

《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20辑  
二〇二三

# 语料库语言学

## CORPUS LINGUISTICS

10年  
第20辑  
2023

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 local grammar word embeddings  
CLEC collocation multifactorial analysis  
AntConc DEAP  
big data corpus  
Brown Crown MDA semantic prosody  
BNC corpus-as-method  
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：王 斌

**Journal Name Calligraphy:** Cui Xiliang  
**Editor:** Xu Jiajin  
**Proofreaders:** Wang Bin

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

2023 年 第 20 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》

2023年 第10卷 第2期

## 目 录

### 语境共选

语域实证研究：基于词汇语义的多维分析法.....	钱玉彬 孙 亚 ( 1 )
视觉动词“看”的多义性：基于语料库的行为特征分析.....	黄静雯 李金妹 胡志勇 ( 15 )
不同水平中国英语学习者虚化动词搭配使用及迁移效应研究.....	闫盛德 高 霞 ( 27 )

### 研究论文

中国抗疫推文特征对海外受众行为参与度的影响研究.....	江进林 王佳慧 ( 43 )
多模态话语分析视阈下的新闻价值研究.....	韩存新 赵宇飞 ( 58 )
基于语料库的中美媒体关于TikTok新闻报道的批评话语分析.....	黄 馨 罗卫华 ( 75 )
基于CIA模型的列举类词习得研究.....	李艳娇 李 齐 庄会彬 ( 89 )
意大利语语料库及其应用研究.....	谭钰薇 余丹妮 ( 107 )

### 研制开发

arGLOBE当代阿拉伯语书面语平衡语料库的创建.....	毛浚语 ( 122 )
faGLOBE当代波斯语书面语平衡语料库的创建.....	李彦军 陈帅楠 胡 奇 周汀鹭 ( 132 )
itGLOBE当代意大利语书面语平衡语料库的创建.....	喻儒辰 董 丹 郭垚一 ( 141 )
MineDEAP矿业工程学术英语语料库的创建.....	张汝莹 ( 150 )
SET多版本高中英语教材语篇语料库的创建.....	陈运良 ( 155 )

### 书刊评介

《如何利用语料库进行语言教学》述评.....	张春青 ( 167 )
《学术界的跨学科实践：写作、教学与评估》述评.....	陈雅刚 ( 174 )
主要文章英文摘要.....	( 180 )



# Corpus Linguistics

Volume 10, Number 2, 2023

## Table of Contents

### Featured column: Contextual co-selection approach to language

- Lexical semantics-based multidimensional analysis of register  
.....*QIAN Yubin & SUN Ya* (1)
- The polysemy of the perception verb “看 (kàn)” : a corpus-based behavioral profile  
.....*HUANG Jingwen, LI Jinmei & HU Zhiyong* (15)
- Usage and L1 transfer effects of delexicalized verb collocations among  
Chinese EFL learners across proficiency levels.....*YAN Shengde & GAO Xia* (27)

### Research articles

- Impact of Chinese anti-COVID-19 Tweets on overseas audience engagement  
.....*JIANG Jinlin & WANG Jiahui* (43)
- Research on news values from the perspective of multimodal discourse analysis  
.....*HAN Cunxin & ZHAO Yufei* (58)
- Comparative corpus-based critical discourse analysis of Chinese and American news  
coverage on TikTok.....*HUANG Xin & LUO Weihua* (75)
- Acquisition analysis of enumerative conjunctions in the CIA model  
.....*LI Yanjiao, LI Qi & ZHUANG Huibin* (89)
- Italian corpora and their applications .....*TAN Yuwei & YU Danni* (107)

### New corpora, tools and methods

- The construction of arGLOBE: A balanced corpus of contemporary written Arabic  
.....*MAO Junyu* (122)
- The construction of faGLOBE: A balanced corpus of contemporary written Persian  
.....*LI Yanjun, ZHOU Tinglu, HU Qi & CHEN Shuainan* (132)
- The construction of itGLOBE: A balanced corpus of contemporary written Italian  
.....*YU Ruchen, DONG Dan & GUO Yaoyi* (141)
- The construction of MineDEAP: A corpus of academic English in mining engineering  
.....*ZHANG Ruying* (150)
- The construction of SET: A corpus of discourses from multi-version high school English  
textbooks .....*CHEN Yunliang* (155)

### Book reviews

- A. Coxhead. *Connecting Corpora and Language Teaching* .....*ZHANG Chunqing* (167)
- L. Buckingham, J. Dong & F. Jiang (eds.). *Interdisciplinary Practices in Academia:  
Writing, Teaching and Assessment* .....*CHEN Yagang* (174)

- English abstracts of major articles..... (180)

# 语域实证研究： 基于词汇语义的多维分析法<sup>\*</sup>

中国科学院大学 钱玉彬 对外经济贸易大学 孙 亚

**提要：**本研究将104个语义特征和67个语法特征纳入考察范围，分析北京奥运对外新闻的语域特征。研究发现8个功能维度，即情感交互性与信息呈现性、地点景观描述性与抽象事物叙事性、权利立场性、主观评价性与客观事实性、人文艺术关切、健康安全关切、设施工具与群体活动关切、技术投入关切。研究表明，基于词汇语义的多维分析法能够有效揭示话语的独特交际功能，有助于弥补传统语料库方法重语法轻语义的不足，对理解语义介入和拓展语域分析技术具有一定的启示。

**关键词：**语域、语料库、多维分析、词汇语法、词汇语义

## 1 引言

依据Conrad (2015: 309)，语域是一种“基于情景的话语类型”，体现特定语境下语言的使用问题及其功能意义。语域研究的基本理念是通过显著的共现语言特征来区分语域类型及功能 (Hymes 1974; Halliday 1988)，而注重频数和概率的语料库技术在识别这些语言特征方面具有天然的优势，其中由Biber (1984)创建的多特征/维度分析框架 (multi-feature/dimensional framework，简称“多维分析法”) 已成为目前语料库学界语域研究的主流研究方法 (McEnery *et al.* 2006; 江进林、许家金 2015)，被广泛应用于口语、书面语、学习者话语、课堂话语、学术话语、商务话语等不同语域和语言变体的研究。尽管如此，以往的多维分析法高度依赖词汇语法而弱化语义，导致语域分析不够全面。据此，本研究尝试将词汇语义纳入多维分析法中，旨在揭示语义如何参与语域构建、如何与语法组合共同实现特定交际功能。

