  <math>P_{\text{negative}} = -0.6000</math> (55.56%)</p>
		<p>predictive + P/PP + China (<math>f = 1</math>);  <math>P_{\text{whole}} = -0.7500</math> (100%);  <math>P_{\text{positive}} = 0</math> (0%);  <math>P_{\text{negative}} = -0.7500</math> (100%)</p>
		<p>V/VP + China's + N/NP + (complement) (<math>f = 1</math>)  <math>P_{\text{whole}} = 0</math> (0%);  <math>P_{\text{positive}} = 0</math> (0%);  <math>P_{\text{negative}} = 0</math> (0%)</p>

(待续)

(续表)

节点词及其 总体语义韵	语义倾向及其 语义韵	类联接 (搭配型式) 及其语义韵
		N/NP + P/PP + China + (P/PP) ( $f = 32$ ) $P_{\text{whole}} = +0.1370$ ( 100% ); $P_{\text{positive}} = +0.6393$ ( 60.87% ); $P_{\text{negative}} = -0.6444$ ( 39.13% ) China + V/VP + O ( $f = 20$ ) $P_{\text{whole}} = +0.1577$ ( 100% ); $P_{\text{positive}} = +0.3778$ ( 69.23% ); $P_{\text{negative}} = -0.3375$ ( 30.77% )
	与他国经济关系 ( $f = 90$ ) $P_{\text{whole}} = +0.1959$ ( 100% ) $P_{\text{positive}} = +0.4794$ ( 70.57% ) $P_{\text{negative}} = -0.4837$ ( 29.43% )	V/VP + (P/PP) + China/China's + (P/PP)/NP ( $f = 19$ ) $P_{\text{whole}} = +0.2200$ ( 100% ); $P_{\text{positive}} = +0.3500$ ( 80.00% ); $P_{\text{negative}} = -0.3000$ ( 20.00% ) (P/PP)/NP + China's + N/NP + (P/PP) ( $f = 18$ ) $P_{\text{whole}} = +0.2400$ ( 100% ); $P_{\text{positive}} = +0.5286$ ( 70.00% ); $P_{\text{negative}} = -0.4333$ ( 30.00% )
China ( $f = 542$ ) $P_{\text{whole}} = +0.1568$ ( 100% ) $P_{\text{positive}} = +0.5482$ ( 65.29% ) $P_{\text{negative}} = -0.5796$ ( 34.71% )		V/VP + China's + N/NP + (complement) ( $f = 1$ ) $P_{\text{whole}} = +0.2500$ ( 100% ); $P_{\text{positive}} = +0.7500$ ( 50.00% ); $P_{\text{negative}} = -0.2500$ ( 50.00% )
		N/NP + P/PP + China + (P/PP) ( $f = 9$ ) $P_{\text{whole}} = +0.0833$ ( 100% ); $P_{\text{positive}} = +0.2500$ ( 66.67% ); $P_{\text{negative}} = -0.2500$ ( 33.33% )
	文化交流、 媒体宣传 ( $f = 27$ ) $P_{\text{whole}} = +0.1939$ ( 100% ) $P_{\text{positive}} = +0.5515$ ( 68.37% ) $P_{\text{negative}} = -0.5790$ ( 31.63% )	V/VP + (P/PP) + China/China's + (P/PP)/NP ( $f = 9$ ) $P_{\text{whole}} = -0.0938$ ( 100% ); $P_{\text{positive}} = +0.5125$ ( 50.00% ); $P_{\text{negative}} = -0.7000$ ( 50.00% ) (P/PP)/NP + China's + N/NP + (P/PP) ( $f = 7$ ) $P_{\text{whole}} = +0.2000$ ( 100% ); $P_{\text{positive}} = +0.7000$ ( 66.67% ); $P_{\text{negative}} = -0.8000$ ( 33.33% )
		China + V/VP + O ( $f = 2$ ) $P_{\text{whole}} = +0.3250$ ( 100% ); $P_{\text{positive}} = +0.3250$ ( 100% ); $P_{\text{negative}} = 0$ ( 0% )

(待续)

(续表)

节点词及其 总体语义韵	语义倾向及其 语义韵	类联接 (搭配型式) 及其语义韵
		N/NP + P/PP + China + (P/PP) ( $f = 12$ ) $P_{\text{whole}} = -0.3833$ (100%); $P_{\text{positive}} = +0.8000$ (22.22%); $P_{\text{negative}} = -0.7214$ (77.78%)
		China + V/VP + O ( $f = 6$ ) $P_{\text{whole}} = -0.0500$ (100%); $P_{\text{positive}} = +0.6500$ (50.00%); $P_{\text{negative}} = -0.7500$ (50.00%)
	内政事务、 外交关系 ( $f = 24$ )	V/VP + (P/PP) + China/China's + (P/PP)/NP ( $f = 4$ ) $P_{\text{whole}} = +0.1518$ (100%) $P_{\text{positive}} = +0.5315$ (65.18%) $P_{\text{negative}} = -0.5590$ (34.82%)
		$P_{\text{whole}} = -0.5000$ (100%); $P_{\text{positive}} = 0$ (0%); $P_{\text{negative}} = -0.5000$ (100%)
		(P/PP)/NP + China's + N/NP + (P/PP) ( $f = 1$ ) $P_{\text{whole}} = -0.5000$ (100%); $P_{\text{positive}} = 0$ (0%); $P_{\text{negative}} = -0.5000$ (100%)
China ( $f = 542$ ) $P_{\text{whole}} = +0.1568$ (100%) $P_{\text{positive}} = +0.5482$ (65.29%) $P_{\text{negative}} = -0.5796$ (34.71%)		predictive + P/PP + China ( $f = 1$ ) $P_{\text{whole}} = +0.6000$ (100%); $P_{\text{positive}} = +0.6000$ (100%); $P_{\text{negative}} = 0$ (0%)
		N/NP + P/PP + China + (P/PP) ( $f = 5$ ) $P_{\text{whole}} = +0.6000$ (100%); $P_{\text{positive}} = +0.6000$ (100%); $P_{\text{negative}} = 0$ (0%)
		V/VP + (P/PP) + China/China's + (P/PP)/NP ( $f = 4$ ) $P_{\text{whole}} = 0$ (0%); $P_{\text{positive}} = 0$ (0%); $P_{\text{negative}} = 0$ (0%)
	国际援助 ( $f = 16$ ) $P_{\text{whole}} = +0.2708$ (100%) $P_{\text{positive}} = +0.5480$ (75.38%) $P_{\text{negative}} = -0.5781$ (24.62%)	China + V/VP + O ( $f = 3$ ) $P_{\text{whole}} = +0.5167$ (100%); $P_{\text{positive}} = +0.9000$ (66.67%); $P_{\text{negative}} = -0.2500$ (33.33%)
		(P/PP)/NP + China's + N/NP + (P/PP) ( $f = 2$ ) $P_{\text{whole}} = +0.3750$ (100%); $P_{\text{positive}} = +0.3750$ (100%); $P_{\text{negative}} = 0$ (0%)

(待续)

(续表)

节点词及其 总体语义韵	语义倾向及其 语义韵	类联接（搭配型式）及其语义韵
China ( $f = 542$ )	国际援助 ( $f = 16$ )	
$P_{\text{whole}} = +0.1568$ (100%)	$P_{\text{whole}} = +0.2708$ (100%)	V/VP + China's + N/NP + (complement) ( $f = 2$ ) $P_{\text{whole}} = -0.8000$ (100%);
$P_{\text{positive}} = +0.5482$ (65.29%)	$P_{\text{positive}} = +0.5480$ (75.38%)	$P_{\text{positive}} = 0$ (0%); $P_{\text{negative}} = -0.8000$ (100%)
$P_{\text{negative}} = -0.5796$ (34.71%)	$P_{\text{negative}} = -0.5781$ (24.62%)	

如表 1 所示，表中  $f$  表示节点词 China 的出现频数，也是类联接个数， $P_{\text{whole}}$  表示总体语义韵强度， $P_{\text{positive}}$  表示积极语义韵强度， $P_{\text{negative}}$  表示消极语义韵强度。 $P_{\text{whole}}/P_{\text{positive}}/P_{\text{negative}}$  后的括号百分比数值表示总体/积极/消极语义词占比。类联接中“V/VP”代表动词或动词短语，“N/NP”代表名词或名词短语，“P/PP”代表介词或介词短语。在 China 一词出现的共现语境中，共识别出 2720 个不同的具有态度意义的搭配词。其中积极搭配词 1776 个，占比 65.29%，消极搭配词 944 个，占比 34.71%。通过统计得出 China 一词共出现 542 次，即共有 542 个包含 China 的类联接出现。同时，根据 500 个索引行的语义特征，研究者将其分为七大类别，即（1）“放开政策”；（2）“动态清零政策”；（3）“对自身经济影响、国内经济建设”；（4）“与其他国经济关系”；（5）“文化交流、媒体宣传”；（6）“内政事务、外交关系”；（7）“国际援助”。另外，研究者还概括了 China 一词在每类语义倾向下的典型搭配型式，即类联接特征。

4.1.1.1 总体语义韵强度

如表 1 所示，China 一词的总体语义韵强度为+0.1568，呈现微弱强度的积极语义韵。需要说明的是，语义韵强度值包括符号和数值两部分。符号代表语义韵的极性类别，正号（+）为积极语义韵，负号（-）为消极语义韵；数值代表语义韵强度等级，数值越大，表示语义韵强度越高。在本研究中，经情感分析技术加成的语义韵强度主要分布区间为-1（消极）到+1（积极），因此，China 一词整体呈现微小强度的积极语义韵。其中，表达积极态度意义的搭配词占比 65.29%，正向情感均值为+0.5482；表达消极态度的搭配词占比 34.71%，负向情感均值为-0.5796。虽然与 China 共现的积极语义词频数远高于消极语义词，其所表达积极态度的情感强度却普遍地低于消极情感强度，说明与 China 一词搭配的积极语义词对应的情感值普遍偏低，而与 China 一词搭配每个消极语义词对应的情感值普遍偏高，表明国外媒体在塑造积极正面的中国形象时，态度极为谨慎拘谨；在塑造消极反面的“中国形象”时，则大放厥词，不知收敛。

综上所述, China 一词总体呈现微弱强度的积极语义韵, 与之共现的积极语义词频数远高于消极语义词, 其所表达积极态度的情感强度却普遍地低于消极情感强度, 说明从 2022 年 12 月 1 日至 2022 年 12 月 31 日, 有关疫情的国际新闻报道中, 中国总体呈现一个积极正面的国家形象, 但对中国的消极评价力度则远远高于积极评价力度。

#### 4.1.2 不同语义倾向中的语义韵强度

分别比较不同语义倾向的积极情感强度和消极情感强度, 可以发现, 无论是积极情感强度还是消极情感强度, “放开政策”的两极情感强度皆为最强, 分别为 +0.5883 和 -0.6229, “与他国经济关系”的两极情感强度皆为最弱, 分别为 +0.4794 和 -0.4837, 说明在 12 月有关疫情的国际新闻报道中, 对于中国形象的七大方面(对应 7 类语义倾向), 国外媒体对于“放开政策”的赞誉与批判最为强烈, 说明中国“放开政策”的实施在国际社会引起了不小轰动, 是一个极具争议性的国际热点话题。相较之下, “与他国经济关系”的国际关注度较小, 呈现“毁誉参半”的局面, 两极情感强度皆为最弱, 说明国外媒体对于“中国与他国经济关系”这一社会议题持相对审慎的态度。

将 7 类语义倾向的积极情感强度与消极情感强度分别进行对比, 可以看出, 除了“动态清零政策”以外, 其他 6 类语义倾向中, 表达积极态度的情感强度都普遍低于表达消极态度的情感强度。说明在 12 月有关疫情的国际新闻报道中, 对于中国形象的七大方面(对应 7 类语义倾向), 国外媒体的批判声整体大于赞誉声。

数据进一步显示, 不同语义倾向下, 虽然与 China 共现的积极语义词频数都远高于消极语义词, 其所表达积极态度的情感强度与负向态度的消极情感强度却“平分秋色”, 差异微小。具体而言, 以 7 类语义倾向中总体语义韵强度最高与最低的“国际援助”与“动态清零政策”为例, “国际援助”的总体语义韵强度最高, 达到 +0.2708, 其中与 China 共现的积极语义词占比 75.38%, 消极语义词占比 24.62%, 可见与 China 共现的积极语义词频数远高于消极语义词, 但“国际援助”所表达积极态度的情感强度却只是略低于表达消极态度的情感强度, 分别为 +0.5480 与 -0.5781。“动态清零政策”的总体语义韵强度最高, 为 +0.1236, 其中与 China 共现的积极语义词占比 61.81%, 消极语义词占比 38.19%, 可见与 China 共现的积极语义词频数远高于消极语义词, 但“动态清零政策”所表达积极态度的情感强度却只是略高于表达消极态度的情感强度, 分别为 +0.5547 与 -0.5311。以上分析说明 7 类语义倾向中, 与 China 一词搭配的积极语义词对应的情感值普遍偏低, 而与 China 一词搭配每个消极语义词对应的情感值普遍偏高, 表明国外媒体对塑造积极正面的中国形象意愿较低, 倾向于使用一些正向情感值较低的情感词, 持谨慎甚至是中立的态度。对塑造消极反面的中国形象的意愿更大, 倾向于使用一些负向情感值较高的情感词, 肆意抹黑中国。



### 4.1.3 不同类联接中的语义韵强度

在 7 类语义倾向中,研究者共发现了 6 种类联接,它们分别是:(1) N/NP + P/PP + China + (P/PP); (2) China + V/VP + O; (3) (P/PP)/NP + China's + N/NP + (P/PP); (4) V/VP + (P/PP) + China/China's + (P/PP)/NP; (5) V/VP + China's + N/NP + (complement); (6) predictive + P/PP + China。

研究数据表明,类联接(或搭配型式)亦会对语义韵分布产生影响,即使是一语义倾向下的不同搭配型式,其共现语境呈现的语义韵强度也会有所差别。以总体语义韵强度最低的“动态清零政策”为例,如表 1 所示,该语义倾向可关联 6 类搭配型式:(1) (P/PP)/NP + China's + N/NP + (P/PP); (2) China + V/VP + O; (3) N/NP + P/PP + China + (P/PP); (4) V/VP + (P/PP) + China/China's + (P/PP)/NP; (5) predictive + P/PP + China; (6) V/VP + China's + N/NP + (complement)。

在 6 种搭配型式中,型式 3 的总体语义韵强度值最高,达到+0.1412,其余 3 类搭配型式的语义韵强度虽然低于型式 3,却也属于微小强度的积极语义韵范畴,最后两类搭配型式则被消极情感主导,总体语义韵皆为接近中等强度的消极语义韵。总体说来,6 种搭配形式按总体语义韵强度值降序排列分别为+0.1412(型式 3)、+0.1202(型式 2)、+0.1152(型式 4)、+0.0881(型式 1)、-0.2500(型式 6)和-0.4000(型式 5)。可以进一步将这 6 种搭配型式分为两组,一组被总体积极语义韵主导(型式 1-4),一组则被总体消极语义韵主导(型式 5-6)。组一中 4 种搭配型式仍呈现较大的组间差异,虽然 4 种搭配型式中与 China 共现的积极语义词频数都远高于消极语义词,但型式 1 与型式 4 所表达积极态度的情感强度却低于表达消极态度情感强度,型式 2 与型式 3 所表达积极态度的情感强度则高于表达消极态度情感强度。组二中型式 5、型式 6 的共现语境则全部由消极语义词构成(即消极词占比为 100%)。综合组一与组二,可得出结论,国外媒体在对中国的“动态清零政策”做出积极评述时,更倾向于使用型式 1—4。这一结果亦与 Hunston(2007: 256)研究发现一致:“只将语义韵归于某个节点词的整体属性可能过于简单化,因为节点词的态度意义可能会受短语结构的影响。”

## 4.2 中国国家形象塑造

根据 China 一词的总体语义韵,可以看出总体呈现微弱强度的积极语义韵,即中国总体上呈现出了一个积极正面的国家形象,但对中国的消极评价力度则远远高于积极评价力度。

语料中含有“China”一词的类联接大多使用了“N/NP + P/PP + China + (P/PP)”型式,该类联接也是五种类联接中积极、消极语义韵强度最高的,反映了外媒在对“放开政策”进行评述时,无论是支持还是反对,都倾向于使用“N/NP + P/PP + China + (P/PP)”来表明自己的态度。

#### 4.2.1 从“创伤”走向“复苏”、与他国经济关系“挑战与机遇”并存的中国

根据语义倾向“对自身经济影响、国内经济建设”及“与他国经济关系”，实行“动态清零”政策后，经济生产受到影响；“放开政策”实施后，经济复苏的苗头初现。如下列代表性语料所示：

(1) But Ackman says there's a growing case for it to abandon that strategy, as **China grapples with a zero-COVID-fueled economic slowdown** and a debt crisis in its property sector. It suggests China's economy weakened more than thought this autumn as the global economy cooled and Covid-19 lockdowns were implemented at many cities.

(2) China stocks getting a short-term boost from the prospects of COVID-19 restrictions being relaxed./China's recovery from Covid-19 is a primary factor for oil demand growth (and the global economy...) going forward. /**As China backs away from its zero-COVID policy** and loosens restrictions business optimism and a slight return to normalcy are welcome changes for citizens and investors.

(3) Instead of kicking off the start of **a resurgence of economic activity in China**, the pandemic could continue to have a negative impact on the Chinese economy in the months ahead.

上述 3 条语料中，两条语料中含有 China 一词的类联接都是“China + V/VP + O”，该类联接也是 6 种类联接中总积极语义韵中强度最高的，反映了外媒在对疫情对中国经济影响、国内经济建设进行评述时，倾向于使用“China + V/VP + O”来表明整体乐观看好的情感态度。

另一方面，中国与他国的经济关系也呈现多元化，“挑战与机遇”并存，如下列代表性语料所示：

(4) There is less agreement when **it comes to China** where big investors seem to be in a wait-and-see mode.