<sup>\*</sup> 本文系中国外语教育基金项目“基于词汇语义的奥运外宣话语语域研究”(ZGWYJYJJ11A127)阶段性研究成果。钱玉彬为本文通讯作者。

作者贡献：

钱玉彬：研究方法、数据收集、讨论结论、初稿撰写、字数占比(90%)、修改润色。

孙亚：选题构思、字数占比(10%)。

## 2 多维分析法

多维分析法在很大程度上受到社会语言学变异研究和多元统计分析的影响（许家金 2019），通过频数统计和因子分析技术识别出高频共现的语言特征作为语域分析的核心（Conrad & Biber 2013）。这些共现特征聚合构成多个因子，即多维分析法的“功能维度”，用于解释和区分语域，而每个维度又通过共现特征这一媒介表征“情景、社会 and 认知功能”（Biber & Finegan 1989: 488）。在早期研究（Biber 1984, 1985, 1986）基础上，Biber（1988）将语言特征拓展至67个，涉及16种主要语法范畴，详细说明了如何采用多维分析法将它们聚类为7个功能维度，即“交互性与信息性表达”“叙述性与非叙述性关切”“指称明晰性与情境依赖型指称”“显性劝说型表述”“信息抽象与具体程度”“即席信息组织精细度”和“学术性模糊表达”，并据此对比口笔语及其多个语类，在推介多维分析法方面具有里程碑意义。自此，大批学者要么依照以上功能维度展开分析（如Biber *et al.* 2004; Biber *et al.* 2006; Biber & Conrad 2009; Gardner *et al.* 2015; Nini 2017; Kruger *et al.* 2019），要么仿照多维分析法的实施步骤探索其他维度（如Grieve *et al.* 2010; Weigle & Friginal 2015; Yan & Staples 2019）。多维分析法的研究内容也从最初的口笔语拓展至语域间的共时对比（Conrad & Biber 2013; Crosthwaite 2016; Sardinha & Pinto 2017; Kruger & Smith 2018; 武姜生 2004; 张一宁等 2018; 李端阳、王志军 2019; 胡春雨、谭金琳 2020; 王伟 2021）、语域的历时演变（Atkinson 1992; Biber & Finegan 1992; Friginal & Weigle 2014; Kytö & Smitterberg 2015; Crosthwaite 2016）、作者/译者风格（Biber & Finegan 1994; Watson 1994; Biber *et al.* 1998; Nini & Grant 2013; 赵朝永 2020）、二语写作（Reppen 1994; Reynolds 2005; 潘璠 2012; 赵朝永、王文斌 2017）、技术工具（McEnery *et al.* 2006; Grieve-Smith 2007; Conrad 2014; Nini 2015）等方面。

多维分析法随着研究数量的增多已趋于成熟，但是存在的共性问题为过度依赖词汇语法而不够重视其语义，造成语域分析因重语法、轻语义而不够完整。为克服这一问题，Xiao（2009）提出的融入语义的多维分析尤其值得关注。Xiao（2009）利用语料库在线分析工具Wmatrix提供的语法和语义赋码功能，对英语国际语料库的5种英语变体、12个语域展开多维分析，共标记141个语言特征，其中包含97个语法特征和44个语义特征；经数据筛选后保留109个特征用于因子分析，解析出9个功能维度：（1）互动休闲式与信息详尽式话语；（2）详尽的即时评估；（3）展示性关注；（4）人物描述与物体描述；（5）未来预测；（6）主观印象和判断；（7）时间或地点焦点缺失；（8）程度和数量；和（9）报道性言论。研究指出，以上维度能够有效区分英语变体。如在维度（1）上，英式英语的互动性最强，印度英语最为详尽，而其他变体，如中国香港、新加坡和菲律宾英语则没有明显差异；在维度（2）上，英式英语的评价性最强，是区分英语母语及其变体的

主要维度。国内目前仅有孙亚、崔子璇（2020）开展过基于语义的类似研究。他们将Wmatrix的语义赋码与Biber（1988）的语法特征相结合，对特定检索词如business、corporation、company的隐喻载体词作语义和语法赋码，然后采用因子分析技术获取基于〔人物运动〕隐喻的信息性表达、基于〔物体增大〕隐喻的非叙述性表达和基于〔物体方位〕隐喻的信息性表达等维度，探讨基于隐喻使用的语域功能维度。本研究在前人基础上，尝试使用新工具设计自由度更高、可重复性更强的多维分析法，这对理解语域中的语义介入和拓展相关技术应用都将具有一定的启示。

### 3 研究设计

#### 3.1 研究语料

研究所用语料取自2001—2022年发表于《中国日报》的北京奥运新闻。经筛选和删除重复项，获取2,577个文本，共975,092形符数。《中国日报》是我国对外宣传的权威新闻媒介，在推进中国媒体走进世界信息体系中扮演中心角色（吴瑛等2015），因此其语料极具代表性和重要性。