(5) Europe is **becoming increasingly reliant on China** for trade and many of its top companies are eager to invest in the world's second biggest economy despite the disruption caused by Covid lockdowns.

(6) Apple Inc. is accelerating its plan to **move iPhone production out of China** in the wake of violent worker protests over COVID regulations at the world's largest iPhone factory, a report by the Wall Street Journal said Saturday.

( 7 ) China is building capabilities that put most U.S. space assets at risk and **China sees the domain as crucial to their military strategy**, the head of the U.S. Space Command said on Friday.

( 8 ) Avatar The Way of Water' could be a **'life saver' for China's cinemas**. /James Cameron's "Avatar The Way of Water" has premiered in China 13 years after the original film took the country by storm. It's expected to be a shot in the arm for the world's second-biggest film market which has been battered by the country's stringent zero-Covid policy.

( 9 ) Given that **China is our major trading partner** that is obviously problematic. We'll continue to engage diplomatically without a loudhailer including on issues where there are differences and we will continue to make representations on behalf of Australian citizens.

上述 6 条语料包含一种类联接“N/NP + P/PP + China + (P/PP)”，该类联接也是 5 种类联接中消极语义韵强度最高的，反映了外媒在对疫情背景下中国与外国经济关系进行评述时，倾向于使用“N/NP + P/PP + China + (P/PP)”来表明整体悲观失望的情感态度。

#### 4.2.2 拥有大国责任担当、坚持底线与原则、自强不息的中国

从语义倾向“国际援助”可以看出，中国拥有大国责任担当，能对疫情期间在国内滞留的德国人提供力所能及的人道主义帮助，面对他国的“援手”，中国能够做到自强不息、“非必要不求助”，从语义倾向“内政事务、外交关系”可以看出，在处理内政事务时，能够坚持底线与原则，坚决反对他国干涉中国内政，如下列代表性语料所示：

( 10 ) He said that the Chinese government formally informed Berlin in a diplomatic note that German citizens can be inoculated with the vaccine which otherwise isn't cleared for use in China. /**China would make BioNTech-Pfizer shots available for expatriates** and said that can “only be a first step” indicating that he hoped the vaccine's use could be expanded to the Chinese public. John Kirby, the White House national security spokesperson, told reporters that China has not requested help at this stage. /The US is prepared to help China with vaccines to help address its Covid-19 outbreak but the government in Beijing hasn't asked for assistance so far, Secretary of State Antony Blinken said.

含有“国际援助”语义倾向的语料中的类联接“China + V/VP + O”在 5 种类

联接中积极语义韵强度最高,表明外媒对于中国在疫情背景下提供的人道主义帮助,倾向于使用“China + V/VP + O”来表达赞同支持态度。含有“内政事务、外交关系”语义倾向的语料的类联接“China + V/VP + O”在 5 种类联接中消极语义韵强度最高,表明对于中国在新冠肺炎疫情背景下与他国的外交关系、处理内政事务的方式,倾向于使用“China + V/VP + O”来表达否定反对态度。

综合上述分析,研究者根据 China 一词的语义韵强度、不同语义倾向与类联接下语义韵强度,总结出了疫情期间中国形象的三个方面:(1)经济层面,中国从“创伤”走向“复苏”,与他国经济关系“挑战与机遇”并存;(2)外交层面,中国拥有大国责任担当、自强自立、坚持底线与原则;(3)文化层面,中国提倡多项文化交流。由“China”一词的总体语义韵可以得出,在有关疫情的国际新闻报道中,中国总体呈现一个积极正面的国家形象,与研究问题(2)总结的中国形象的三个方面相呼应。

## 5 结语

本文基于 Coronavirus 语料库与扩展意义单位模型,尝试运用情感分析技术优化和精化语义韵强度测量,并使用新算法计算以 China 为节点词的 3 类语义韵强度值,以及对比分析 China 在不同语义倾向和类联接中的语义韵特征异同,由此归纳出外媒眼中的中国形象。研究发现,China 一词总体呈现微弱强度的积极语义韵,与之共现的积极语义词频数远高于消极语义词,其所表达积极态度的情感强度却普遍低于消极情感强度,说明 2022 年 12 月 1 日至 2022 年 12 月 31 日,在有关疫情的国际新闻报道中,中国总体呈现一个积极正面的国家形象,但对中国的消极评价力度则远远高于积极评价力度。上述“他者之语”也对中国形象的对外传播提供了相关启示,中国应重构大国形象宣传逻辑,实现话语体系创新调整,坚定人类命运共同体理念,讲好负责任的中国故事,澄清、抵御污名形象,坚定、及时表明大国立场。

本研究旨在提供一种新的交叉研究视角,探讨语义韵强度更精细测量的可行性和操作性。毋庸讳言,本文只是对基于情感分析的语义韵强度计算的初始研究,尚存许多不足之处,比如情感词典收录的情感词是否全面、情感倾向赋值是否合理、如何选择或建构合适的情感词典等亦会影响本文语义韵强度计算的准确性。在后续研究中,可以尝试把语言情感分析技术应用于不同领域、不同文本类型(尤其是弱态度文本)的语义韵强度计算,进一步检验辅以情感分析技术的语义韵强度测量的适用性和普及性。同时,未来研究还可以从扩展意义单位视角出发,扩充现有情感词典的态度/情感词数据,使语言情感分析技术能更好适用于扩展意义单位分析。

## 参考文献

- ELLIS C, FREY E, JALKANEN I. The psycholinguistic reality of collocation and semantic prosody[C]//RÖMER-BARRON U, SCHULZE R. Exploring the lexis-grammar interface. Amsterdam: John Benjamins, 2009: 89-114.

- FIRTH R. Selected papers of J. R. Firth, 1952-59[M]. Bloomington: Indiana University Press, 1968.
- HUNSTON S. Semantic prosody revisited[J]. International Journal of Corpus Linguistics, 2007, 12: 249-268.
- LOUW B. Contextual prosodic theory: bringing semantic prosodies to life[C]//HEFFER C, SAUNTON H, FOX G. Words in context: a tribute to John Sinclair on his retirement. Birmingham: University of Birmingham, 2000: 48-94.
- SINCLAIR J. The search for units of meaning[J]. Textus, 1996, 9(1): 75-106.
- SINCLAIR J. Trust the text: language, corpus and discourse[M]. London & New York: Routledge, 2004.
- WEI N, LI X. Exploring semantic preference and semantic prosody across English and Chinese: their roles for cross-linguistic equivalence[J]. Corpus Linguistics and Linguistic Theory, 2014, 10(1): 103-138.
- YANG L, LI J, CUNNINGHAM P, et al. Exploring the efficacy of automatically generated counterfactuals for sentiment analysis[C]//Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), online: ACL-IJCNLP. 2021: 306-316.
- 高歌, 卫乃兴. 汉英翻译界面下的语义韵探究: 来自《红楼梦》英译本的证据[J]. 解放军外国语学院学报, 2019 ( 1 ): 48-56.
- 胡雪婵, 吴长安. 汉语成语语义韵的冲突及其消解[J]. 古汉语研究, 2018 ( 1 ): 59-69.
- 李华勇. 汉译口语中“被”字句的语义韵: 基于复合语料库的分析[J]. 浙江外国语学院学报, 2020 ( 5 ): 82-90.
- 李晶洁, 胡奕阳, 陶然. 基于情感倾向分析的语义韵强度算法探析[J]. 外国语, 2022 ( 5 ): 65-74.
- 李晓红, 卫乃兴. 双语视角下词语内涵义与语义韵探究[J]. 现代外语, 2012 ( 1 ): 30-38.
- 刘鼎甲. 新冠肺炎疫情中美国媒体涉华报道的语料库历时分析[J]. 外国语, 2021 ( 6 ): 52-64.
- 孙莉. 中国英语学习者近义词使用情况的语义韵新解——对比 affect 和 influence[J]. 当代外语研究, 2017 ( 1 ): 57-60.
- 陶源, 胡谷明. 致使动词语义韵翻译研究——基于俄汉平行语料库的专业文本[J]. 武汉大学学报 ( 人文科学版 ), 2015 ( 1 ): 119-124.
- 王婷, 杨文忠. 文本情感分析方法研究综述[J]. 计算机工程与应用, 2021 ( 12 ): 11-24.

通信地址: 215026 江苏省苏州市 苏州大学外国语学院

# 基于语料库印度媒体“人类命运共同体”新闻话语命名策略研究<sup>\*</sup>

西安外国语大学 李梦翔

**提要：**本文选取话语—历史分析法中的命名策略视角，基于自建印度媒体“人类命运共同体”新闻报道语料库（2013—2022年），对检索词的索引行、高频共现名词以及语义韵进行分析，结果表明印度媒体在不同时期对“人类命运共同体”的态度经历了既不关注也不认可，开始关注、客观报道，高度关注、有褒有贬，减少关注、负面偏见的历时变化，这种变化主要受到国际关系、意识形态和国际环境等因素的影响。该研究旨在了解印度媒体对人类命运共同体的报道态度和传播情况，为制定人类命运共同体的国际对外传播策略提供参考，发挥语言研究为国家战略服务的作用。

**关键词：**人类命运共同体、命名策略、印度媒体、名词

## 1 引言

“人类命运共同体”旨在倡导追求本国利益时要兼顾他国关切。人类只有一个地球，共处共享同一个地球，如今国际社会面临百年未有之大变局，各国相互联系相互依存的局面日益加深，该理念契合人类可持续发展的需要，对推动全球治理体系改革、构建新型国际关系和国际秩序的共同价值规范发挥重要作用（高金萍 2023）。自该理念提出后，习近平主席在多个重要会议及国际场合阐释和倡导推动构建人类命运共同体，中国通过践行新发展理念、共商共建共享“一带一路”、全球发展命运共同体等实际行动不断践行和丰富发展人类命运共同体。十年来，越来越多的国家了解、支持并加入到人类命运共同体的实践中，这一理念逐渐得到国际社会的广泛关注。中印两国均为世界文明古国，两国人民在政治、贸易、文化、农业、国防、科技等领域的交往与合作密切，同时两国互为邻国，均是人口大国和发展中国家。由于中印两国这种复杂多变的关系，印度媒体密切关注中国的发展动向，随着“人类命运共同体”理念日益成为中国开展双边和多边外交的主线，印度媒体对这一理念进行了相关报道（张忞煜 2021）。在中国坚持推动构建“人类命运共同体”理念

<sup>\*</sup> 本文系西安外国语大学研究生科研基金项目“基于语料库‘人类命运共同体’新闻报道话语策略分析——以 G20 国家为例”（2023SS008）阶段性研究成果。



的今天,考虑到中印两国的关系发展局势,有必要对印度媒体关于人类命运共同体的新闻报道进行研究。本研究能够了解该理念在印度的传播、接受及构建情况,进而完善该理念的对外传播策略。

国内外学者对“人类命运共同体”进行了相关研究,但主要集中在中国政治与国际政治、经济学、新闻与传播学、教育学等学科(Smith 2018; Nathan & Zhang 2022; Eisenman 2023; 王公龙 2021; 周玲、李佳欣 2021; 刘建萍、周金鑫 2022; 吴志成 2023)。由于国际关系、国家利益和意识形态的差异,不同国家基于本国利益对“人类命运共同体”的新闻报道差异明显。针对“人类命运共同体”新闻报道,国外鲜有学者对其进行研究,而国内学者研究了国内主流媒体如何呈现和传达“人类命运共同体”理念(朱鸿军、李喆 2023),以及美国、南非、德国、印度媒体对“人类命运共同体”理念的报道情况、传播特点等(高金萍、余悦 2020; 郭石磊、孙有中 2020; 许涌斌、高金萍 2020; 朱莉 2021),但此类研究主要基于政治学和新闻传播学的角度,从语言视角对该话题开展的研究并不多见,且主要局限于对“人类命运共同体”的英译研究,例如邓海丽(2020)研究了针对国外媒体“人类命运共同体”的英译策略和传播情况;自正权、吴洁(2022)研究了“人类命运共同体”英译在东盟国家媒体的传播特点和接受情况。综观国内外学者对“人类命运共同体”的研究发现,语言学领域的研究较少,研究的话题和学科领域有待扩展。

基于以上综述结果,本文拟从语言学领域对人类命运共同体开展研究。本文依据 Wodak 话语—历史分析法中的命名策略,基于自建的印度媒体“人类命运共同体”新闻报道(2013—2022年)语料库,运用定量定性结合的方法,对印度媒体关于“人类命运共同体”新闻报道中的命名策略进行分析,具体分析以下两个问题:(1)印度媒体中关于“人类命运共同体”新闻报道中运用的命名策略及其语言实现形式是什么?(2)印度媒体对“人类命运共同体”新闻报道的态度及其历时变化如何?

## 2 理论框架

话语—历史分析法(Discourse-Historical Approach,简称DHA)是批评话语分析的研究范式之一,该范式是Wodak及其团队在分析语篇的过程中逐步形成的,为批评话语分析提供了新的研究视角。话语—历史分析法注重从历时的角度考察社会问题和分析话语,要求尽可能多地利用语料及有关历史背景,多角度地考察某一类话语在特定时期的历时变化(刘曼 2020),其目的是通过分析话语表征以及内群体和外群体的话语建构情况,进而阐明话语中蕴含的意识形态和态度倾向。该研究范式的分析框架包括3个研究步骤:一是确定语篇的具体内容或主题;二是研究语篇与语篇之间和文本与文本之间的关系;三是分析语体和文本(田海龙 2009),此步骤的重点是对话语策略进行分析。在话语—历史分析法的分析框架中,第三步是关键环节,即运用话语—历史分析法研究文本时,重点对其进行话语策略分析。话语

一历史分析法包括命名策略、述谓策略、辩论策略、视角化策略和强化弱化策略 5 种话语策略。命名策略可以用于构建和再现社会活动者,如可以通过成员分类的方法,将一个人划归为属于自己一方的“组内成员”或属于对方的“组外成员”(田海龙 2009)。该策略旨在对社会活动者、物体、现象、事件、过程和行为进行话语构建,其运用的手段包括进行成员归类、使用暗喻、转喻、提喻等修辞手法、用特定词语表征过程和动作(Wodak & Meyer 2015: 33),语言实现形式主要为指示词和名词。本研究基于命名策略,对印度媒体新闻报道中用于建构“人类命运共同体”的高频共现名词进行分析,解析印度媒体所构建的“人类命运共同体理念”的特点。

### 3 研究设计

#### 3.1 研究方法

本研究基于语料库开展定量定性相结合的研究,通过更大的数据集呈现的结果更可靠有效(Baker & Levon 2015),在一定程度上提高了研究的可信度和客观性。本研究涉及的语料库方法有词频统计、索引行分析和语义韵分析。通过词频统计能够得出高频出现的词语,词语的使用能直观地反映语言使用的显著特征,在一定程度上还可以反映作者的态度倾向和意识形态。对特定高频词索引行的分析便于批量观察语言现象,找出其中的规律,帮助研究者更好地分析语言的使用特点及背后动因(孙秀丽 2021)。通过观察索引行周围的共现词,还可扩展至语义韵的研究,语义韵在一定程度上和作者的态度倾向有关(Baker 2006: 87)。研究者结合语境对特定词语索引行进行定性分析,由此得出语言的使用特点及暗含的态度倾向和意识形态。

#### 3.2 研究语料

本研究选取 LexisNexis 新闻数据库,检索词选择“人类命运共同体”的常用英文形式,主要有以下 4 种: a community with a shared future for mankind(十九大报告译法), a community of common destiny(十八大报告译法), a community of shared future for mankind/humanity(部分媒体译法)和 a community of human destiny(部分媒体译法),考虑到不同媒体对“人类”一词翻译方式的不同,本文在实际检索中选取 a community with a shared future、a community of common destiny、a community of shared future 和 a community of human destiny 为该研究检索词。国家选择印度,语言选择英语,时间跨度选择 2013—2022 年,新闻来源需要除去 TendersInfo-News、China Government News 等非印度专属媒体的来源。以上述 4 种表达为检索词搜索相关新闻报道,并人工逐一剔除重复及无关新闻,将待研究的新闻仅保留标题和正文内容,保存为文本文档,得到该研究的专用语料库——印度媒体“人类命运共同体”新闻报道语料库(Indian News Corpus of a Community with a Shared Future for

Mankind, 简称 INC), 该语料库共 142 篇新闻, 合计 92,209 词, 语料库中年度报道量情况如图 1 所示。本研究使用的语料库分析软件是 AntConc 4.1.4。

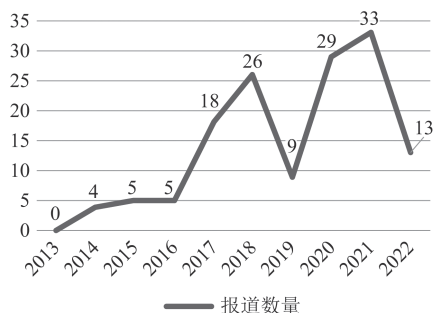


图 1 INC 语料库 2013—2022 年年度报道量分布情况

图 1 的升降变化能够表明印度媒体对“人类命运共同体”的新闻报道和关注情况。2017 年, 中共十九大召开, 十九大报告中提出的“十四个坚持”中包括“坚持推动构建人类命运共同体”, 习近平新时代中国特色社会主义思想被写入党章, 吸引印度媒体关注并报道, 故该年报道量显著增加。2020 年, 新冠肺炎疫情的爆发以及中印边境冲突事件促使印度媒体对中国提出的“人类命运共同体”倡议频繁报道, 故报道量经过 2019 年的回落之后再次显著增加。2021 年为近十年印度媒体对“人类命运共同体”报道量最多的一年, 统计发现, 仅 2021 年 4 月相关报道数量便达到 10 篇, 占比将近三分之一, 主要关注该月在中国海南召开的博鳌亚洲论坛。博鳌亚洲论坛年会每年举行一届, 但 2020 年的年会因为新冠肺炎疫情被取消, 2021 年的年会是新冠肺炎疫情全球爆发以来的首届年会, 在此大环境下构建“人类命运共同体”的必要性日益显著, 故印度媒体对此进行了大量报道, 随后报道量减少。2022 年相比 2021 年的报道量减少将近三分之二, 说明印度媒体对该倡议关注减少。根据新闻报道量与社会历史语境, 笔者将印度媒体 2013—2022 年有关“人类命运共同体”的新闻报道分为 4 个时期。第一时期为 2013—2016 年, 第二时期为 2017—2019 年, 第三时期为 2020—2021 年, 第四时期为 2022 年。本文构建了对应的四个子语料库。INC1 (2013—2016 年) 共 11,926 词, INC2 (2017—2019 年) 共 28,101 词, INC3 (2020—2021 年) 共 42,469 词, INC4 (2022 年) 共 9,713 词。

## 4 结果与讨论

命名策略探究如何从语言学角度命名或指称话语中的人、物、现象、事件、过程和行为, 本研究中其实现形式主要为名词。利用检索分析软件 AntConc 4.1.4, 使用上文提及的人类命运共同体实际检索词 a community with a shared future、a

community of common destiny、a community of shared future 和 a community of human destiny, 运用该软件的 KWIC 功能, Context Size 设定为 15 tokens, 分别检索 INC1、INC2、INC3、INC4 四个子语料库, 生成以“人类命运共同体”为中心词的索引行, 分别对每个子语料库检索行中的词语进行词频统计, 设置停用词, 提取前后语境中与“人类命运共同体”共现搭配的高频名词进行统计分析, 考虑到 INC1 和 INC4 词数远小于 INC2 和 INC3, 故选取高频名词的标准为  $\text{Freq} \geq 2$  (第一时期和第四时期)、 $\text{Freq} \geq 3$  (第二时期和第三时期)。结果发现, 与检索词高频共现的名词主要包括“人类命运共同体”的指称名词(如 idea、concept、initiative、vision、goal 和 objective)、“人类命运共同体”的理念内涵相关名词(如 prosperity、development、cooperation、diplomacy、fairness 和 power)、国家名称及地理范畴相关名词(如 China、Pakistan、India、Asia、neighbour 和 world)、人物范畴名词(如 Xi、Wang、leaders、president 和 ambassador)和其他特定类别名词(如 Belt、Road、covid 和 pandemic)。本文将按照共时历时相结合的分析方法, 结合中心词的索引行和语义韵对其高频共现名词进行定量定性分析, 进而探究印度媒体关于“人类命运共同体”的新闻报道在不同时期运用的命名策略和态度及其变化情况。

#### 4.1 第一时期(2013—2016 年)

运用统计分析软件 AntConc 4.1.4 对 INC1 检索词索引行的高频名词( $\text{Freq} \geq 2$ )进行统计分词, 研究发现本时期印度媒体对人类命运共同体的命名策略中无高频指称名词, 仅使用过 idea 一词一次, 用于指称该倡议, 无其他指称名词, 与该倡议理念内涵相关的高频共现名词有 prosperity、development、diplomacy、interests 和 power, 地理范畴名词有 China、neighbour(hood)、areas、Asia、Bandung 和 country(ies), 无人物范畴名词和其他特定类别名词。统计结果表明, “人类命运共同体”的指称名词和该倡议理念内涵相关名词较地理范畴名词数量少。深入考察索引行及语义韵, 发现印度媒体此时并不看好且并不过多关注该倡议。地理范畴名词, 尤其是 neighbour(hood) 的高频出现, 表明印度媒体较为关注在“人类命运共同体”框架下中国和印度周边国家的国际关系、合作交往。例如:

(1) **A community of common destiny** is a saleable **idea** in the neglected peripherals.

(2) Chinese Premier promised his Pakistani counter-part that, his country is all set to establish “**a community of common destiny**” with its **neighbour**.

例(1)中印度媒体认为, 人类命运共同体在被忽视的外围地区是一个“畅销”的想法, 用 idea 指称“人类命运共同体”, 认为该倡议只是一个抽象的、停留在思维

层面的想法,且该想法只在被忽视的外围地区才畅销,表明印度媒体此时并不认可人类命运共同体。例(2)间接引用中方总理对巴基斯坦总理的承诺,中国坚持同邻国构建人类命运共同体,本是中巴两国领导之间的对话,印度媒体却关注并报道,此类报道的多次出现表明其非常关注中国与印度周边国家在“人类命运共同体”框架下开展的交流与合作,而印方对此类合作交流持谨慎态度。通过考察这一时期命名策略索引行及其扩展语境,发现印度媒体对“人类命运共同体”的态度是不关注、不认可。

#### 4.2 第二时期(2017—2019年)

运用统计分析软件 AntConc 4.1.4 对 INC2 检索词索引行的高频名词(Freq ≥ 3)进行统计分词发现,本时期印度媒体对人类命运共同体的命名策略中,人类命运共同体高频指称名词有 concept、initiative、goal、idea、thought 和 vision。与该倡议理念内涵相关的高频共现名词有 cooperation、relations、development、governance、respect、diplomacy、fairness 和 justice 等。地理范畴名词有 China、countries、world、Pakistan、neighbourhood 和 Zhejiang。人物范畴名词有 Xi (Jinping)、people、leaders 和 Wang。其他特定类别名词有 Belt 和 Road。比较第二时期和第一时期的高频词发现,“人类命运共同体”的指称名词以及理念内涵相关名词数量和种类均显著增加,且出现了其他特定类名词 The Belt and Road/One Belt One Road (“一带一路”倡议),“一带一路”倡议是构建“人类命运共同体”的具体实践。以上名词的变化表明印度媒体在该时期开始大量关注“人类命运共同体”。地理范畴名词的频率和种类也有所增加。本时期印度媒体除了关注印度周边国家之外,也会在“人类命运共同体”框架下放眼世界(如 world)和聚焦中国部分省份(如 Zhejiang)开展的事务。人物范畴名词在本时期高频出现,通过观察索引行和高频动词 said/saying 发现,印度媒体较多引用权威人士的讲话描述人类命运共同体,权威人士话语的引用增强了新闻的客观性,表明本时期印度媒体对人类命运共同体多进行客观报道。例如:

(3) He also said that the initiative of building **a community with a shared future for mankind** is being transformed from a **concept** into **action**.

(4) Building **a community with a shared future for mankind** is to build a world of “lasting **peace**, universal **security**, common **prosperity**, **openness and inclusiveness**, **cleanliness and beauty**,” he remarked.

(5) In the face of such global unrest, China’s “**One Belt, One Road**” network under the realm of **a community of shared destiny** is destined to integrate people round the world without posing any threats to their security and existence.

例(3)中印度媒体间接引用习近平主席在中国共产党与世界政党高层对话会议



中的讲话,客观报道了构建“人类命运共同体”倡议不只是一个概念 (concept),而是实际的行动 (action)。例(4)引用一名外交使者的话语,说明构建“人类命运共同体”就是建设持久和平 (peace)、普遍安全 (security)、共同繁荣 (prosperity)、开放包容 (openness and inclusiveness)、清洁美丽 (cleanliness and beauty) 的世界,客观叙述了该倡议的含义。例(5)客观报道了面对全球动荡,人类命运共同体范畴下的“一带一路”倡议不会对世界人民的安全和生存构成威胁,本时期“One Belt, One Road”在报道中的出现表明印度媒体开始关注“人类命运共同体”理念下的具体实践,表明其对“人类命运共同体”有所关注。通过考察这一时期命名策略索引行及其扩展语境发现,印度媒体开始关注并客观报道“人类命运共同体”。

#### 4.3 第三时期 (2020—2021 年)

2020 年新冠肺炎疫情席卷全球,本时期印度媒体对“人类命运共同体”命名策略中与新冠肺炎疫情相关名词必然高频出现,运用统计分析软件 AntConc 4.1.4 对 INC3 检索词索引行的高频名词 (Freq  $\geq 3$ ) 进行统计发现,本时期特定专有名词中和新冠疫情相关高频名词有 pandemic、covid、epidemic,该时期其他类特定名词还有 Belt、Road;“人类命运共同体”的指称名词有 vision、concept、idea、initiative,其中 vision 出现频率最高,达 13 次,印度媒体认为“人类命运共同体”是一种想象 (vision),甚至是一种信仰 (belief, 出现 1 次),部分媒体还认为该倡议是一个教条 (doctrine, 出现 2 次),表明印度媒体对该倡议极度不认可,甚至带有批判的意味。和该倡议理念内涵相关的高频共现名词有 cooperation、development、solidarity、friendship、relations、economy 和 power。地理范畴名词有 China、world、countries、Afghanistan、Beijing、Nepal 和 Pakistan。人物范畴名词有 Xi (Jinping)、people、president、Wang 和 Guterres。“人类命运共同体”指称名词和理念内涵相关名词的高频出现表明印度媒体对该倡议高度关注,但对该倡议的态度由于社会环境因素的影响有褒有贬。新冠肺炎疫情期间印度失守时,中国支援印度对抗疫情,印度媒体也进行了正面报道。地理范畴名词中 Afghanistan、Nepal 和 Pakistan 高频出现,三国均是中国周边国家,而同为中国邻国的印度却鲜有出现,疫情期间中国对上述国家均提供了援助,再次说明印度媒体高度关注中国与周边国家的国际关系和合作来往,但印度对中国在“人类命运共同体”理念下积极推进的对外合作持有戒备心态 (张忞煜 2021)。例如:

(6) China will continue to uphold the vision of **a community with a shared future for mankind** and engage in more **cooperation** with the **international community**.

(7) In the spirit of building **a community with a shared future for**



**mankind**, China has also been working tirelessly to advance **international cooperation** against such major threats as climate change.

例(6)引用耿爽(2020年时担任外交部发言人)在中国同欧亚和南亚地区国家举行新冠肺炎疫情防控问题视频会议中的发言,表示中国会继续构建人类命运共同体,与国际社会(international community)开展更多合作(cooperation),在该篇新闻报道中,印度国家卫生部门官员对新冠肺炎疫情期间中国对印度的帮助表达了感谢,反映出此时印度媒体对“人类命运共同体”的态度具有正面倾向。例(7)报道了中国本着构建“人类命运共同体”的精神为应对气候变化等重大威胁的国际合作(international cooperation)做出巨大努力,表明印度媒体对中国构建“人类命运共同体”的高度赞扬。通过考察这一时期命名策略索引行及其语境发现,该时期印度媒体对“人类命运共同体”高度关注但其态度褒贬不一。

#### 4.4 第四时期(2022年)

运用统计分析软件 AntConc 4.1.4 对 INC4 检索词索引行的高频名词(Freq ≥ 2)进行统计发现,本时期印度媒体对“人类命运共同体”的命名策略中,“人类命运共同体”高频指称名词只有 initiative。和该倡议理念内涵相关的高频共现名词有 development、contribution、cooperation、economy 和 relations。地理范畴名词有 world、China 和 Asia。无高频人物范畴名词和其他特定类别名词。结合本时期新闻报道数量和社会历史语境分析发现,2021年印度媒体对“人类命运共同体”的关注度达到顶峰之后,随着新冠肺炎疫情的逐渐好转以及中印加勒万河谷边境冲突的缓和,印度媒体对“人类命运共同体”的关注有所降低。通过分析检索词的索引行及语义韵发现,由于国际环境、历史遗留问题等,印度媒体对该倡议的态度并未好转。通过考察这一时期命名策略索引行及其语境发现,印度媒体对人类命运共同体报道偏负面,对该倡议存在异议和偏见。

通过对检索词的索引行及高频共现名词的分析考察了四个时期命名策略的历时变化,结果发现印度媒体在四个时期对人类命运共同体的态度经历了既不关注也不认可,开始关注、客观报道,高度关注、有褒有贬,减少关注、负面偏见的历时变化。

## 5 结语

本文从 Wodak 话语—历史分析法话语策略中的命名策略视角,自建印度媒体人类命运共同体新闻报道语料库(2013—2022年),根据新闻报道量和社会历史语境将其分为4个时期,通过对检索词的索引行、高频共现词以及语义韵的分析,运用定量定性结合的方法,考察了不同时期印度媒体对人类命运共同体进行新闻报道运用的命名策略以及态度变化情况。结果发现,索引行中与人类命运共同体共现搭配

的高频名词主要包括“人类命运共同体”的指称名词、“人类命运共同体”的理念内涵相关名词、国家名称及地理范畴相关名词、人物范畴名词和其他特定类别名词。结合文本语境和社会历史语境,分别对4个时期的5类高频共现名词分析,结果表明印度媒体在4个不同时期对人类命运共同体的态度经历了既不关注也不认可,开始关注、客观报道,高度关注、有褒有贬,减少关注、负面偏见的历时变化,这种变化会受到国家关系、意识形态和国际环境等因素的影响。

通过对印度媒体关于人类命运共同体的新闻报道情况和态度进行分析,在国际上积极传播“人类命运共同体”理念任重而道远。要结合国际环境、不同国家的国情等制定合适的传播策略,积极传播该理念,必要时回应部分媒体对该理念的污名化和不实报道。作为语言学习者,要提高解读外媒报道的能力,发挥语言研究为国家和社会服务的作用。

### 参考文献

- BAKER P. Using corpora in discourse analysis[M]. New York: Continuum, 2006.
- BAKER P, LEVON E. Picking the right cherries? A comparison of corpus-based and qualitative analyses of news articles about masculinity[J]. Discourse & Communication, 2015, 9(2): 221-236.
- EISENMAN J. Locating Africa in China's community of shared future for mankind: a relational approach[J]. Journal of International Development, 2023, 35(1): 65-78.
- NATHAN A, ZHANG B. "A shared future for mankind": rhetoric and reality in Chinese foreign policy under Xi Jinping[J]. Journal of Contemporary China, 2022, 31(133): 57-71.
- SMITH S. Community of common destiny: China's "new assertiveness" and the changing Asian order[J]. International Journal, 2018, 73(3): 449-463.
- WODAK R, MEYER M. Methods of critical discourse studies[M]. London: Sage, 2015.
- 邓海丽. 中国时政话语的翻译策略——以“人类命运共同体”的英译及其传播为例[J]. 理论月刊, 2020(8): 21-30.
- 高金萍, 余悦. 美国媒体视域下“人类命运共同体”理念的呈现[J]. 新闻爱好者, 2020(3): 15-18.
- 高金萍. 人类命运共同体的全球传播图景——基于18国媒体相关报道分析[J]. 国际新闻界, 2023(3): 6-25.
- 郭石磊, 孙有中. 在地化呈现与跨文化阐释: 南非媒体涉“人类命运共同体”的报道研究[J]. 现代传播, 2020(12): 47-50.
- 刘建萍, 周金鑫. 人类命运共同体理念的国际传播研究综述[J]. 北京联合大学学报(人文社会科学版), 2022(4): 57-67.

- 刘曼. 日本主流报刊对“一带一路”的认知变化研究——基于语料库的批评话语分析[J]. 外语电化教学, 2020(5): 108-113.
- 孙秀丽. 基于语料库的能源企业身份建构的批评话语分析[J]. 山东外语教学, 2021(5): 21-30.
- 田海龙. 语篇研究: 范畴、视角、方法[M]. 上海: 上海外语教育出版社, 2009.
- 王公龙. 构建人类命运共同体: 引领新型经济全球化的中国方案[J]. 上海行政学院学报, 2021(5): 4-13.
- 吴志成. 人类命运共同体理念的中华文化基础[J]. 政治学研究, 2023(1): 24-33.
- 许涌斌, 高金萍. 德国媒体视域下的“人类命运共同体”理念研究——语料库辅助的批评话语分析[J]. 德国研究, 2020(4): 151-167.
- 张忞煜. 印度主流媒体“人类命运共同体”理念报道探析[J]. 国际传播, 2021(3): 59-66.
- 周玲, 李佳欣. 人工智能时代的高等教育与人类命运共同体[J]. 苏州大学学报(教育科学版), 2021(4): 8-15.
- 朱鸿军, 李喆. 主流媒体对人类命运共同体的政治传播分析[J]. 上海交通大学学报(哲学社会科学版), 2023(1): 67-86.
- 朱莉. “人类命运共同体”理念在印度的传播——基于2020年1-8月部分印度主流媒体的分析[J]. 青年记者, 2021(16): 120-121.
- 自正权, 吴洁. “人类命运共同体”英译在东盟国家的传播与接受研究[J]. 西安外国语大学学报, 2022(2): 96-101.