#### 3.2 研究步骤和工具

多维分析一般遵循以下3个技术步骤：提取语言特征、对语言特征作描述统计、因子分析（Biber 1988）。首先，本研究的语言特征包括词汇语法特征和词汇语义特征，分别采用Stanford Tagger（ST）和UCREL Semantic Analysis System（USAS）工具实现自动识别和标注。ST是斯坦福大学自然语言处理研究组研发的词性赋码工具，可复现Biber（1988）中的67个词汇语法特征。与其他同类型工具如TreeTagger、TnT和SVMTool相比，ST采用依存网络（dependency network）和最大熵模型，具有双向推理、特征描述全、准确度高等优势（Toutanova *et al.* 2003）。由于模型经过《华尔街日报》文本训练，因此ST也非常适用于标注新闻话语。USAS由兰卡斯特大学语料库研究中心研发，基于《朗文当代英语词典》，可区分232个词汇语义特征（Rayson *et al.* 2004）。这些语义特征存在层级式嵌套关系，如一级语义“总体和抽象”包含“情感”和“分类”等二级语义，“情感”又包含“调整、改变”等三级语义。在选取语言特征时，应避免重复并满足因子分析的数据要求，即变量数目不宜过多，同时得到解释力强的主要因子。据此，我们将语义特征按照层级关系合并，并抽取104个特征与语法特征组合，共同构成171个语言特征。接着，使用统计工具计算各语言特征频数及其标准化数据。最后，参考Biber（1988）的多维分析步骤，使用SPSS工具对上述标准化数据作因子分析。具体而言，需将语言特征视为变量，使用方差最大化正交旋转降维数

据并获取因子。接着依据碎石图和载荷系数分别裁剪因子及其变量（通常而言，碎石图中的自然断点对应须保留的因子数）：排除载荷系数低于 $|0.30|$ 的变量，若该变量同时出现在多个因子中，则按照其载荷系数的最高值确定归属因子（Biber 1988：87-88）；最后，从因子内不同变量共同实现的交际目的出发定义因子。

## 4 结果与讨论

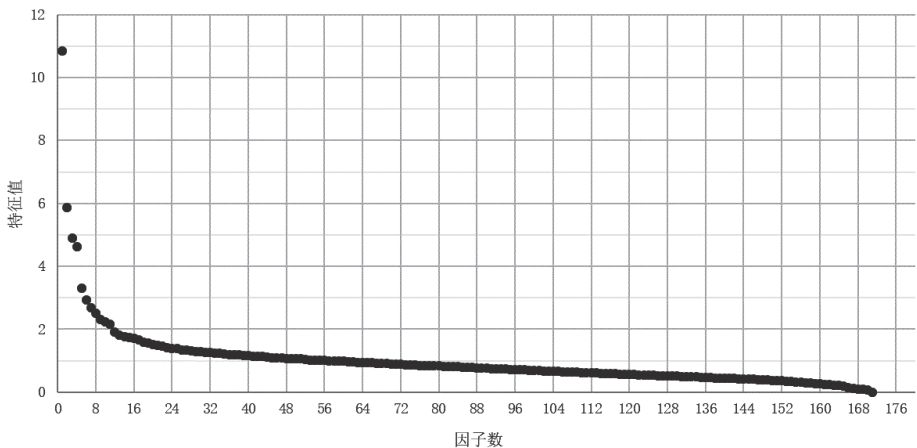


图1 北京奥运外宣话语因子分析碎石图

如图1所示，碎石图中存在多个自然断点，对应的因子数在1—12个之间。若因子提取数量不足则会排除大量语言特征而丢失信息，若提取过多则会造成语言特征的“混乱图景”（Biber 1988：88），使定性分析变得困难。据此，我们按照Costello & Osborne（2005：3）的建议，在手动设置6—12个因子数并依次比较变量数量 and 解读难度之后，提取出8个因子作进一步分析。表1—表8呈现8个因子及其变量；在分析共现语言特征交际功能的基础上，将各因子界定为8个功能维度。

将因子1定义为“情感交互性与信息呈现性”维度，共包含26个语言特征。如表1所示，载荷值为正数的语言特征（以下简称“正向特征”）共23个，大致分为3种类型：一是情感态度类，如存在含义、表语形容词、可能性含义、可能类情态动词、语气强调词、排他含义等；二是人际互动类，如第一人称代词、条件状语从句从属连词、第二人称代词等；三是信息模糊类，如*be*做主动词、分析否定词、缩略词等。正向特征共同形成较明显的情感交互而弱化细节信息，如例（1）的*if*、*you*、*interested*、*want*、*drive*、*please*等词明确揭示北京奥运新闻的互动特色。对比而言，载荷值为负数的语言特征（以下简称“负向特征”）共3个：

表1 因子1：情感交互性与信息呈现性

<i>be</i> as main verb <i>be</i> 做主动词	0.717	Existential <i>there</i> 存在句 <i>there</i>	0.424
Being 存在含义	0.716	Exclusivizers/particularizers 排他含义	0.424
Present tense 现在时	0.659	Pronoun <i>it</i> 代词 <i>it</i>	0.421
Analytic negation 分析否定词	0.653	Second-person pronouns 第二人称代词	0.412
Adverbs 副词	0.635	Private verbs 私用认知动词	0.407
Contractions 缩略词	0.626	Causative adverbial subordinators 原因 状语从句从属连词	0.352
First-person pronouns 第一人称代词	0.602	Mental actions and processes 心理活动 和过程含义	0.340
Predicative adjectives 表语形容词	0.584	Pro-verb <i>do</i> 助动词 <i>do</i>	0.328
Probability 可能性含义	0.577	Demonstratives 指示词	0.325
Degree 程度含义	0.550	Nouns 名词	-0.611
Demonstrative pronouns 指示代词	0.476	Prepositional phrases 介词短语	-0.427
Possibility modals 可能类情态动词	0.460	Phrasal co-ordination 并列短语连词	-0.377
Emphatics 语气强调词	0.440		
Conditional adverbial subordinators 条件状语从句从属连词	0.434		

名词、介词短语和并列短语连词，均是信息高度密集的标志，如例（2）的 on、in、and、Shougang、Group、partner 等词披露首钢参与冬奥建设的相关信息。

（1）*If you are interested* in urban renewal, *if you want* to know how the Olympic Games *drive* the urban development... then *please* look around.

（2）*On* June 5, 2018, *Shougang Group* officially became a cooperative *partner* of the official urban renewal *service* of the Beijing *Winter Olympic Games and Winter Paralympic Games in* 2022.