通信地址: 710128 陕西省西安市 西安外国语大学英文学院

# frGLOBE 当代法语书面语平衡语料库的创建\*

北京外国语大学 殷俐恺

**提要：**frGLOBE 当代法语书面语平衡语料库是“北外全球语料库集群”项目的子课题之一，旨在收集近十年的法语书面语文本，建成百万词级的平衡语料库。本文首先简述了法语语料库的建设历程，在此基础上讨论 frGLOBE 语料库的建库模式及建库过程，并展望了基于 frGLOBE 的相关研究及后续语料库建设。

**关键词：**frGLOBE、法语书面语平衡语料库、法语教学与研究

frGLOBE 当代法语书面语平衡语料库（简称 frGLOBE 语料库）是“北外全球语料库集群”项目下的百万词级法语子库。“北外全球语料库集群”项目旨在覆盖北外开设的 101 个语种，借鉴布朗语料库的采样方案，建成 GLOBE（Global Languages Out of BFSU Expertise）系列的当代书面语平衡语料库。GLOBE 系列语料库可与现有英汉布朗家族语料库进行对比，进而开展双语或多语对比研究。frGLOBE 语料库旨在收集 2013—2023 年首次出版或发表的原创法语文本，其设计规模为 100 万词。

## 1 法语语料库建设历程

法语语料库建设的早期尝试主要应用于词典编纂。20 世纪 60 年代，法国国家科学研究中心召集约 100 名语言学、计算机科学、图书情报学等领域的研究人员，编纂了《法语宝典》，旨在反映 16—20 世纪法国语言使用情况，成为第一部利用计算机技术编成的词典。这部大型词典共 16 卷，约 2.3 万页，由保罗·安布斯（Paul Imbs）和伯纳德·凯马达（Bernard Quemada）主持编纂。该词典收录 10 万词条，每个词条提供详尽的定义、用法、历史背景及词源信息，所有形态和语义分析均以引文考证为基础，深入探讨了词汇在不同文化、历史和社会背景下的含义变化。纸质版《法语宝典》于 1971—1974 年出版，数字化法语宝典于 1990—2001 年完成，涵盖 10 万单词及其词源历史、27 万定义、43 万示例，共计 3.5 亿字符，并进行了语

\* 本文系北京外国语大学 2022 年度“双一流”重大标志性项目“多语种词典编纂理论与实践研究”（2022SYLZD015）和北京外国语大学 2022 年度“双一流”重大标志性（培育）项目“全球语料库集群建设与研究”（2022SYLPY004）的阶段性成果。

法、语义、文体标注。数字化法语宝典界面友好、操作简便，可提供简单查找、辅助查找、复杂查找三种检索模式，满足了从学习者到专业研究人员的不同需求。

在法语宝典的编纂过程中，为了便于查找、收集、分析文献资料，法兰西语库（Frantext）应运而生。这一大型语料库项目于1992年开展，自1998年上线以来定期更新，截至2023年12月，法兰西语库收录了从1180—2013年的5,658个文本，约2.7亿字，其中90%为文学文本，包括小说、回忆录、戏剧、诗歌、散文等体裁；10%为科技文本，涉及建筑学、生物学、经济学等领域，成为法国最大、最具代表性的语料库。该库中所有文本均经过词形还原与词性赋码处理，提供索引分析、词表生成等功能，使用者可以根据作者、日期、体裁等元数据进行检索。

随着计算机技术的发展，规模不一、用途多样的法语语料库层出不穷，包括但不限于：（1）大规模参照语料库。以当代法语参照语料库为代表。该语料库涵盖了1945—2014年的法国法语，总词数达到3.1亿词，其中，口语语料与书面语料各占总词数的二分之一，包含演讲、舞台剧、影视字幕、学术论文、新闻、杂志等多种体裁类型（Siepmann *et al.* 2017）。（2）口语语料库。如聚焦于巴黎及其周边地区的巴黎法语口语语料库、重点考察语音的当代法语音系语料库、关注不同交流情境下成人与儿童互动的法语口语语料库等。（3）历时语料库。例如收录9—15世纪末古法语和中古法语文本的中世纪法语语料库，以及针对14—18世纪法语语法著作及相关评论的大型法语语法、评论与论著语料库等。（4）学习者语料库。如学生写作语料库、聚焦学术写作的高阶读写能力语料库等。（5）专门用途语料库。包括政治推文语料库、新闻语料库、网络语料库等。

法语语料库的多样性与代表性渐趋成熟，这不仅为语言研究、语言教学、自然语言处理提供了诸多资源，也为保护语言文化遗产发挥了重要作用。尽管如此，当代法语语料库的平衡性建设仍有待推进。在法兰西语库和数字化法语宝典当中，绝大多数文本为文学类文本，未严格采用平衡语料库的建库模式，而是依据“权威性原则”收录文本（Grieve-Smith 2019），导致许多文本来源于同一作者。目前法兰西语库中可供检索的最新语料仅至2013年，难以全面反映当代法语的真实使用情况。此外，目前相关法语语料库大多采用独立的取样标准，难以与现有的英语（如布朗语料库）及汉语（如CROWN语料库）语料库进行对比研究。本文介绍的frGLOBE语料库项目可以为填补上述空缺提供有益补充。

## 2 frGLOBE 语料库的创建

frGLOBE语料库是按照布朗语料库模式创建的百万词级平衡语料库，主要收集2010年之后出版和发表的法语书面语文本。该库包括生语料、词性赋码、词形还原三个版本，其中，词性赋码与词形还原采用spaCy标注工具。目前，frGLOBE语料

库已上传至“北外 CQPweb 多语种语料库平台”( <http://114.251.154.212/cqp/> )。该在线平台可提供索引分析、搭配计算、词表生成和主题词分析等功能，并随库提供法语词表 ( Word List ) 和短语列表 ( Phrase List )，可用于教学与研究。

2.1 采样方案

frGLOBE 语料库借鉴布朗语料库的采样方案，所收文本类型及文本数量参见表 1。

表 1 frGLOBE 语料库文本类型及文本数量

体裁大类	体裁类型	子体裁代码	子体裁类型说明	文本数量（篇）	
信息类 （374 篇）	新闻	A	新闻报道	44	
		B	社论	27	
		C	报刊评论	17	
	通用	D	宗教	17	
		E	日常技艺及消遣爱好	36	
		F	通俗读物	48	
		G	传记、回忆录等	75	
		H	政府或机构公文及文宣	30	
		学术	J	学术	80
			K	一般小说	50
虚构类 （126 篇）	小说	L	侦探小说	12	
		M	科幻小说	12	
		N	历险悬疑小说	13	
		P	言情小说	30	
		R	幽默	9	
合计				500	

frGLOBE 语料库大体沿用布朗语料库的采样原则，涉及新闻、通用、学术、小说 4 种体裁类型，在 4 种体裁下进一步细分为 15 个子类。对于个别子体裁类型的语料采样，课题组根据法语语言文化状况进行了适当调整，例如在子体裁类型 L “侦探小说”类别下，frGLOBE 语料库主要收集本土侦探/惊悚小说，符合法语小说类型划分，且与法国语言文化传统相契合。此外，在语料采样方面，部分收录时事热点话题及相关评论文本，主要集中于子体裁类型 A、B、C 下，从而保证并维护 frGLOBE 语料库的平衡性与代表性。

2.2 语料采集

按照上述适用于法语的采样方案，frGLOBE 语料库采集 500 个 2,000 词左右的



法语文本，并进行初步加工。语料采集过程主要包括文本收集与取样、文本录入以及语料库元信息标注 3 个主要环节。

2.2.1 文本收集与取样

frGLOBE 语料库所收文本的作者均为法语本族语者，且文本均为原创法语语料，不涉及由其他语言译入法语的翻译文本。此外，由于法语文本整体流通总量低于英语文本，考虑到语料收集的可操作性，相比于布朗家族语料库而言，frGLOBE 语料库扩大了语料的时间范围。新闻类语料发布时间集中于 2020—2022 年，其他 3 类体裁（通用、学术、小说）主要收录首次出版时间为 2013—2023 年的文本，仅存在个别文本发布时间早于 2013 年，所选文本均发布于 2010 年之后。所收语料的时间与数量分布可见图 1。

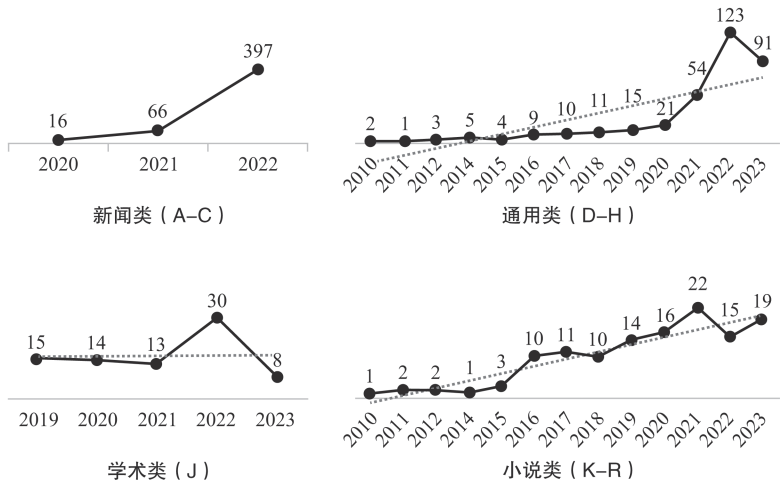


图 1 4 类体裁所收语料时间数量分布

在建库过程中，除考虑语料的时间数量分布外，课题组还兼顾了语料来源及主题的多样性。新闻类语料文本源于法国各大规模新闻媒体，覆盖全国性大报、地区小报、新闻周刊、综合报刊、专题报刊、线上新闻，涵盖政治、经贸、社会、文化、科技、体育、时尚等多种话题。学术类语料主要来自于法国论文检索门户及各学科期刊，涉及数学、医学、文学、历史等多个学科及研究领域。通用类、小说类中的书籍取自多家出版社，体现不同主题。通用类下的子体裁 H“政府或机构公文及文宣”源于法国多个政府部门、多家企业、机构与高校。此外，针对篇幅较长的文本，课题组选取其前、中、后三部分拼接成一个 2,000 词左右的文本，以体现语料的代表性。

2.2.2 文本录入

frGLOBE 语料库所收文本长度为 2,000 词左右，通过正则表达式[a-zA-ZÀ-

Üà-øoù-ÿEœ0-9-] + 进行计数。对于篇幅较短的语料体裁类型,如新闻类、通用类等,则将多篇同类型文本拼接合并,以满足收录词数要求。用于拼合的语料单独保存,并在文件名末尾添加 A/B/C 等以示区分。例如,frGLOBE 语料库所收的第一篇新闻文本由同一报社中的 5 篇新闻拼合而成,因此存储为五个文件:FrA01A、FrA01B、FrA01C、FrA01D、FrA01E,共同构成表 1 子体裁 A 下的第一个语料文本。所有语料均经过人工校对,以 UTF-8 编码格式存储为 txt 文本文件。

2.2.3 元信息标注

总体来看,frGLOBE 语料库共包含 500 个长度为 2,000 词左右的文本,存储为 1034 个语料子文件,并遵循“两位字母语种代码——位字母体裁编码—数字编号—字母编号”的命名格式,便于从文件名中了解语料属性信息,如 FrA01A 表示法语新闻类第一篇语料文本的第一个子文件。除文件命名信息外,frGLOBE 语料库还使用 Excel 表格记录元信息,包括文件编码 (File Name)、体裁类型 (File Type)、字数统计 (Word Count)、出版商 (Publisher)、书名或文章标题 (Article Name)、作者 (Author)、出版年份 (Time)、资源链接 (URL) 8 项基本信息。

2.3 语料版本及应用

frGLOBE 语料库包含生语料、词性赋码、词形还原 3 个版本。生语料指未添加任何语言学标注的文本,词性赋码和词形还原是经过自然语言处理工具 spaCy 的“fr\_core\_news\_lg”模型处理后的语料文本,该模型基于大规模语料训练,具有较高的准确性,能够更好地捕捉法语词汇变化及语法特征。各版本语料示例如表 2 所示。

表 2 3 种语料版本示例

语料版本	语料示例	来源文件
生语料	Les chiffres sont néanmoins alarmants	FrA34A
词性赋码	Les_DET chiffres_NOM sont_VER néanmoins_ADV alarmants_ADJ	FrA34A_POS
词形还原	le chiffre être néanmoins alarmant	FrA34A_LEMMA

表 2 以 FrA34A 语料文件中的一个小句为例,展示了 3 个版本语料之间的区别。其中,词性赋码版语料以“单词\_词性码”的格式呈现,例如冠词的词性赋码为 DET、名词为 NOM、动词为 VER、副词为 ADV 等,可用于快速检索特定法语句式。词形还原版语料则将文本中所有发生屈折变化的单词逐一替换为该单词的原形,如将阳性复数定冠词 les 替换为阳性单数定冠词 le,将阳性复数名词 chiffres 替换为阳性单数原形 chiffre,将直陈式现在时复数动词 sont 替换为原形 être,将阳性复数形容词 alarmants 替换为阳性单数原形 alarmant,该版本语料可以减少屈折语素的干扰,提

高词频统计和搭配检索的准确性。

上述 3 个版本的 frGLOBE 语料均可在“北外 CQPweb 多语种语料库平台”在线使用。平台提供简单查询和 CQP 语法检索两种检索模式,既能满足普通语料库用户快速查询特定单词、短语的需求,又能够支持对语料库进行更加灵活、精准的检索和分析。

### 3 研究展望

#### 3.1 基于语料库的法语研究

frGLOBE 语料库的代表性与平衡性为开展法语研究提供了丰富语料。通过检索 frGLOBE 全库,可以深入探究 2010 年之后法语的词汇、短语、句法、篇章等特征。此外,由 frGLOBE 生成的词频表、短语表,可应用于法语学习与教学,以及基于语言使用实例的词典编纂。同时,frGLOBE 语料库中的新闻、通用、学术、小说 4 类体裁文本也可单独成库,为开展文体学研究、语域变异研究提供了实证基础。

frGLOBE 语料库与布朗家族语料库采样方案的对应性有助于开展对比研究,可与布朗家族语料库和其他 GLOBE 语料库组合使用。例如,frGLOBE 可与 CROWN、CLOB 组合使用,开展法英对比研究;可与 ToRCH 系列汉语语料库(2009、2014、2019)组合使用,开展法汉对比研究;也可与上述英汉语料库及北外全球语料库集群中的其他语种语料库(如 deGLOBE 德语语料库、itGLOBE 意大利语语料库)组合使用,开展类型学研究。此外,frGLOBE 中的通用类、小说类文本还可与法国最大语料库法兰西语库中的同类型文本进行组合,由于法兰西语库收录 1180—2013 年的文本,组合使用有助于推动法语历时变化研究。

#### 3.2 后续语料库建设

目前已经上线的 frGLOBE 语料库可以进一步完善。一方面,参考 deGLOBE 的语料采集方案(周顾盈等 2022),后续版本可进一步增加或更新语料元信息,例如新闻语料文本所属栏目、学术文本的 DOI 编码、小说文本的 ISBN 编码及章节信息等。另一方面,后续版本应进一步降低 2014 年前文本的比例,还可以考虑收录社交媒体文本等近十年的新兴体裁类型,使语料库更精确地呈现近十年的法语书面语使用情况。此外,目前 frGLOBE 语料库仅包含书面语文本,在条件允许的情况下,后续可以适当加入口语语料,开展口笔语对比研究。另外,由于遵循相同的采样方案,frGLOBE 语料库还可与上文中提到的 CROWN、CLOB、ToRCH 语料库及其他 GLOBE 语料库共同搭建多语种平行可比语料库,促成多种语言之间的对比分析研究。

## 4 结语

尽管现有的法语语料库类型广泛、用途多样，但当代法语书面语，尤其是 2014 年后法语书面语的平衡语料库仍然具有建设价值。在“北外全球语料库集群”项目的支持下，frGLOBE 语料库基本沿用布朗语料库的采样方案创建，研究者可以基于 frGLOBE 独立开展研究，也可以直接与布朗家族语料库、GLOBE 系列语料库及其他采用相近取样方法的语料库组合使用，为法语教学与研究提供了丰富资源。frGLOBE 语料库可通过“北外 CQPweb 多语种语料库平台”在线使用，研究者可根据实际需求使用平台提供的检索、词表等功能。frGLOBE 语料库遵循“共建共享”理念，希望并倡导更多同行加入语料库建设和使用，共同促进法语教学与研究的发展。

### 参考文献

- GRIEVE-SMITH A. Building a representative theater corpus[M]. New York: Springer International Publishing, 2019.
- SIEPMANN D, BÜRGEL C, DIWERSY S. The corpus de référence du français contemporain (CRFC) as the first genre-diverse mega-corpus of French[J]. International Journal of Lexicography, 2017, 30(1): 63-84.
- 周顾盈, 宋瑛明, 舒哲, 等. deGLOBE 当代德语书面语平衡语料库的创建[J]. 语料库语言学, 2022 (2): 136-144.

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

# 人教社中小学教材语料库 (PEPTC) 的研制\*

人民教育出版社 黄攀伟 胡 源 杨亚欣 张博翎

**提要：**人教社中小学教材语料库 (PEPTC)<sup>1</sup> 是人民教育出版社自行研发、集语料和语料分析工具于一体的教材语料库，分为教材语料、非教材古代汉语、非教材英语三个子库，规模近 2,500 万字节，具有数据质量高、标注针对性强、功能完备且易于使用等特点。它采用浏览器-服务器 (B/S) 的体系架构，为用户提供信息概览、语料分析、资源管理和数据智能分析工具等功能，主要服务于人教社教材、辞书、教育图书的编研出版，同时也为教材语言研究提供帮助、为数字出版积累资源。本文详细介绍了 PEPTC 的建设背景、设计架构、数据处理原则、语料情况、检索和使用方法，总结了该库的特点、价值和发展方向。

**关键词：**教材语料库、人教社中小学教材语料库、PEPTC、语料检索、分析工具

## 1 引言

语料库是语言数据资源的具体形式之一，是语言服务和语言信息处理的重要支撑。近年来，语料库研究得到越来越多的关注和重视，随着技术更新换代，“从最开始的面向语言学研究的言语材料集合，到如今支撑知识挖掘和发现的深度标注知识资源，语料库及相关研究在深度和广度两方面都得到了充分的探索”（黄水清、王东波 2021：4）。

教材语料库是语料库中的一个特殊门类，主要指以教材中的课文和练习为建库对象的语料库，其语料来源多为书面语或用书面语转写的口语，库内的语料根据需要进行赋码和标注，并具有检索功能（罗庆铭 2017）。我国教材语料库的构建工作有着起步早、发展慢的特点。20 世纪 80 年代，北京语言学院（今北京语言大学）和北京师范大学都曾建设收录当时部分语文教材的语料库。21 世纪初，教育部语信司与厦门大学联合成立厦门大学国家语言资源监测与研究教育教材中心，建设了现代汉语语料库，收录 2016 年以前出版的部分对外汉语教材、中小学语文教材及其他学科教材语料。近些年来，一些高校学者也敏锐地意识到了教材语料库的重大意义，因而自建语料库进行相关研究。总的来说，这些中小学教材语料库主要存在以下几

\* 黄攀伟为本文通信作者。

作者贡献：

黄攀伟：选题构思、初稿撰写、研究方法、数据分析、讨论结论、修改润色、字数占比（60%）；  
胡源、杨亚欣、张博翎：系统开发、架构设计、批量数据处理、修改润色、字数占比（40%）。

方面问题：（1）年代久远，语料陈旧；（2）取材时段短，难以提供历时研究线索；（3）覆盖学科少，以语文为主；（4）大库套小库，大库语料来源包含教材，但不能单独检索分析教材语料；（5）标注不详，对每一条语料的具体来源未作说明；（6）提供的检索和分析工具有限，不能满足研究者导出、统计、量化分析等需求；（7）未开放权限，难以调用。

教材语料对于新教材的编写修订、教材语言的研究、学生工具书的编写等具有重要价值。有鉴于此，人民教育出版社（简称人教社）自主开发了人教社中小学教材语料库（People's Education Press Textbook Corpus，简称 PEPTC）。作为集教材编写和研究于一体的出版单位，人教社一方面在开发建设的过程中具有天然优势——基于社内多年教材、工具书和教育类图书的编修积累，获取一手资源；另一方面，该库的建成也对人教社图书品质的提升大有裨益——通过检索、统计、分析获得可靠数据，反哺图书的策划、编写、出版和数字资源的深度开发。

## 2 系统架构设计

从资源管理的角度来说，PEPTC 依托人教社“资源管理及智能服务平台”进行开发（见图 1）。该平台通过建立统一的出版资源管理中心，以内容资源资产作为核心要素，以“统一存储、分级管理、共享使用”作为建设原则，采用基础层、能力层及应用层的三层式架构，提供资源类、应用类及工具类的三重服务，使整个业务运转过程围绕内容资源的创造、转化、更新和增值进行，既可以为教材编研出版提供必要的资源服务和知识服务，还可以为后端数字产品的开发提供完整的内容资源储备，进而实现出版资源的数字化、知识化、可管理和再利用。PEPTC 的资源来源主要是该平台的中小学教材 XML 文件和非教材 PDF 文件。

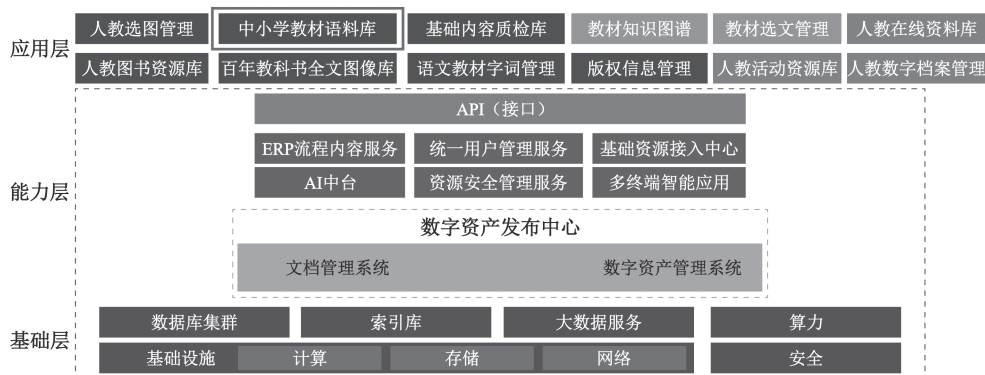


图 1 教材语料库在资源管理及智能服务平台中的定位

从架构设计来看，PEPTC 采用微服务的分布式架构，应用客户端与业务逻辑实现层采用服务调用模式，并同时考虑微服务架构下的服务设计、协同、调度、编排、



治理、数据分片及同步、事务完整性等方面。PEPTC 从下到上共分为数据源、数据湖、应用服务、API 服务、门户 6 个层面。其中，数据源就是系统数据的源头，分为教材和非教材两个部分，教材以 XML 文件为主，非教材以 PDF 文件为主。数据湖主要用于发现更多标引数据，增强后续的分析、提炼、查询等能力。数据湖主要用于存储经过数据挖掘、机器分析的语料。应用服务主要为上层 API 服务提供语料数据智能分析的处理能力（包括一些 AI 模型），包括分词与词性分析、词汇分级、情感分析、实体抽取、实体关系抽取、文章关键字抽取、文章概要抽取、图片描述生成、图片标签生成、图像物体识别、文本纠错功能。API 层的主要目的是为业务场景提供给前台应用的接口能力。门户指的是系统和用户交互的终端，方便用户通过各类终端使用系统功能。图 2 描述了系统各层的功能以及层与层之间的关系，从而可以直观地反映整个系统的架构。

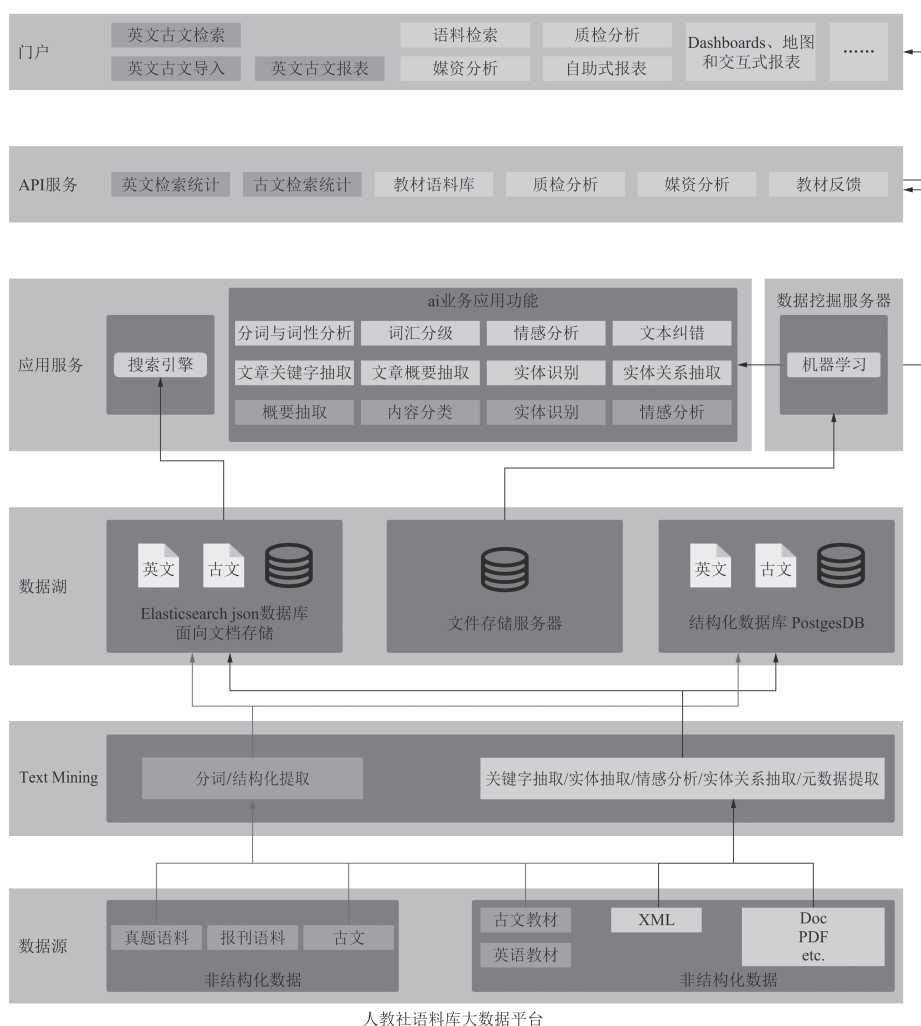


图 2 语料库与大数据分析平台整体架构

### 3 数据处理原则

从数据处理的角度来说,以教材 XML 的数据处理流程为例(见图3),PEPTC 首先通过“资源管理及智能服务平台”的多媒体资产管理系统(Digital Asset Management,简称 DAM)接口,根据图书选题编号读取未入库的教材 XML 结构化文件,并提取其结构化内容,通过大数据平台分析后保存到 Elastic Search 数据库中,为后续检索提供基本数据。另一方面,教材 XML 的结构化数据和教材相关的元数据则被存入 PostgreSQL 数据库(PG),并通过 PostgreSQL 数据库的 ID 建立 ES 与 PG 的关联,方便回溯。

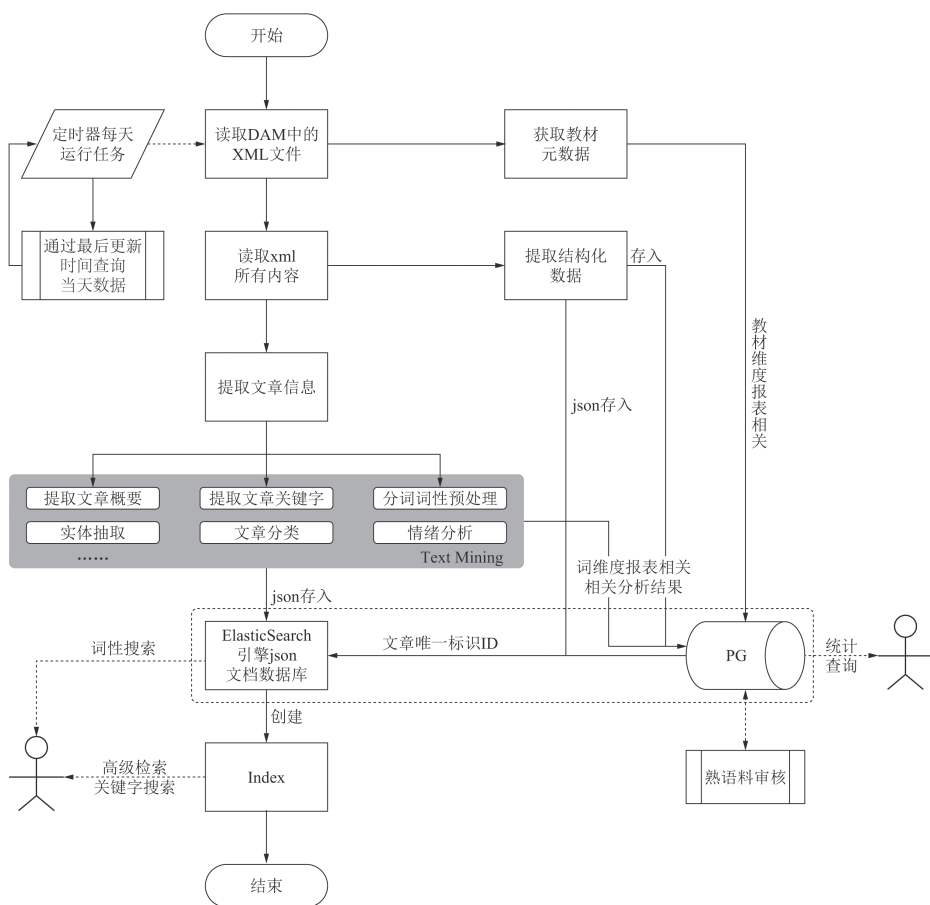


图3 教材语料数据处理流程

总体来说, PEPTC 的数据资源建设遵循以下原则:

(1) 元数据规范。语料库的数据源以教材 XML 文件和社内基础资源为主,通过多种分析方式进行数据挖掘,提取出目标对象的结构化信息,并根据既定的完整内

容及元数据规范，对所有素材资源进行 XML 结构化处理，实现资源碎片化，提供内容的精确定位，并结合多元化元数据为教材数字出版的后续数字化转型和创新提供基础支持。

（2）海量存储。语料库依托平台使用了面向海量数据的存储技术，保障了海量素材资源的集中安全存储，并根据内容使用规则，实现内容的自动归档和备份。

（3）引入数据智能分析软件。平台引入了大数据分析软件（Magellan）和人教社文档管理系统（ECM）、多媒体资产管理系统（DAM）实现原生集成，通过搭建数据智能分析应用平台，提供数据智能分析服务能力，实现数据采集、规整、运算和存储，建立人教教材语料库，利用人工智能技术，对存量教材语料及大量素材进行训练，构建适合中小学教材的相关算法模型（见图 2）。

（4）面向服务。平台使用标准接口，提高了系统的兼容性和开放性。针对辞书编辑室及其他学科编辑室的业务需求，平台提供定制化的业务接口。同时，平台支持通过 API 形式向内部及外部机构提供在线服务，提高了资源的可重复使用率和交互操作性。

## 4 语料情况

整个语料库包含“教材语料、非教材古代汉语、非教材英语”3 个子库，3 个子库的资源在后台单独存储。其中，教材语料是整个语料库的核心，目前已入库教材近 300 种，语料近 2,500 万字节，字种数 6,000 余个，词种数 11 万余个，涵盖语文、历史、政治、数学、物理、化学、生物等 18 个学科。随着教材资源数字化的不断推进，还将有更多现行教材和以往教材不断入库（见图 4）。此外，非教材古代汉语语料以先秦诸子、二十四史、古代诗词的常见选本选篇为主，非教材英语语料库以历年中高考真题、权威英语报刊资料、英美文学等为主。为方便客户使用，3 个子库集成于同一页面，用户可通过首页侧栏进行选择，调取相应子库的语料。

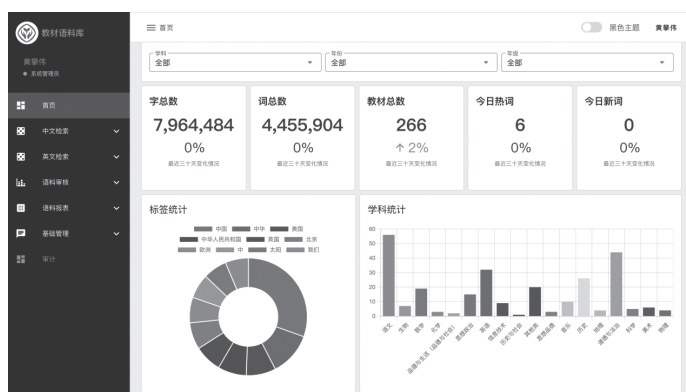


图 4 PEPTC 首页

在语料加工方面，为满足不同类型的研究需要，语料库提供“生语料”和“熟语料”两个选项。生语料检索得到的是语料的自然文本状态，仅包含文本的篇章分类和少量信息标注。熟语料检索得到的是经过机器自动切分和人工干预后得到的分词和词性标注结果。比如：使用正则表达式“高??兴”进行检索时，如果选择“生语料”，则凡是符合检索要求的形式都会出现在检索结果中，既有“看过戏的少年们也都聚拢来了，高高兴兴地来讲戏”“我们只应该高高兴兴地多说几遍”，也有“作者于宋高宗绍兴三十二年（1162）南归”“新巴比伦王国被伊朗高原上兴起的波斯帝国消灭”；如果选择“熟语料”，由于经过了分词处理，则只会出现“高高兴兴”的例子。

5 功能实现

为满足使用者需要，PEPTC 采用浏览器—服务器的体系架构、语料库与语料库分析工具合二为一的模式，通过友好的交互界面，为用户提供信息概览、语料分析、资源管理和数据智能分析等功能。

5.1 系统概览

PEPTC 首页提供“学科”“年份”“年级”3 个基本检索栏，同时展示语料库的整体概况（见表 1）。

表 1 PEPTC 首页概览信息

概览信息		代表信息
数据统计	字总数	语料库规模
	词总数	
	教材总数	
	今日新词	语料的当日检索情况
	今日热词	
标签统计		自动标注的高频标签
学科统计		入库教材的学科和数量
编订统计		教材修改时间和次数的关系
词云画像		教材高频词语
入库统计		语料入库时间和数量的关系
非教材英文语料情况		不同类型语料的数据和占比
非教材古文语料情况		不同类型语料的数据和占比

5.2 检索和分析

传统的检索型语料库正在逐步向分析型语料库转变，语料分析工具功能多样、结果准确已经成为语料库评价的重要标准之一。PEPTC 依托前期的架构设计和基础资源，可以提供的分析工具包括：多重条件过滤、多模式检索、多维度统计和对比、近义词辨析等，所有统计和分析结果均可一键导出。

（1）多重条件过滤。检索前，用户可选定“作者”“年代”“查询对象（课文/注释/标题）”“语料范围（生语料/熟语料）”等过滤条件，并决定是否“去重”、范围是否限定在“统编三科”或“最新教材”。初步检索后，可以对“学科”“年级”“年份”“阶段”“册次”“使用要求（必修/选修/选择性必修）”“教材细类”“使用季别”等信息进行进一步筛选（见图 5）。



图 5 PEPTC 教材语料中文检索界面

（2）多模式检索。汉语语料支持 5 种检索模式，即关键字检索、模式检索、词性检索、正则检索和知识检索（见表 2）。检索结果中会呈现教材名称、年份、年级、学科和责任编辑等信息，同时提供全文预览和出处链接。

表 2 PEPTC 汉语语料检索方式一览

检索方式	检索举例	结果举例
关键字检索	中国人	天安门广场升起第一面五星红旗 <u>中国人民</u> 从此站起来了！（生语料） 于是有人慨叹曰： <u>中国人</u> 失掉自信力了。（熟语料）
模式检索	AABB	他的诗读起来 <u>往往朗朗</u> 上口（生语料） 接着走出一个小旦来， <u>咿呀呀</u> 地唱。（熟语料）

（待续）

(续表)

检索方式	检索举例	结果举例
词性检索	打 n	你是景阳冈打虎的好汉 <sup>2</sup>
正则检索	高??兴	新巴比伦王国被伊朗 <u>高原上兴起的</u> 波斯帝国消灭 (生语料) 我们只应该高 <u>高兴兴地</u> 多说几遍(熟语料)
知识检索	汉字	……丰富多样的文字语言文字是文化的重要载体， 是人们交流的工具。我国的通用语言文字是普通话 和规范汉字，推广普及国家通用语言文字是宪法规定 的责任。……