因子2界定为“地点景观描述性与抽象事物叙事性”维度，共包含11个语言特征。如表2所示，正向特征共6个，将地点和房屋、建筑物等语义特征，与名词化、推断类情态动词和未带施动者被动句等交互性弱而信息性强的语法特征组合在一起，用于描述地理景观，如例（3）的 street、community、supermarket 和 park 等词描述因筹办奥运而改善的景观环境。负向特征共5个，由指称宽泛的特征，如能力含义、宇宙含义和第三人称代词，和指示时间的特征，如过去时和线性顺序含义共同组成，展现北京奥运人或物故事发展的历史进程，具有较鲜明的叙事



性，如例（4）的 *series*、*shed* 等词提供北京夏奥会成功举办及赛后社会效应的历史线索。

表2 因子2：地点景观描述性与抽象事物叙事性

Places 地点含义	0.496
Nominalizations 名词化	0.483
Predictive modals 推断类情态动词	0.459
Agentless passives 未带施动者被动句	0.423
Architecture and kinds of houses and buildings 房屋和建筑物含义	0.368
Gerunds 动名词	0.359
Ability 能力含义	-0.519
Past tense 过去时	-0.501
The universe 宇宙含义	-0.471
Linear order 线性顺序含义	-0.443
Third-person pronouns (excl. <i>it</i> ) 第三人称代词（除 <i>it</i> ）	-0.407

（3）The *street* lights in the *community* are brighter, the *supermarket* is closer, and the *park beside* the Yongding *River* has become a good place for a walk.

（4）With a *series* of celebrations for the 10th anniversary taking place across China, the thoughts of leading figures in the sports community *shed* light on the huge effect the event has had on the country.

表3 因子3：权利立场性

Speech acts 言语行为	0.702
Public verbs 公用认知动词	0.657
Subordinator <i>that</i> deletion 主从连词 <i>that</i> 省略	0.539
<i>that</i> verb complements 动词补语 <i>that</i>	0.467
Social actions, states and processes 社会行动、国家和过程含义	0.383
Relationship 关系含义	0.367
Government, Politics and elections 政府政治和选举含义	0.359
[no negation features] 无负向特征	□

因子3是“权利立场性”维度，共7个正向特征，无负向特征。如表3所示，这些特征包含社会行动、国家和过程、关系、政府政治和选举等多个涉及权利的主体和客体、获取和实施等语义；语法特征则包括公用认知动词、主从连词*that*和动词补语*that*等，它们经常出现在公开演讲类话语中，实现阐明价值观、表达意见或事理的交际功能。

表4 因子4：主观评价性与客观事实性

Attributive adjectives 定语形容词	0.614
Affect 影响含义	0.426
Type-token ratio 类形符比	0.388
Evaluation 评估含义	0.375
Comparing 比较含义	0.371
Helping/hindering 帮助/阻碍含义	0.364
Mental object 认知心理含义	0.331
Time 时间	-0.497
Numbers 数字	-0.401

因子4是“主观评价性与客观事实性”维度，包含9个语言特征。如表4所示，正向特征共7个，包括指示词汇丰富度和内容清晰度的类形符比，和表示个人情感、评价和态度的多种语言特征，如定语形容词、影响含义、评估含义、比较含义等。以上正向特征共同展现奥运新闻的情感态度基调，如例（5）的*definitely*、*another*和*boost*等词呈现鲜明的主观色彩。负向特征共2个：时间和数字，均用于描述细节以增强客观事实性，如例（6）的*one*、*180,000*、*2013*、*25 percent*等词反映场馆人数、时间、增长率等细节信息。

（5）The Olympic bid will *definitely* provide *another boost* for the sport, attracting more people and more funds.

（6）*One* of the club’s commercial rinks, in the suburbs of north Beijing, was visited by almost *180,000* skaters in *2013*, a *25 percent* increase from the previous year.

因子5是“人文艺术关切”维度，仅包含7个正向语义特征。如表5所示，这些特征主要描写精神文化相关的行为活动或精神风貌，如美术和工艺、戏剧、演出、宗教、音乐、快乐/悲伤情绪等。如例（7）的*concerts*、*shows*、*staged*、*set*



和 example 等词围绕多种文化现象，展现北京奥运人文艺术的生命力及其激发的赛后效应。鉴于奥运在经济、政治、哲学、文化等诸多方面的影响力，北京奥运新闻凸显人文艺术维度，彰显我国主张的人文奥运理念。

表5 因子5：人文艺术关切

Arts and crafts 美术和工艺含义	0.491
Drama, theatre and show business 戏剧、演出和娱乐业含义	0.461
Seem 似乎含义	0.436
Open/closed; Hiding/Hidden; Finding; Showing 打开/关闭；躲藏/隐藏；寻找；显示含义	0.381
Religion and the supernatural 宗教与超自然含义	0.369
Music and related activities 音乐及相关活动含义	0.359
Happy/sad 快乐/悲伤含义	0.301
[no negation features] 无负向特征	□

( 7 ) With the transformation work for 2022 and commercial events such as *concerts* and entertainment *shows* being *staged*, Beijing has *set* an *example* for Olympic host cities in the post-Games operation of permanent facilities.

表6 因子6：健康安全关切

Crime, law and order 犯罪、法律和秩序含义	0.456
Calm/Violent/Angry 冷静/暴力/愤怒含义	0.451
Life and living things 生命和生物含义	0.375
Health and disease 健康与疾病含义	0.336
Fear/bravery/shock 恐惧/勇敢/震惊含义	0.317
Warfare, defense and the army; weapons 战争、国防和军队；武器含义	0.311
[no negation features] 无负向特征	□

因子6界定为“健康安全关切”维度，仅包含6个正向语义特征。如表6所示，它们主要涵盖生命生物、健康疾病等与百姓健康福祉息息相关的语义，或法制法规、冷静/暴力/恐惧等情绪、战争国防等事关生命财产安全活动的语义。如例（8）的 COVID-19、pandemic、virus 和 protocols 等词反映北京充分考虑区域性人员流动带来的潜在疫情防控风险，适时采取措施保障赛事期间人员的生命健康

安全。因子6与因子5互为补充，共同体现以人为本的人文奥运理念。

(8) Despite the unprecedented challenges posed by the **COVID-19** pandemic... we have to adjust the event plans reasonably to get the action going while sticking to the **virus-prevention protocols** in our country.