英语语料支持 8 种检索方式，包括关键词检索、任意词检索、通配符检索、同义词检索、词性检索、全部形态检索、“非”检索等（见表 3）。检索之后，可以选择“频率”模式优先查看频率，也可以选择“上下文”模式优先查看文本信息。

表 3 PEPTC 英语语料检索方式一览

检索方式	检索举例	结果举例
关键词检索	one student	How does <u>one student</u> educate his mother about waste ?
任意词检索	give * to	Did someone <u>give it to</u> you ?
通配符检索	f?r	得到包含 faster, for, finger 等词的例句
同义词	see	得到包含 look 等同义词的清单, 勾选后出现频率或例句
词性检索	see/PRP	see them, see it, see us 等选项，及其对应文本链接
全部形态检索	see	saw, seeing, seen 等选项，及其对应文本链接
“非”检索	look/NNP <sup>3</sup>	look for, look forward, look carefully 等选项，及其对应文本链接

（3）多维度统计和对比。语料库的统计功能主要服务于工具书编写和教材语言的研究，PEPTC 提供频率统计和搭配统计工具。

频率统计方面，其功能单独放置在“语料报表”板块，可以根据选定的语料范围对教材中文字词、非教材中文字词、教材英语词汇、非教材英语词汇的情况进行统计，并详列每条数据的具体来源、频次、频率、累计频率等。

搭配统计方面，汉语的搭配统计选项在检索页面中，可分别显示被检索词的左搭配词和右搭配词，以及搭配词的相关信息，包括词性、词总数、占比、词频、语义韵等（见图 6）。



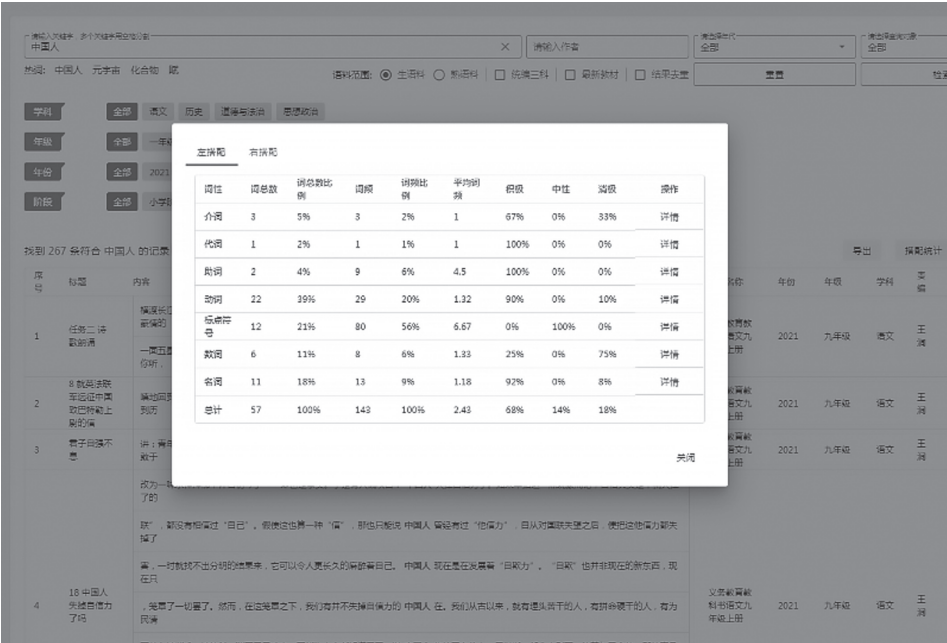


图 6 PEPTC 汉语语料的搭配统计

英语搭配的统计和对比功能合并单独的选项卡下，用户可以对两个单词或短语前后搭配的相关情况，包括成分性质、词频等（见图 7）。



图 7 PEPTC 英语语料的搭配统计

系统支持按照使用年份、教材细类、阶段、学科、教材、课文、年代等条件对教材语料中的字进行统计分析（见图 8），显示每个字出现的教材总数量、该字出现

的总次数、频次/字总数、累计频率，其中累计频率是将之前的字频率累加，当累加值达到总频率的 90%时，将之前的所有字定义为高频字。语料库中词汇的统计方法与此类似。此外，系统还支持导出功能，为教材编写者和研究人员提供参考数据。

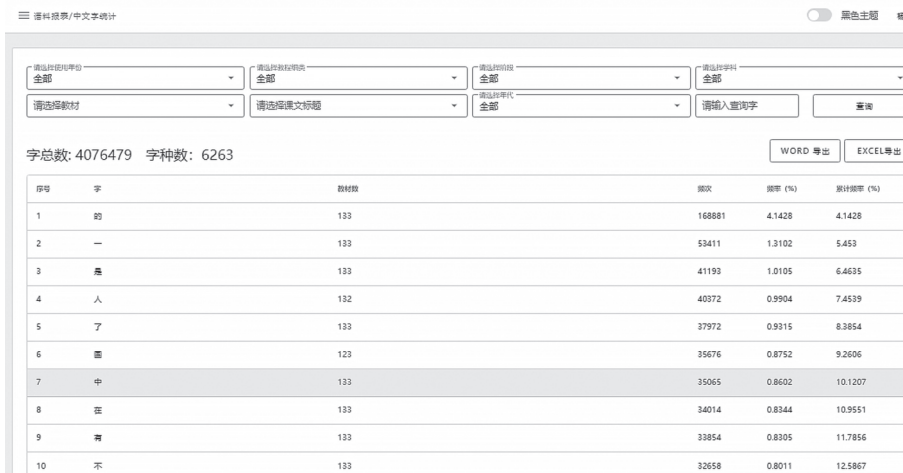


图 8 PEPTC 中文教材语料的字统计

(4) 英语近义词辨析。通过对已有辨析材料的数字化，用户可在该选项卡下检索，查询英语近义词在词义、用法等方面的区别，并获得相关短语搭配和例句（见图 9）。



图 9 PEPTC 英文语料的近义词辨析

5.3 系统管理

(1) 语料审核与资源管理。为了提高语料库的优化效率，PEPTC 打破了开发者和使用者之间的界限，为用户开放人工干预和语料导入权限。一方面，受限于目前的自然语言处理技术，语料库标注仍存在一些偏误。对此，普通用户可以通过方便

友好的人工干预界面，编辑或删除模型生成的标签、实体对象、实体关系等，解决机器分词、词性标注、命名实体识别等过程中出现的问题，管理员对这些修改进行审核。另一方面，语料库也开放了语料导入功能（见图 10），对于格式符合要求、质量有所保障的语料，用户可以自行导入。

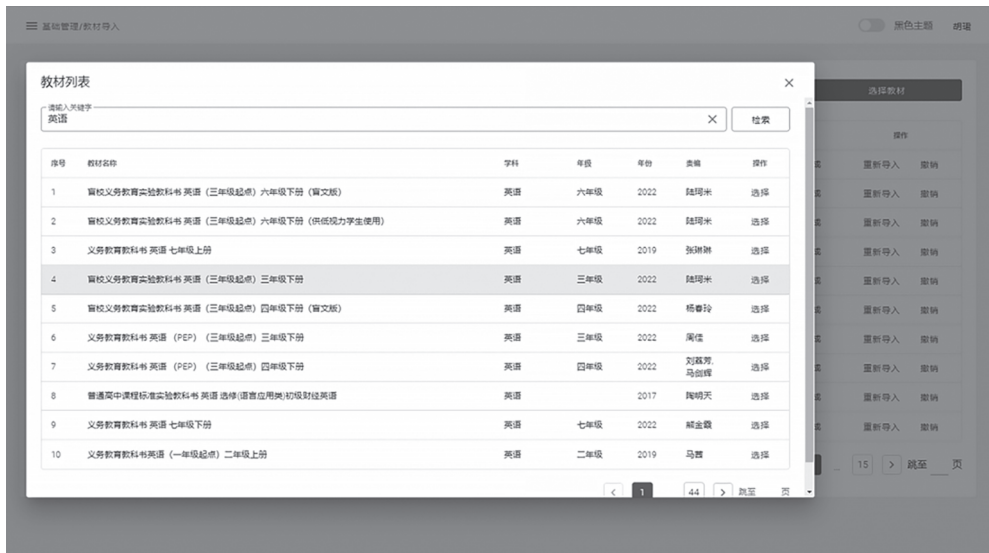


图 10 PEPTC 教材语料的导入

（2）权限管理。为了能够灵活配置角色的权限，系统设计了权限组，可以根据实际需要执行配置角色所对应的权限组，权限组包括首页权限组、报表权限组、审核权限组、管理权限组、检索权限组、英文权限组和工具权限组，再将用户与角色绑定，即可实现不同用户的权限管理，如图 11 所示。

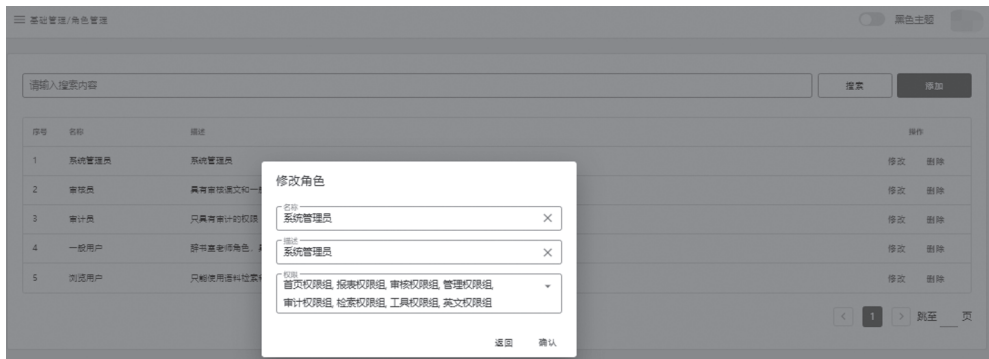


图 11 PEPTC 的权限管理

## 6 建设特点

从语料库语言学工程实践的角度来看，PEPTC 具有以下特点。

(1) 规模中等。罗庆铭(2017)调研发现，目前市面上的教材语料库规模差距很大，一般从十万字符到几百万字符不等。相对而言，PEPTC 在教材语料库中属于规模较大的一个，但与动辄上亿字符的通用大型语料库相比，其规模仍然较小，这也是各类教材语料库的普遍特征。PEPTC 的建库规模主要受以下 3 点原因制约。第一，自身定位为中小学教材语料库，语料选取范围受限；第二，部分语料仍在入库过程中，建设尚未完成；第三，考虑到版权等相关问题，目前仍以人教社出版物语料为主。

(2) 语料质量过硬。这一方面得益于原始资源内容质量高，教材质量检查要求远高于一般图书<sup>4</sup>，因而保证了原始数据的高质量；另一方面得益于规范的数据处理流程，原始文件通过转码或 OCR 技术转为 XML 文件后，再辅以人工纠错，从而获得了高精度的数据资源。

(3) 标注针对性强。为满足后期检索和各类分析的需要，PEPTC 对教材信息和语言信息进行了尽可能详细的预处理和标注。教材信息包括学段、学科、版次、年级、学期、册次、使用季别、内文类别（课文、注释、标题等）、责任编辑等。语言信息包括语种、时代等基本信息，以及分词、词性、实体、语义韵等语言学信息。

(4) 功能完备、易于使用。张永伟、吴冰欣(2023)认为，语料库分析工具能对大规模语料实现更高效的多样化分析才算“功能完备”，便于操作且使用者无需编写程序或复杂的查询表达式才算“易于使用”。PEPTC 已经满足了这两项要求，这主要得益于语料库规划之初具有前瞻性的顶层设计和服务理念，以及在对底层数据进行处理、储存、分析时所采用的先进技术，特别是近年来人工智能技术在自然语言处理领域取得的成果。比如，人教社技术团队使用训练数据集，根据不同业务需求，对包括 BERT、LSTM 等神经网络模型在内的模型进行训练（见图 12），逐步调整参数、提取特征及模型收敛，以达到最佳的模型训练结果，并使用 20% 未经训练的数据对训练好的模型进行交叉验证，从而完成汉语分词、词性标注、实体识别等任务，满足语料库使用者的需求。

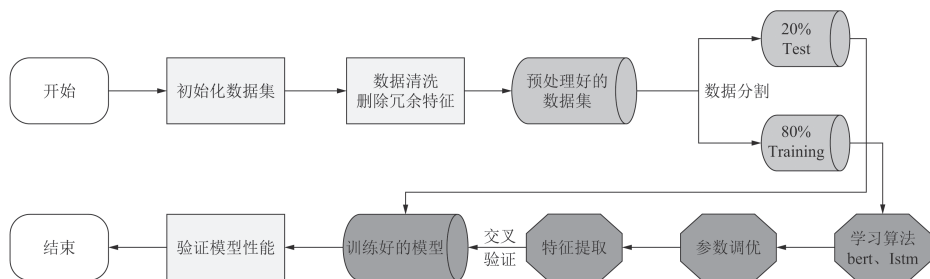


图 12 模型训练全流程

（5）研发与应用同步。在 XML 文件入库、网页基本框架搭建完成之后，社内编辑成为第一批用户，试用系统、反馈改进意见、将语料分析结果应用到编校实践当中。信息技术部门的工程师则三线并行：导入新语料、根据反馈调整平台、按照原定计划推进语料库建设。这种开发模式使语料库资源能够第一时间发挥价值，同时也能及时得到调整和完善。

## 7 结语

### 7.1 语料库的价值

正如詹卫东等（2019）所述，经过数十年的发展，语料库在教材编写、词典编纂、语言教学等语言应用领域，以及语言本体研究等方面都发挥了重要作用。不仅如此，掌握大量优质的精标数据，也是为后续开发利用提早布局、储备资源。

首先，PEPTC 直接服务于人教社教材、辞书、教育图书的编研出版，特别是学生工具书的编研出版。实际上，“在计算机出现之前，研究者——特别是词典编纂者，也有语料库，只是规模小、范围窄，因而难以在学术界形成气候”（黄昌宁、李涓子 2002：4）。计算机出现之后，早在 20 世纪 60 年代，美国的 Heritage 出版社就曾为编纂中学生词典，专门设计了 AHI 语料库。到了 1987 年，柯林斯出版社在 COBUILD 语料库的支持下出版了《柯林斯 COBUILD 英语大词典》，被认为是语料库词典的发端。国内也从 20 世纪 80 年代开始了双语语料库辅助英汉词典编写的探索。PEPTC 的建设对提升人教社学生工具书的编纂质量功不可没，以《新编学生字典》（第 3 版）、《新编学生词典》为代表的系列工具书，借助 PEPTC 在实现教材字词全覆盖、完善释义、增补教材用例、选配教材名物插图、设置学习板块等方面得到显著提高。

其次，PEPTC 也为教材语言的研究提供了可靠依据。在用字、用词、用语统计分析，字词表对比分析、字词表分布分析、词语搭配抽取及分析等方面，PEPTC 均能提供可靠的数据和完备的分析工具，从而为教材字词语难度设计、中小学字词教学方法、学生语言习得规律、教材跨学科表述等领域的研究提供支持。

最后，以出版社为研发主体的语料库建设过程，也是实现优质出版资源标准化、规范化管理的过程。在这个过程中，出版社一方面获得了精标细注的宝贵数字资产，为后端数字产品的开发提供了完整的内容资源储备；另一方面也打破了系统内部的信息孤岛、促进了数据的互联互通，为激发和提升数据资源应用价值、融入数据智能分析应用研发成果、开启教材出版新模式赋能。

### 7.2 发展方向

对于 PEPTC 未来的发展，从近期来看，项目三期及未来规划的主要目标包括如



下方面。

(1) 扩充语料。语料的质量和规模是语料库的核心资产。精细标注的语料需要花费较多时间,因而 PEPTC 的计划内语料按照需求程度的高低排期入库。在教材语料、古代汉语语料、英语语料等方面都进行大规模扩充的同时,将部分人教自主版权的工具书纳入语料库。

(2) 完善功能。便捷易用是语料库建设和应用之间良性循环的重要基础。未来 PEPTC 一是要进一步加强服务应用的开发,逐步实现英汉双语对照检索与分析、各类英汉工具书的统一检索与个性化展现等功能。二是要提升模型性能,目前语料库中对现代汉语分词和词性标注的结果尚可,但对近代汉语和古代汉语的分词能力还十分有限,依靠近年来古代汉语分词模型研究的新进展,PEPTC 有望在这方面获得更好的分词结果。三是进一步增加和改进分析工具,如增加繁简转换、汉英互译、音视频转录等智能工具。

(3) 尝试构建学科互通的知识库。从数据点状分布的简单语料库,走向树结构的数据库,再走向图结构的知识库,是语言资源建设的大势所趋。与一般的结构性语料库相比,语言知识库需要对语料进行更为深入的语义和句法标注。得益于近年来预训练模型的快速发展,PEPTC 已经在这方面进行了尝试,并开放了“知识检索”功能,但为达到预期效果仍需开展大量工作。未来 PEPTC 要继续利用知识图谱和预训练模型技术,对工具书和教材内部的知识元进行关联和抽取,一方面优化“知识检索”标签下的检索结果,另一方面通过图的形式,将知识元之间的关联直观地呈现出来。

### 注释

- 1 PEPTC 语料库的建设工作得到了人民教育出版社领导黄强、付天华、朱于国、王强等的关心和支持。在研制过程中,有多位老师和技术人员给予大力支持,包括人民教育出版社辞书室谢仁友、由明智、金欣欣、周彩萍、周煜、程少峰、李雪、宋晓蓓、张祝香等,以及人民教育出版社信息技术部李尚通、王兴华、胡珺等。此外,测试用户也对 PEPTC 语料库提出过很多宝贵的改进意见和建议。在此一并表示诚挚的谢意。
- 2 词性检索以计算机中文分词和词性标注为基础,所以只提供熟语料选项,但从结果来看,目前的差错率较高,仍有进一步提升的空间。
- 3 第一个“N”表示 not, look/NNP 即查询 look 后面所加成分不是 NP(名词性成分)的情况。
- 4 图书质量检查中,一般图书要求差错率低于 1/10000,教材要求差错率低于 0.25/10000。

### 参考文献

黄昌宁,李涪子.语料库语言学[M].北京:商务印书馆,2002.



- 黄水清, 王东波. 国内语料库研究综述[J]. 信息资源管理学报, 2021 (3): 4-17.
- 罗庆铭. 教材语料库的建构与应用: 以新加坡小学华文教材为例[M]. 北京: 中国社会科学出版社, 2017.
- 詹卫东, 郭锐, 常宝宝, 等. 北京大学 CCL 语料库的研制[J]. 语料库语言学, 2019 (1): 71-86.
- 张永伟, 吴冰欣. 基于网络的第四代语料库分析工具核心功能评介[J]. 当代语言学, 2023 (4): 611-624.

**通信地址:** 北京市 人民教育出版社课程教材研究所辞书研究中心

# 《历时构式语法的节点与网络》 述评\*

北京邮电大学 房印杰

Lotte Sommerer & Elena Smirnova (eds.). 2020. *Nodes and Networks in Diachronic Construction Grammar*. Amsterdam: John Benjamins. xxvii + 363pp.

## 1 引言

历时构式语法研究侧重以构式网络模型解释语言的历时变化 (Traugott & Trousdale 2013), 但其在理论建构与研究路径上均存在进一步拓展空间。譬如, 语言历时演变中的构式化与构式演变的相关性 (Barðdal & Gildea 2015), 构式网络中的节点与联结的主导性问题 (Hilpert 2018)。Sommerer 与 Smirnova 主编的论文集《历时构式语法的节点与网络》对上述问题进行了深入探讨, 代表了该领域的最新成果。该文集在理论层面, 尝试厘清历时构式语法的诸多概念及假设。在方法层面, 基于大规模历时语料阐述构式的纵向/水平联结关系, 构式节点内与节点间的联结变化, 并尝试基于神经网络模型量化构式联结的多维性。在研究对象上, 突破了传统意义上的近义构式, 转向基于类比推理的构式变体与语法范式。

## 2 主要内容

文集共收录 10 篇论文, 分为 4 个部分。第一篇论文针对历时构式语法研究的诸多问题与争议作了深入的理论梳理, 指出了亟需回答的 7 个问题, 涵盖构式节点、构式的纵向/水平联结、构式网络模型的拓展。其余 9 篇论文则分别基于大规模历时语料分析典型构式的历时变化, 或通过缜密的内省思辨, 尝试回答上述问题。

第一章全面梳理了历时构式语法中的争议, 涵盖节点的本质, 构式网络的架构, 构式语法视角下语言演变的实质。基于此, Sommerer 与 Smirnova 提出亟待解决的 7 个问题: (1) 如何以网络模型解释构式节点的产生与消亡? (2) 如何辨析构式化

\* 本文系北京市社科基金项目“基于统计模型的中国学习者英文文体特征研究”(20YYB006)的阶段性成果。

与构式演变？（3）在构式网络中，节点之间存在何种联结？（4）构式变体、构式家族、范式的理论地位如何？（5）如何以现有网络模型阐释节点外部联结的重构？（6）一般认知能力（如类比推理、常规化、抽象化、范畴化）对构式网络的结构与重构有何影响？（7）如何拓展现有网络模型？本章重点阐述了一系列基本概念的模糊性问题。譬如，作者提出，构式化定义违背构式的基本定义（形式-意义的配对）（Goldberg 2006），其与构式演变的界定存在重合，在实证研究中缺乏操作性标准。第二部分包括两篇实证性论文，探讨了构式节点的产生、演变和消亡。第二章聚焦构式化概念的模糊性，即构式化概念既可以指构式转变的过程，亦可指构式转变的最终结果，即生成新的构式节点。通过对 into-致使构式的历时演变分析，Flach 发现该类构式中 into 之后的成分呈现“名词—动名词—动名词小句—动词”的演变过程。据此，她提出构式化应专指新构式节点的生成，而渐进性演变过程则应称为构式浮现（constructional emergence），二者均属于构式演变。通过重新定义上述概念，Flach 侧重凸显构式节点内部与节点间的动态性联结。第三章侧重探讨构式节点的消亡过程。Sommerer 基于古英语语料，分析了“指示词 + 形容词性所有格”结构的逐步消亡，从而对构式节点的消解过程进行阐释，她发现构式节点的消亡源自构式节点之间的竞争。在该过程中，基于频次效应所获得的高认知固化节点会逐步获得使用优势，而低认知固化节点则会逐步消失。从古英语的早期到晚期阶段，“指示词 + 形容词性所有格”结构的使用频次逐步下降，并最终被“指示词 + 名词”与“形容词性所有格 + 名词”所取代。该过程源自频次效应与类比推理所导致的构式网络演变。

第三部分包括 5 篇实证性论文，探讨了构式网络中的纵向与水平联结。其中，第四、五章侧重阐述节点的纵向联结，第六至八章聚焦节点的水平联结。第四章讨论了荷兰语中的强化性反身动结构式所经历的构式化过程，Gyselinck 聚焦该类构式在纵向联结（抽象度层面）上的变化，即一系列子图式的抽象度会出现不同程度的上升与下降，该变化同步发生。同时，同级别的子图式会呈现迥然不同的能产性与抽象度变化。对于构式节点与构式纵向/水平联结，她提出应进一步从联结主义计算模型视角探讨构式节点的多种联结关系，但不应完全抛弃构式节点概念，这与 Schmid（2016）完全基于构式联结关系讨论构式变化的观点不同。第五章着力区分抽象度与能产性的异同，作为一组近义概念，学界往往将二者等同视之。通过分析 way 构式的历时变化，Perek 认为构式抽象度应包括两个层面，构式内部成分（如槽位）的抽象度和构式整体的抽象度。前者与能产性直接相关，但后者与能产性的相关性并不明显。换言之，内部成分的抽象度与构式整体抽象度变化并不同步。第六章阐述构式网络与受益型与格构式历时演变的相关性。Zehentner 和 Traugott 基于构式变体概念与构式网络中的水平联结，分析早期现代英文中的目标构式。通过从 EEBO 与 PPCME 语料库中检索双宾与格语句、for 型与格语句、to 型与格语句，然后进行人工判定。将提纯后的语料切分为 4 个时期，然后通过显著共现词分析分析不同时期中动词与所在构式之间的预测程度变化。结果显示，双宾与格语句、for

型与格语句、to 型与格语句通过水平维度的类比推理机制逐步形成了与格构式变体。

第七章聚焦次要谓语构式。基于中古英语 PPCME2 语料库, Percillier 分析了 to/for/as/into 4 个介词引导的目标结构的构式网络变化, 及其潜在相关性。结果显示: (1) 从古英语到早期现代英语, to 型结构的使用倾向逐步被 as 型结构所取代; (2) 4 个介词所引导的目标结构已形成一种近义构式, 可以被判定为构式变体; (3) 语言接触所导致的外来语可与本土英语根据语义、用法的相似性形成构式变体。构式变体的形成涉及构式节点在水平层面的类比推理, 其内部成员在纵向联结层面由一个更为抽象的构式位所统辖。

第八章探讨动词不定式标记 to 的缩略结构。通过分析历时美式英语语料库 (COHA) 中的目标语料, Lorenz 发现: (1) 动词不定式充当宾语时的缩略结构并不普遍适用于所有动词, 而是仅限于上述 3 个目标动词; (2) 高频使用使得 3 种缩略结构从其对应的非缩略结构中脱离出来, 虽然 3 者最初的使用倾向、演变过程不尽相同, 但最终均指向相似地使用型式, 并转向同一个构式型式; (3) 构式节点间联结的重构具有动态性、渐进性。

第四部分包括两篇论文, 分别尝试从理论与实证层面拓展现有网络模型, 以增强对构式演变的解释力。第九章尝试对现有的构式网络模型进行理论拓展, 提出应将 (语法) 范式 (如时态) 纳入其中。Diewald 分析了语法化、构式化、范式三者的相关性, 指出: (1) 范式是语法化的前提, 它是由一系列构式组成的超级构式 (hyperconstruction); (2) 在该构式节点中, 内部构式的意义/功能通过其他成员界定, 该超级构式的意义/功能则由全体成员界定; (3) 范式内部同样包含纵向/水平联结。

第十章力图从研究方法层面拓展、丰富构式网络模型。Budts 与 Petré 尝试通过神经网络模型客观量化构式网络的各种联结。通过分析 be going to 与助动词 do, 他们发现: (1) 构式的纵向/水平联结均包含组合与聚合关系; (2) 水平联结促使 [be [go]ing] [to INF] 的组构性逐步下降, 并最终形成形式与意义的解放, 生成新的构式节点; (3) 基于使用型式的相似性, 类比推理机制使得 do 与其他情态动词 (如 can) 的语义/功能日益趋近, 并最终与情态动词组成近义构式群, 同属一个抽象构式图式。

### 3 要点简评

#### 3.1 节点与联结关系的重构

作为构式语法中的重要概念, 学界一直尝试论证节点与联结的关系。文集从以下两个方面对这一问题做了相关阐释: (1) 节点与联结的主导性; (2) 构式网络中联结关系的多样性。Goldberg (2006) 强调构式是语言的基本单位, 不同构式作为节点, 彼此间形成一个构式网络。根据该论述, 节点显然被作为了构式网络的基础,

而节点间的联结关系则处于从属地位 ( Hilpert 2018 )。与之相比,文集侧重凸显联结关系对于构式网络形成与变化的核心作用,强调以基于多种联结关系的动态网络取代以节点为基础的静态模型。基于该假设,构式节点的形成与内部演变源自其内部成分在形式/意义层面联结关系的变化,构式网络的演变则源自节点间的多种联结关系的变化。这一观点与 Schmid ( 2016: 25 ) 相一致,均凸显多种联结关系对于构建构式网络的核心作用。

针对构式网络中的联结关系,文集聚焦联结关系的复杂性、多维性,突破了前人研究对纵向继承关系的过度强调,转向凸显纵向继承关系与水平竞争关系的同等重要性。前人研究多着力阐述具体构式与抽象构式之间的潜在继承关系,但对构式间的水平竞争关系着墨不多。然而,相关研究 ( 如 Sommerer & Baumann 2021 ) 显示,部分构式节点的生成似乎主要依托与原有构式节点的水平联结,而非纵向联结。换言之,近义构式群之间可能并不存在一个高度抽象的原型性图式,而是由一系列具有家族象似性的半抽象子图式所统辖。

文集重点论述了水平联结关系对构式网络形成与演变的解释力。水平联结被具化为构式变体与范式两种分析路径,前者基于构式间的近义异形,后者则基于构式间的近形异义。第六章和第八章的研究均侧重论述水平联结关系对于构式间联结关系的潜在影响。譬如, wanna/gonna/gotta 三者所形成的近义构式关系主要源自三者在使用型式层面的相似性。对纵向/水平联结的并重显著增强了对构式节点变化、网络演变的解释力。此外,文集将纵向/水平联结与传统语法分析中的聚合/组合概念相融合,第十章就指出纵向/水平联结中均各自包含聚合/组合关系。更为重要的是,文集初步证明构式网络关系极为复杂,具有多维性,纵向/水平联结的二维模型只是对构式网络关系的高度简化。

### 3.2 构式化与构式演变间二分对立的消解

自 Traugott & Trausdale ( 2013 ) 提出构式化与构式演变的概念,如何客观界定、厘清二者的内在联系便成为历时构式语法研究的一个重点/难点。文集梳理了二者在概念层面的重合性,在实证操作层面的模糊性,而后尝试将二者在概念层面的二分对立转变为构式节点内部变化的动态性、渐变性。换言之,构式演变是语言演变的普遍机制,而所谓的构式化则只是针对构式演变的一个分析性概念,对后者的操作界定标准取决于具体的研究目的或问题。基于该论断,一切构式变化均属于构式演变,而构式化只是构式演变过程中量变导致质变的结果。无论是 Flach 所提出的构式浮现与构式化 ( 第二章 ), 还是 Sommerer 的“新构式化”与“替代性构式化” ( 第三章 ), 均凸显构式演变的普遍性、动态性、渐进性,原有的概念层构式化则降级为具体操作层面对特定构式节点形成的判定。

将构式演变作为语言演变的普遍机制,解决了构式化与构式演变在概念层面的



重合性,在操作节点标准上的模糊性。基于此,对构式化的判定标准应基于不同研究对象、研究目的而设定,在具体操作层面则依托统计测量手段(如搭配构式分析)(如 Sommerer & Baumann 2021)。

### 3.3 研究方法的创新

全书除去第一章与第九章,其他 8 个章节均依托大规模历时语料开展实证研究。在具体研究方法上,由简单的描述统计向复杂统计方法(如神经网络模型)转变。特别值得注意的是,部分研究(第七和第十章)已开始依托向量化的统计模型分析特定语言对象的历时演变机制。复杂统计方法能够更为客观、有效地呈现构式网络演变的动态性,构式节点联结关系的多维性。文集所集中论述的纵向/水平联结只是对构式网络联结关系的高度简化,并不能真实反应其多维性,而基于神经网络模型的复杂统计方法为深入挖掘这种多维联结关系提供了可能。

文集对历时构式语法的诸多核心问题做了深入、客观的探讨,但仍存在以下不足。(1)在第二部分,文集论述了节点的诞生与消亡,但并未涉及当代英文中存在的半消亡节点,如 *he was gone* 这种以“BE + done”表达完成体的中古英语残留问题。这使得对节点的演变形式论述不够完整;(2)针对节点之间的纵向/水平联结关系,没有考虑到节点之间的错位继承关系,如 BE 过度生成句便是对“主语 + BE + 静态谓语”结构的错位继承。应该讲,错位继承可以出现在纵向关系,也同样会出现在横向关系中;(3)在论证纵向联结关系时,文集使用了多个术语(如 *allostruction*、*homostruction*、*hyperconstruction*、*megaconstruction* 和 *constructeme*),对上述术语的指代异同,文集并未明确交待。

概言之,该文集的研究初步消解了构式化与构式演变的对立,凸显了构式(内部与节点间)演变的动态性、渐进性,指明了构式网络中联结关系的多维性,着力论证了纵向继承关系与水平竞争关系的重要作用,并尝试将其与组合/聚合关系相融合,凸显多种联结关系在界定构式节点与网络中的关键作用。在理论建构与方法拓展上,该文集的研究做了有益尝试,框定了历时构式语法研究的核心问题,指明了未来的研究方向。

### 参考文献

- BARÐDAL J, GILDEA S. Diachronic construction grammar: epistemological context, basic assumptions and historical implications[C]//BARÐDAL J, SMIRNOVA E, SOMMERER L, et al. Diachronic construction grammar. Amsterdam: John Benjamins, 2015: 1-50.
- GOLDBERG A. Constructions at work: the nature of generalization in language[M]. Oxford: Oxford University Press, 2006.



HILPERT M. Three open questions in diachronic construction grammar[M]//COUSSE E, ANDERSSON P, OLOFSSON J. Grammaticalization meets construction grammar. Amsterdam: John Benjamins, 2018: 21-39.

SCHMID H. Entrenchment and the psychology of language learning[M]. Berlin: Mouton de Gruyter, 2016.

SOMMERER L, BAUMANN A. Of absent mothers, strong sisters and peculiar daughters: the constructional network of English NPN constructions[J]. Cognitive Linguistics, 2021, 32(1): 97-131.

TRAUGOTT E, TROUSDALE G. Constructionalization and constructional changes[M]. Oxford: Oxford University Press, 2013.

通信地址: 100876 北京市 北京邮电大学人文学院

# 《语料库设计与评估：针对语料库代表性的实用性框架》述评\*

上海外国语大学 张琮然 徐海铭

Jesse Egbert, Douglas Biber, & Bethany Gray (eds.). 2022. *Designing and Evaluating Language Corpora: A Practical Framework for Corpus Representativeness*. Cambridge: Cambridge University Press. xii + 284pp.