表7 因子7：设施工具与群体活动关切

Measurement 测量含义	0.388
Vehicles and transport on land 陆路交通和运输含义	0.377
Physical attributes 物理属性含义	0.336
Substances and materials generally 物质和材料含义	0.311
People 人民含义	-0.407
Education in general 教育含义	-0.402
Groups and affiliation 团体和隶属含义	-0.394
Sports and games generally 运动和竞技含义	-0.379

因子7是“设施工具与群体活动关切”维度，共8个语言特征，且均为语义特征。如表7所示，正向特征主要表示测量、陆路交通和运输、物理属性、物质和材料等含义，涉及设施标准、场地设备、建筑材料、交通运输等从工程建设到资源供给等多个方面，贯穿奥运赛事前后全过程的可持续发展观，体现独特的绿色办奥理念，如例（9）的 **venue**、**water**、**rainwater** 和 **reservoirs** 等词表达水资源的重复利用。负向特征由人民、教育、团体和隶属、运动和竞技等语义组成，表现广大公众支持和参与奥运、共同谱写奥运篇章的浓厚热情，如例（10）的 **singer**、**actor**、**skiing** 和 **games** 等词展现社会各界为北京奥运助力添彩。

(9) The **venue**'s design features an efficient **water** recycling system, which will collect **rainwater** and store melted snow at two high-altitude **reservoirs**.

(10) Pop **singer** David Zee Tao ... **actor** Lu Han record their copy of Welcome to the Great Wall for **Skiing**, one of the theme songs of Beijing's bidding for the 2022 Olympic Winter **Games**.

因子8是“技术投入关切”维度，共5个正向语义特征，体现商务、金钱、信息技术和计算、数量、传播等含义。这些语义与过去二十年，特别是2008年夏奥

会至2022年冬奥会期间，北京秉持的科技办奥理念是分不开的，即抓住历史契机支持高新技术产业，大力发展身份认证、智能交通、5G通信等先进技术，保障科技成果的市场转化和开发。如例（11）的China Unicom、inked、Huawei、company和internet等词描述中国联通等企事业单位为提升赛事管理的高效性和安全性而提供的5G通信技术保障。

表8 因子8：技术投入关切

Business 商务含义	0.553
Money generally 金钱含义	0.546
Information technology and computing 信息技术和计算含义	0.449
Quantities 数量含义	0.360
Communication 传播含义	0.308
[no negation features] 无负特征	□

（11）*China Unicom* announced that it has *inked* strategic deals with *Huawei Technologies Co, Panasonic, Beijing Sport University* and 13 ski *resorts* to form an alliance to promote winter sports in China... The *company* will support the Olympic Games with a secure, smart network that allows superfast *internet* speeds.

5 结论

研究表明，基于词汇语义的多维分析法能够有效揭示语域特点。本研究以北京奥运对外新闻报道为例，对104个词汇语义和67个词汇语法展开多维分析，成功获取该语域的8个功能维度：（1）情感交互性与信息呈现性；（2）地点景观描述性与抽象事物叙事性；（3）权利立场性；（4）主观评价性与客观事实性；（5）人文艺术关切；（6）健康安全关切；（7）设施工具与群体活动关切；（8）技术投入关切。

与以往的多维分析相比，本研究的语言特征更为完整，实现了将Biber（1988）的词汇语法拓展至语义层面，一定程度上克服了重语法而轻语义的不足。另一方面，采用基于词汇语义的多维分析法，能够揭示分类更加精细的语域功能维度，也能够揭示其特有的交际功能。例如，人文艺术关切和健康安全关切维度体现以人为本的办奥理念，设施工具与群体活动关切维度则展现绿色办奥理念，技术投入关切维度则传递科技办奥理念。可以看出，这些依赖词汇语义的功能维度很难

用语法特征作出解读，它们将北京的人文奥运、绿色奥运和科技奥运等核心理念，以及“同一个世界、同一个梦想”“一起走向未来”的友邦精神予以充分展现，彰显出奥运对外新闻定位北京、传播中国声音的特色。

最后，本研究虽然同时考察了语法和语义特征，但是没有根据语料特色对词汇作出筛选，也没有平衡两类特征的数量，因此得出的结果存在局限性。此外，词汇语义包含多层级的嵌套关系，没有分别考察各层级语义，因此未来仍需进一步区分不同层级语义对多维分析结果的潜在影响。

### 参考文献

- ATKINSON D. The evolution of medical research writing from 1735 to 1985: the case of the Edinburgh Medical Journal [J]. *Applied Linguistics*, 1992, 13(4): 337-374.
- BIBER D. A model of textual relations within the written and spoken modes [D]. Los Angeles: University of Southern California, 1984.
- BIBER D. Investigating macroscopic textual variation through multifeature/multidimensional analyses [J]. *Linguistics*, 1985, 23(2): 337-360.
- BIBER D. Spoken and written textual dimensions in English: resolving the contradictory findings [J]. *Language*, 1986, 62(2): 384-414.
- BIBER D. Variation across speech and writing [M]. Cambridge: Cambridge University Press, 1988.
- BIBER D, CONRAD S. Register, genre, and style [M]. Cambridge: Cambridge University Press, 2009.
- BIBER D, CONRAD S, REPPEN R. Corpus linguistics: investigating language structure and use [M]. Cambridge: Cambridge University Press, 1998.
- BIBER D, CONRAD S, REPPEN R, et al. Representing language use in the university: analysis of the TOEFL 2000 spoken and written academic language corpus [M]. Princeton, New Jersey: Educational Testing Service, 2004.
- BIBER D, DAVIES M, JONES J, et al. Spoken and written register variation in Spanish: a multi-dimensional analysis [J]. *Corpora*, 2006, 1(1): 1-37.
- BIBER D, FINEGAN E. Drift and the evolution of English style: A history of three genres [J]. *Language*, 1989, 65(3): 487-517.
- BIBER D, FINEGAN E. The linguistic evolution of five written and speech-based English genres from the 17th to the 20th centuries [C]//RISSANEN M, IHALAINEN O, NEVALAINEN T. History of Englishes: New methods and interpretations in historical linguistics. Berlin: Mouton, 1992: 688-704.
- BIBER D, FINEGAN E. Multi-dimensional analyses of authors' styles: some case studies from the eighteenth century [C]//ROSS D, BRINK D. Research in humanities

- computing. Oxford: Oxford University Press, 1994: 3-17.
- CONRAD S. Expanding multi-dimensional analysis with qualitative research techniques [C]// SARDINHA T, PINTO M. Multi-dimensional analysis, 25 years on: a tribute to Douglas Biber. Amsterdam: John Benjamins Publishing Company, 2014: 273-295.
- CONRAD S. Register variation [C]//BIBER D, REPPEN R. The Cambridge handbook of English corpus linguistics. Cambridge: Cambridge University Press, 2015: 309-329.
- CONRAD S, BIBER D. Variation in English: multi-dimensional studies [M]. New York: Routledge, 2013.
- COSTELLO A, OSBORNE J. Best practices in exploratory factor analysis: four recommendations for getting the most from your analysis [J]. Practical Assessment, Research, and Evaluation, 2005, 10(1): 1-8.
- CROSTHWAITE P. A longitudinal multidimensional analysis of EAP writing: determining EAP course effectiveness [J]. Journal of English for Academic Purposes, 2016, 22: 166-178.
- FRIGINAL E, WEIGLE S. Exploring multiple profiles of L2 writing using multi-dimensional analysis [J]. Journal of Second Language Writing, 2014, 26: 80-95.
- GARDNER S, BIBER D, NESI H. MDA perspectives on discipline and level in the BAWE corpus [R]. Corpus Linguistics 2015, Lancaster, UK, 2015.
- GRIEVE J, BIBER D, FRIGINAL E, et al. Variation among blogs: a multi-dimensional analysis [C]//MEHLER A, SHAROF S, SANTIN M. Genres on the web: corpus studies and computational models. New York: Springer-Verlag, 2010: 45-71.
- GRIEVE-SMITH A. The envelope of variation in multidimensional register and genre analyses [C]//FITZPATRICK E. Corpus linguistics beyond the word: corpus research from phrase to discourse. Leiden: Brill Rodopi, 2007: 21-42.
- HALLIDAY M. On the language of physical science [C]//GHADESSY M. Registers of written English: situational factors and linguistic features. London: Printer, 1988: 162-178.
- HYMES D. Foundations in sociolinguistics: an ethnographic approach [M]. Pennsylvania: University of Pennsylvania Press, 1974.
- KRUGER H, SMITH A. Colloquialization versus densification in Australian English: a multidimensional analysis of the Australian diachronic Hansard corpus [J]. Australian Journal of Linguistics, 2018, 38(3): 293-328.
- KRUGER H, VAN ROOY B, SMITH A. Register change in the British and Australian Hansard (1901-2015) [J]. Journal of English Linguistics, 2019, 47(3): 183-220.
- KYTÖ M, SMITTERBERG E. Diachronic registers [C]//BIBER D, REPPEN R. The Cambridge handbook of English corpus linguistics. Cambridge: Cambridge University Press, 2015: 330-345.

- MCENERY T, XIAO R, TONO Y. Corpus-based language studies: an advanced resource book [M]. New York: Routledge, 2006.
- NINI A. Multidimensional analysis tagger (version 1.3) [EB/OL]. <http://sites.google.com/site/multidimensionaltagger>, 2015.
- NINI A. Register variation in malicious forensic texts [J]. *International Journal of Speech, Language & the Law*, 2017, 24(1): 99-126
- NINI A, GRANT T. Bridging the gap between stylistic and cognitive approaches to authorship analysis using systemic functional linguistics and multidimensional analysis [J]. *International Journal of Speech, Language & the Law*, 2013, 20(2): 173-202.
- RAYSON P, ARCHER D, PIAO S, et al. The UCREL semantic analysis system [R]. Paper presented at the proceedings of the beyond named entity recognition semantic labelling for NLP tasks workshop, Lisbon, Portugal, 2004.
- REPPEN R. Variation in elementary student language: a multi-dimensional perspective [D]. Flagstaff: Northern Arizona University, 1994.
- REYNOLDS D. Linguistic correlates of second language literacy development: evidence from middle-grade learner essays [J]. *Journal of Second Language Writing*, 2005, 14(1): 19-45.
- SARDINHA T, PINTO M. American television and off-screen registers: a corpus-based comparison [J]. *Corpora*, 2017, 12(1): 85-114.
- TOUTANOVA K, KLEIN D, MANNING C, et al. Feature-rich part-of-speech tagging with a cyclic dependency network [R]. *Proceedings of the 2003 human language technology conference of the north American chapter of the association for computational linguistics*, 2003.
- WATSON G. A multidimensional analysis of style in Mudrooroo Nyoongah's prose works [J]. *Text-Interdisciplinary Journal for the Study of Discourse*, 1994, 14(2): 239-286.
- WEIGLE S, FRIGINAL E. Linguistic dimensions of impromptu test essays compared with successful student disciplinary writing: Effects of language background, topic, and L2 proficiency [J]. *Journal of English for Academic Purposes*, 2015, 18(3): 25-39.
- XIAO R. Multidimensional analysis and the study of world Englishes [J]. *World Englishes*, 2009, 28(4): 421-450.
- YAN X, STAPLES S. Fitting MD analysis in an argument-based validity framework for writing assessment: Explanation and generalization inferences for the ECPE [J]. *Language Testing*, 2019, 37(2): 189-214.
- 胡春雨, 谭金琳. 中美企业致股东信语域特征的多维分析[J]. *外语与外语教学*, 2020 ( 6 ): 66-75.
- 江进林, 许家金. 基于语料库的商务英语语域特征多维分析[J]. *外语教学与研究*, 2015 ( 2 ): 225-236.