## 1 引言

作为当代语言学研究的一个分支，语料库语言学因其实证数据的丰富性和可靠性，吸引了越来越多学者的研究兴趣。然而，语料库语言学所涉及的几个相关概念，例如库容、代表性、平衡性和抽样过程，一直是语料库语言学学科探讨的主要话题（Wong *et al.* 2011）。那么，语料库语言学中的“代表性”内涵为何？它与一般应用语言学研究中的数据代表性有何不同？根据 Egbert *et al.*（2022：11）的定义，“代表性”指语料库能够准确概括目标语或话语场域中的典型定量语言模式的程度。我们认为，在基于语料库进行语言学研究时，研究者有必要考虑“代表性”这一数据库特征，因其决定了研究结论的可靠性。如果研究中所使用的语料库没有足够的代表性，研究将无法对话语场域中的典型定量语言模式进行准确概括，数据结果和研究结论也将不再可靠，最终，语料库语言学将难以作为一门语言学的分支学科而合法存在。鉴于此，学界需要对语料库代表性形成统一定义，语言学研究也需要一个具有实用性的操作指南来设计语料库并评估其代表性。由 Egbert、Biber 和 Gray 主编的《语料库设计与评估：针对语料库代表性的实用性框架》作为第一本以语料库代表性为主题的书籍，对语料库语言学中的代表性分析颇具开创和引领性质，旨在构建一个理论与实践并重的框架，以便在语料库语言学研究中之能够使代表性的相关概念更具可操作性。本书针对语料库的代表性，为语料库语言学研究提供了一套统一且可操作的理论框架，并演示了如何在实践中设计与评估语料库，以使其代表

\* 徐海铭为本文通信作者。

作者贡献：

张琮然：选题构思、讨论结论、初稿撰写、字数占比（70%）；

徐海铭：修改润色、字数占比（30%）。

性达到最优水准。下文将首先简介本书各章主要内容，随后对其优缺点进行简要评析。

## 2 内容简介

本书共 7 章，第一章初步介绍“代表性”的定义，后续章节更加深入地探讨相关概念，包括前人研究代表性的方法、场域和分布考虑维度及二者关系，以及在实践中实现代表性的具体方法。本书的主要观点是：如果语料库中的样本具有高度代表性，那么基于其所得出的参数估计就可以准确可靠地作为一定场域内语言特征分布的近似值。

在第一章中，作者首先给出了语料库的可操作化定义（Egbert *et al.* 2022: 7），其次初步定义了语料库代表性，最后明确了本书的 3 类目标读者群体。

第二章关注语料库语言学研究处理代表性的方法。本章介绍了先前研究人员对语料库代表性的概念化过程，并总结了现有研究基础上的两点不足，即学界对于代表性的概念存在较大分歧，且目前有关代表性的定义不具备较好的可操作性（Egbert *et al.* 2022: 41）。在此前的研究中，大多数研究人员忽视了语料库代表性，默认所使用的语料库必然具有代表性，因此在论证中缺乏对这一点的探讨，甚至有悲观者认为具有完美代表性的语料库永远不可能存在。语料库语言学中普遍存在一种观点，即代表性是“一个永远无法实现的理想”（Phillips & Egbert 2017: 1594），因此不应该去尝试实现它。然而，本书的主要目标就是反驳这些消极观点，这一点至关重要，因为其关乎此前所有语料库研究的可信度和可靠性。如果所使用的语料库代表性不足，将其应用于特定语言学的意义将大大贬值，先前的相关研究结论很可能遭受质疑，语料库语言学这门分支也将难以在学界立足。完美的代表性自然是一个无法实现的理想，但特定的操作步骤可能有助于研究者“逼近”这个理想（Mukherjee 2004: 114）。此外，语料库研究结果的可靠性因各语言特征而异，同一个高质量语料库可能对某一语言特征具有代表性，而对另一特征却不具有足够代表性。因此，本书的目标不是证明某语料库实现了代表性，而是立论证明某语料库样本在特定语言特征上具有代表性（Phillips & Egbert 2017: 1607）。总之，本书在此章的主要贡献在于，作为关注并研究语料库代表性的先锋，其重新将人们的注意力吸引到了代表性上，证明了代表性的重要性，并将针对其提供实现手段，以便为语料库语言学日后的发展夯实基础。

第三章构建了语料库代表性的概念与方法框架。本章指出语料库语言学分析的终极目标是使语言参数能够概括场域（Egbert *et al.* 2022: 56），代表性是一个持续概念（Egbert *et al.* 2022: 62）。本章为语料库代表性提供了一个较为全面的综合性理论框架。例如，开篇部分首先介绍了语料库设计的“两根支柱”：场域考虑维度和分布考虑维度。接下来，在对语言参数估计进行评估时，本书将“准确性”与“精

度”区分开来。“准确性”是语料库代表性的终极目标,意味着基于语料库分析的统计度量接近场域中语言变量的实际值。“精度”则是实现“准确性”目标的前提条件,它评估语料库库容是否足够大,以代表可操作场域中定量语言度量的值。两者的差异性还体现在:准确性不仅取决于场域考虑维度,还取决于分布考虑维度,而精度仅仅取决于样本大小,并可能受到“取样不足”(Egbert *et al.* 2022: 124)等要素影响。此外,本书还提出了其他理论术语,以描述语料库代表性的相关概念和实现途径,例如“场域边界”“场域内部范畴”等,这些术语都收录于书末的术语表中,以便读者查阅学习。总之,本书在此章的主要贡献在于:其为语料库代表性构建起了一个理论框架,清晰地解释了框架内的相关概念及术语,使得研究者能够将代表性作为一个具体的研究问题加以考察和探讨,而不是在研究中一笔带过,有助于使语料库语言学研究更加全面、系统和可信。

第四章聚焦场域考虑维度,关注定义场域边界的难度及重要性,提出了具体的操作步骤,并辅以案例分析来演示如何将场域考虑维度融入语料库建设中。在场域考虑维度下,本章首次为研究者提供了系统化的操作步骤,以分析特定场域的定性特征和非语言特征。这3个主要步骤分别是“场域描述”“场域可操作化处理”和“文本抽样”,分别与场域考虑维度相关的3个层级对应,即“目标场域”“可操作场域”和“语料库”。本书详尽展示了一例有关“学术期刊语域语料库”的案例研究,说明了如何在实际操作中实施上述3个步骤。此外,作者还指出两种可能偏差的存在,分别为目标场域和可操作场域之间的“覆盖偏差”以及可操作场域和语料库之间的“选择偏差”。最后,本书基于此认识设定了待实现的研究目标,即让这两种偏差最小化,依次考量了抽样单位、分层、层级的相对容量、随机性和非随机抽样法等方面,以实现场域考虑维度下的高代表性。总之,本书在此章的主要贡献在于:从场域考虑维度出发,为研究者提供了一套实现语料库代表性的框架,其说明既系统详尽,又具可操作性。

第五章重点阐述分布考虑维度,提出了样本容量的决定因素和分析方法,并得出了实现语言类型精确分析的方法。在分布考虑维度下,本书首创地运用统计学方法来确定创建新语料库所需的样本大小(Egbert *et al.* 2022: 130),说明为了精确地分析罕见的语言特征或变化较多的语言特征,所构建的语料库需要更大的样本容量。同时,该公式能够明确计算出研究特定语言特征所需的样本大小,以便语料库创建者对实际参数(场域均值)落在估计参数(语料库均值)的特定范围内有一定置信度。相同的定量方法也适用于确定现有语料库的精度,本书将统计学公式应用于估计“精度”的计算,即根据固定大小的现存语料库计算出定量语言特征估计的“精度”。总之,本书在此章的主要贡献在于:其从分布考虑维度出发,结合统计学知识,运用量化手段精确地计算出语料库实现代表性所需的样本量大小,并运用量化手段使精度这一抽象概念具体化,为研究者提供了具有参考价值的示范。

第六章探讨了场域和分布考虑维度对提升代表性的共同作用,并分析了二者间

的相互影响,主要通过对维基百科文章的3个场域和分布分析案例进行说明(Egbert *et al.* 2022: 126)。

第七章为研究者提供了创建新语料库的具体步骤,并通过Yelp网站餐厅评价语料库和YouTube平台视频播客语料库两个具体案例,展示了设计和评估语料库代表性的实用方法。

### 3 简评

概括而言,本书的第一个优点在于其应用了目标导向的分类方法。在练习和讨论部分,作者用特殊符号标记了目标受众,包括“语料库建设者”“语料库分析者”和“语料库研究使用者”,使得本书对目标受众更具针对性,也与语料库研究实践更密切相关。此外,通过向目标受众提出一系列可供讨论的问题,如“该语料库旨在代表哪个目标场域?”,本书引导目标读者自主得出结论,以解释语料库代表性的相关概念和操作步骤,有益于培养目标受众的批判性和逻辑思维能力。

本书的另一个优点在于其教学价值,尤其体现在其对类比法的运用。本书将语料库语言研究中构建语料库的过程类比为实证科学研究中的抽样调查,将“语料库”比作“样本”,将“语言使用的场域”比作“总体”。代表性即是“样本”可代表“总体”的程度,因此其重要性不容小觑。Biber(1993: 243)指出,语料库设计中的代表性是指“样本包括总体中所有变体的程度”。本书还通过对比法探讨了语料库语言学和其他实证科学研究之间的区别:语料库研究人员通常会为多个研究问题重复使用同一个“样本”,即同一个语料库,因此语料库代表性对于各个具体研究的适用性至关重要,研究者更应该严肃对待语料库代表性这一问题。此外,纵观全书,作者始终站在读者视角,将“代表性”视作一个抽象概念,并用简明而富有启发性的语言对其进行了详尽阐释,具有一定的教学价值,如在最后一章,作者通过实例引导读者评估某一现存语料库对于特定研究问题的适用性,有助于读者在使用语料库进行研究之前,考量其对所研究语言对象的代表性,从而有益于读者未来在语料库语言学领域的学术研究与写作。

该书存在如下局限。首先,该书将场域考虑维度和语言分布考虑维度作为两个截然不同的方面加以分析可能会引起混淆。这两个维度相互交织,以实现准确估计语言参数作为终极目标,但场域考虑维度不能提供准确的定量语言分布估计,而分布考虑维度也无法说明语料库所包含的文本种类。因此,两者对于语料库代表性的全面权衡都至关重要。然而,针对这两个相互关联的考虑视角,该书尚未明确界定彼此之间的界限,这一点有待后续阐明。其次,作者在谈及样本量时提到,研究者需要一个“循环过程”来确定样本量大小,即研究者既需要标准差来估算语料库所需的库容,也需要一个语料库来度量标准差(Egbert *et al.* 2022: 129),如果该书能



够更详细地解释其原理和具体操作，可能会更便于读者理解。最后，该书在分布考虑维度方面的论述也可以更加简洁，比如，对于一些基本的统计学术语，如“置信区间”“标准差”等，不必进行详细的定义和解释。

整体来看，本书详略组织得当，时效性强，作为入门教学参考书，具有极其有效的作用。最为重要的是，本书构建了一个相对全面的语料库代表性框架，使得研究者重新关注这一先前常被语料库研究者忽视的重要特征，开创了语料库代表性概念化和操作化的先河。

### 参考文献

- BIBER D. Representativeness in corpus design[J]. *Literary and Linguistic Computing*, 1993, 8(4): 243-257.
- EGBERT J, BIBER D, GRAY B. Designing and evaluating language corpora: a practical framework for corpus representativeness[M]. Cambridge: Cambridge University Press, 2022.
- MUKHERJEE J. The state of the art in corpus linguistics: three book-length perspectives[J]. *English Language and Linguistics*, 2004, 8(1): 103-119.
- PHILLIPS J, EGBERT J. Advancing law and corpus linguistics: importing principles and practices from survey and content analysis methodologies to improve corpus design and analysis[J]. *Brigham Young University Law Review*, 2017(6): 1589-1619.
- WONG W, LIU W, BENNAMOUN M. Constructing specialized corpora through analyzing domain representativeness of websites[J]. *Lang Resources & Evaluation*, 2011, 45(2): 209-241.

通信地址：201620 上海市 上海外国语大学英语学院



# English abstracts of major articles

---

## The analysis of the confrontational features and discursive strategies in Chinese and American media’s COVID-19 virus origin tracing discourse

.....*WEI Naixing & JIN Bixi* (1)

During the COVID-19 pandemic, tracing the origin of COVID-19 became a critical discursive event in international media. Based on a self-compiled corpus, this study explores the discursive representation of this event in US and Chinese media. News reports were collected from January 2020 to December 2021. Basic corpus linguistic analytical techniques were utilized including concordance analysis, collocation analysis, semantic preference and semantic prosody analysis, aiming to explore the characteristics of discourse conflict, game-playing, and contention between the two sides. Data analyses showed that the Chinese and US media employed a series of “attack-defense” discursive strategies embodied in three diametrically opposing sets of semantic preferences and prosodies, revealing a strong profile of confrontational discourses bounded up with divergences in national stances and ideologies. This article discusses the important linguistic characteristics and discursive strategies of confrontational discourse, offering insights into the conceptual construction and methodological aspects of discourse analysis.

## An analysis of conflict discourse in the Iraq War from the perspective of speech act theory

.....*LI Xinyue, MA Xiaolei & LI Dekun* (15)

This research compiles and analyzes the variety of discourses from different groups, as reported both domestically and internationally during the Iraq War. It applies Searle’s Speech Act Theory to categorize these complex discourses. The study further explores the conflicting discourses within the Iraq War, examining the intentions behind the United States’ use of specific types of speech acts. Findings indicate that the discourse of conflict in the war centered on specific areas of contention; the US military’s strategic employment of various speech acts aimed to manage the war’s progression, thereby enforcing its hegemonic and power-driven policies.

## Research on the generation and application of sense lists assisted by large language models

..... *GE Xiaoshuai & WANG Meng* (26)

This article introduces the concept of a sense list derived from dictionaries, which expands the vocabulary of a corpus at the semantic level. It explores the presentation format of sense lists and verifies the feasibility of utilizing ChatGPT to assist in their generation. The findings suggest that the ChatGPT model is highly accurate in differentiating senses and possesses strong generalization capabilities, making it an ideal tool for creating dictionary-based sense lists. The article concludes by discussing the practical value of sense lists and illustrating their utility through a specific example.

## A review of the empirical alternation studies

..... *WANG Menglan & JIANG Guiying* (38)

This paper examines the emerging research trajectory of “Linguistics+” through the lens of empirical alternation studies, which integrate linguistic research with advanced statistical techniques to enhance cognitive plausibility and quantitative analysis. It begins with an overview of the theoretical frameworks, primary themes, and methodologies employed in alternation research. Significant progress has been identified in several key areas: (1) comparative analysis of constraining factors, (2) investigations into cognitive plausibility, and (3) the use and development of sophisticated statistical methods. Moving forward, the paper advocates for the expansion of research into native languages, the inclusion of a broader set of influencing factors, the adoption of online experimental paradigms, and a reinforced emphasis on the use of the R language for statistical analysis.

## Multidimensional scaling and cluster analysis of Shakespeare’s sonnets and their Chinese translations using word frequency indices

..... *CHEN Guoxing & XU Dekuan* (49)

This study conducts a one-way ANOVA on eight word frequency distribution indices to explore the similarities and differences between Shakespeare’s sonnets and five Chinese translations. Using multidimensional scaling and cluster analysis, the findings indicate that while no Chinese version is a perfect match to the original, three versions are notably closer. The data-driven analysis confirms that, although translating poetry poses significant challenges, certain translations can quantitatively approximate the original. This is corroborated by qualitative evaluations of some versions.

## The construction and dissemination strategies of the Yellow River Basin's media image from the perspective of appraisal theory

..... *WANG Haifeng & LI Wenyu* (68)

An integral part of national image construction—the construction and dissemination effectiveness of the Yellow River Basin's external discourse—has become an important research topic in academia. This study utilizes English reports about the Yellow River Basin from the official website of *China Daily*, applying appraisal theory as its analytical framework. Starting from a positive discourse analysis perspective and employing a corpus linguistics quantitative analysis paradigm, the study explores the discursive meanings of attitude resources in the dissemination of news discourse about the Yellow River Basin. The findings indicate that: (1) the current key themes in media discourses about the Yellow River Basin have shifted from “drought and flood disasters”, “floods”, and “environmental pollution” to “environmental governance”, “ecological protection”, and “high-quality development”; (2) positive emotions, judgments, and appreciations construct a positive, objective, and active linguistic environment in news discourses, highlighting a constructive approach to the positive external discourse space of the Yellow River Basin, and presenting a high-quality, ecologically sound media image; (3) in response to the current status of the Yellow River Basin's media image, this study proposes a combined strategy of self-shaping and other-shaping. This strategy integrates overt and covert communication, and aligns segmented and multimodal dissemination methods, aiming to provide references for the external discourse communication of the Yellow River Basin.

## A study of the construction of China's image through prosodic valence analysis

..... *TANG Huitao & WANG Jun* (87)

Employing the extended unit of meaning model and semantic analysis technology, this study adopts a quantitative approach to explore the depiction of China in the media. It calculates the prosodic valence from 500 instances where “China” is the focal word in the Coronavirus Corpus. The findings reveal that although foreign media generally portray China in a positive light, negative evaluations are significantly more prevalent.

## A corpus-based study on nomination strategy of “a community with a shared future for mankind” in Indian media coverage

..... *LI Mengxiang* (103)

This paper employs the nomination strategy from Ruth Wodak's discourse-historical approach to analyze the notion of "a community with a shared future for mankind" as covered in Indian media. A corpus of reports related to this initiative, spanning from 2013 to 2022, was compiled. The analysis focused on concordance lines, high-frequency co-occurring nouns, and the semantic prosody of key search terms. Based on socio-historical contexts, the findings suggest that the portrayal of "a community with a shared future for mankind" in Indian media has evolved through stages: initial neglect and disagreement, the onset of attention and balanced reporting, heightened attention with mixed responses, and a subsequent decline in attention with predominantly negative reporting. This study aims to understand the attitudes towards and dissemination of this concept in Indian media, providing insights for shaping international dissemination strategies of the initiative and demonstrating the role of language research in national strategy.

# 语料库语言学

CORPUS LINGUISTICS

## 要 目

- |                           |             |
|---------------------------|-------------|
| 中美媒体新冠病毒溯源话语的对抗性特征与话语策略分析 | 卫乃兴 金碧希     |
| 言语行为理论视角下伊拉克战争冲突性话语分析     | 李新月 马晓雷 李德坤 |
| 大语言模型辅助下基于词典的义项表生成及应用研究   | 葛晓帅 王 梦     |
| 交替现象实证研究现状与展望             | 王梦岚 江桂英     |
| 基于语义韵强度算法的中国形象建构研究        | 汤慧桃 王 军     |
| frGLOBE 当代法语书面语平衡语料库的创建   | 殷伶俐         |
| 人教社中小学教材语料库 (PEPTC) 的研制   | 黄攀伟 等       |

外研社·期刊出版分社

电话: 010-88819267

E-mail: qkzx@fltrp.com

网址: www.bfsujournals.com



记载人类文明  
沟通世界文化  
www.fltrp.com



北外学术期刊



iResearch 微信公众号

责任编辑: 赵 雪  
责任校对: 蔡 喆  
封面设计: 锋尚设计

ISBN 978-7-5213-5433-1



9 787521 354331 >

定价: 60.00元