- 李端阳, 王志军. 基于语料库的海关新闻英语语域特征多维分析[J]. 西安外国语大学学报, 2019 (1): 27-32.
- 潘璠. 中国非英语专业本科生和研究生书面语体的多特征多维度调查[J]. 外语教学与研究, 2012 (2): 220-232.
- 孙亚, 崔子璇. 基于隐喻使用的多维法与语域分析[J]. 外语学刊, 2020 (3): 12-19.
- 王伟. 英语 TED 演讲语篇语域特征多维分析[J]. 外语教学, 2021 (2): 23-28.
- 武姜生. “学术交流 e-mail” 文体特征的多维度分析[J]. 外语与外语教学, 2004 (2): 53-57.
- 吴瑛, 李莉, 宋韵雅. 多种声音一个世界: 中国与国际媒体互引的社会网络分析[J]. 新闻与传播研究, 2015 (9): 5-21.
- 许家金. 美国语料库语言学百年[J]. 外语研究, 2019 (4): 1-6.
- 张一宁, 孙彩慧, 李晔. 中外语言类期刊高被引论文英文摘要语言特征多维分析[J]. 外语电化教学, 2018 (4): 64-71.
- 赵朝永. 基于语料库的《金瓶梅》英文全译本语域变异多维分析[J]. 外语教学与研究, 2020 (2): 283-295.
- 赵朝永, 王文斌. 中国英语学习者语域变异多维分析: 英汉时空特质差异视角[J]. 外语电化教学, 2017 (4): 71-78.

通信地址: 100049 北京市 中国科学院大学外语系 (钱玉彬)  
100029 北京市 对外经济贸易大学英语学院 (孙亚)



# 视觉动词“看”的多义性： 基于语料库的行为特征分析<sup>\*</sup>

北京语言大学 黄静雯 天津师范大学 李金妹 四川外国语大学 胡志勇

**提要：**本文基于语料库的行为特征分析法，从共时的角度研究了视觉动词“看”的多义性，包括分析“看”的语义特点以及基于聚类分析结果确定不同义项间的关系。研究发现，视觉动词“看”主要包括两大类语义。第一大类为视觉义通过认知隐喻引申得到的“社交活动义”，第二大类为认知义与视觉义相关联的语义。第二大类语义间的关系复杂，体现了视觉动词“看”的语义从视觉义向认知义转变的进程。在这部分语义中，视觉义“眼睛感受外界事物”与认知义“被视作，对待”的联系最为紧密，其他认知义“观察并判断”“观察并分析，认为”“取决于”与视觉义“眼睛感受外界事物”的联系由密到疏。最后，本研究根据聚类分析的结果，构建了相对完整的“看”字多义语义网络，实现了基于行为特征分析法量化研究汉语多义现象的研究目标。

**关键词：**行为特征分析法、语料库、视觉动词“看”

## 1 引言

多义词普遍存在于各种语言中，一词多义现象体现了语言的灵活性、适应性和创造性，展现了人类语言的强大生命力（李福印 2008），因此一直是学界的研究热点和重点。不同学者曾对词类的多义性作出解释并提出了相关的假说与模型，其中颇具影响力的是多义词的辐射网络模型（Rosch 1973），即以一个原型义为中心，相关联的派生义从中心向四周发散开来。虽然这种模型对词类的多义性有一定解释力，但目前大部分基于辐射网络模型的研究在确定多义词本义以及分析义项间关系时都是凭借主观经验的定性研究，缺乏足够的说服力。甚至在有些研究中还存在多义词的本义确定不准确、义项区分不明等问题（Geeraerts 1993）。

<sup>\*</sup> 本文系国家社科基金项目“因果构式力动态实证研究”（20XYY001）、天津市哲学社会科学规划重点项目“工具因果关系事件的理论模型构建及汉语表征研究”（TJYY22-009）阶段性成果。李金妹为本文通讯作者。

作者贡献：

黄静雯：数据收集、数据分析、初稿撰写、字数占比（80%）；

李金妹：选题构思、研究方法、讨论结论、初稿撰写、字数占比（20%）、修改润色。

胡志勇：数据分析。



值得注意的是，确定多义词原型义以及义项间的关系是多义词研究的重要任务，因为它揭示了多义词的语义核心与不同义项的演变情况。已有的定性研究无法准确地完成这些任务，因此恰当引入定量研究可以解决目前存在的问题。其中一种定量研究方法，即基于语料库的行为特征分析法（corpus-based behavioral profile analysis，以下简称BP分析法）备受学界瞩目，学界运用该方法对多义词的研究已经取得了许多理论成果，如吴淑琼等（2021）、江艳艳（2022），但对多义词“看”的BP分析研究尚付阙如。

“看”属于感知动词的范畴，Viberg（1983）曾对感知动词进行了分类与等级排序，他认为人的感知主要遵守“sight > hearing > touch > smell/taste”的顺序。通过这个排序可以看出视觉信息是最重要且直观的，而且视觉信息具有很强的主观性和主动性。由于视觉感知动词的特殊性，深入研究视觉动词一直是认知语义学的热点，国外许多学者曾从不同角度对其进行了细致的研究。其中比较普遍的是对视觉动词see特点的探究，例如Stamenkovic（2010）曾对塞尔维亚语的视觉动词gledati和英语的视觉动词look所对应的下义词进行了探究，以便更好地揭示视觉动词的语义转移方式与路径。另外，也有从教学角度出发的研究，如Sato（2015）曾借助实验比较了58名日本高中生在基于模式教学与基于翻译教学两种情境下对英语视觉动词look和see的掌握情况，并指出第一种教学模式更高效。

英语视觉动词see、watch、look在汉语中对应视觉动词“看”，而目前国内对“看”的探究成果十分丰富，主要包括以下4个方面：（1）从语义系统、语义分布的角度出发，致力于描绘“看”的多义语义网络，帮助学习者克服理解和学习困难（王文斌、周慈波 2004；欧德芬 2014）；（2）从语义结构、语义角色的角度出发，分析并探究与“看”有关的常见构式的语义，包括“看看”“看了看”“看过”等（杨霞 2010；陈辰 2020）；（3）从语义演变的角度出发，结合丰富的历时语料，对“看”的古今语义演变进行研究与考察（张云 2009；段颖玲 2010；张子华 2018）；（4）从对外汉语教学的角度出发，利用偏误分析等方法研究母语非汉语学习者的学习情况与教学策略，这部分研究起步相对较晚，研究成果比较局限且大多为思辨研究（陈梦 2015；石莹莹 2017；王宇 2019）。

总的来说，对英语视觉动词探究的角度多样、研究内容丰富，大部分为定性研究。在对汉语本体的研究中，对视觉多义词“看”的研究主要集中在语义分析、语法化历程、隐含视觉动词的对比等方向。在对外汉语的相关研究中，研究重心主要集中在对“看”及其近义词的偏误分析上，部分研究运用了语料库，但只限于获取文章分析的语料，没有采用量化分析的方法，而且这部分研究对“看”的语义网络构建不够完整，义项之间的关系也没有清晰指出，因此本研究致力于弥补这些研究的不足之处。

鉴于此，本文将利用基于语料库的行为特征分析法对北京大学CCL语料库中

与“看”相关的随机语料进行分析、标注，通过一定的统计手段构建相对完整的“看”的多义语义网络，并对其义项之间的关系进行分析，揭示“看”的多义语义网络内部细节，实现量化研究汉语词类多样性的目的。

2 研究步骤

本文所采取的量化研究方法为基于语料库的行为分析法，该方法主要运用于词汇语义研究中，致力于揭示词语在特定语境的使用特点，如形态、句法、语义、功能层面的特征，并进一步借助统计工具揭示语言现象背后的规律。利用该方法研究词类语义一般包括四个步骤，即前三步的数据加工以及第四步的数据评价。

第一步，从语料库中检索与本文研究对象“看”有关的所有用例。本文利用北京大学现代汉语语料库（CCL），以“看”为关键词检索并获得相关语料。从获得的20,000条语料中，随机取3次1,000条语料，共计3,000条语料。按照语料中的“看”必须为动词，且如果每个语义下达到200条语料则不再取更多例句的标准，最终筛选得到2,177条语料进入第二步骤的人工标注与分析环节。其中，如果“看”的读音不同则意义不同，包括/k ā n/和/k à n/两类，由于随机选取的3,000条语料中读作一声“看”的语料仅有13条，语料数量不足，而且根据李仕春（2020）借助CCL语料库建立的“看”字平衡语料库可知，一声“看”对应的义项“守护，照料”和“看押，监视”约占全部语料总数的1.87%，占比较小，统计学意义不足。更重要的是，一声“看”的分布与四声“看”相比更为局限，大部分情况下需搭配主语和宾语，例如“你在这把犯人看好了”，而四声“看”则能在各种语境下使用，因此，一声“看”与四声“看”应分开研究。本文仅研究四声“看”。

第二步，对符合筛选标准的2,177条语料进行人工标注，即分析与研究对象“看”相关的形态、句法、语义等特征，具体对视觉动词“看”的标注特征分析包括语义特征、形态句法特征<sup>1</sup>。

（1）语义特征

表1 视觉动词“看”的义项

序号	义项	示例
看1	眼睛感受外界事物	你喜欢不喜欢[看]电影？
看2	看望，拜访	明天我们去[看][看]老师。
看3	安排，准备（旧）	双料春爷喝道：“还不赶紧与压寨夫人[看座]！”
看4	诊治	同理，大夫说“我去[看]病”是给人[看]病，病人说就是让人给他[看]病。

（待续）

(续表)

序号	义项	示例
看5	观察并分析	就我国目前情况来[看], 人人享有受教育的机会, 只能对基础教育而言。
看6	被视作, 对待	口语被[看]成为不登大雅之堂的俚言俗语, 不予重视。
看7	观察并得出结论	成己的同时, 一定要[看]到还要成人。
看8	取决于	一个人对于社会的有用与否, 完全[看]遗传如何。

(2) 标识码类别和标识码水平

表2 标识码类别与标识码水平

标识码类型	标识码	标识码水平
形态	重叠式	VV、V—V
	复合式	V1+V2
	词尾	“成”类、“作”类、“到”类、“着”类
	词缀	“来”类
	零形态	
句法	句子结构	主谓结构、主谓宾结构、动宾结构、动补结构、其他结构
	及物性	及物、不及物
	是否为否定	是、否
	主语	有生命(第一、第二、第三人称)、无生命、零主语
	宾语	有生命名词、有生命名词词组、无生命名词、无生命名词词组、人称代词、指示代词、小句、零宾语
	状语	程度副词、频度副词、范围副词、时间副词、其他副词、无副词修饰
	补语	结果补语、趋向补语、状态补语、可能补语、零补语

第三步, 将人工筛选并标注好的语料制作成一个绝对频率共现表, 如表3所示。

表3 部分视觉动词“看”绝对频率共现情况

标识码	标识码水平	眼睛感受外界事物	看望, 拜访	安排, 准备 (旧)
重叠式	VV式	22	2	0
	V—V式	7	0	0
复合式	V1+V2式	44	3	0
	“成”类	2	0	0
词尾	“作”类	0	0	0
	“到”类	157	0	0
	“着”类	19	0	0
词缀	“来”类	41	0	0
零形态		276	4	2

为了进一步统计与研究, 需要将上表中的绝对频率转换为相对频率, 如表4所示。

表4 部分视觉动词“看”相对频率共现情况

标识码	标识码水平	眼睛感受外界事物	看望, 拜访	安排, 准备 (旧)
重叠式	VV式	0.03873	0.22222	0
	V—V式	0.01232	0	0
复合式	V1+V2式	0.07746	0.33333	0
	“成”类	0.00352	0	0
词尾	“作”类	0	0	0
	“到”类	0.27640	0	0
	“着”类	0.03345	0	0
词缀	“来”类	0.07218	0	0
零形态		0.48591	0.44444	1

第四步, 运用层次聚类分析法 (hierarchical agglomerative cluster, 简称HAC) 对数据进行定量分析。本文用到的层次聚类分析法是一种探索性统计方法, 主要采用自下而上的方式对数据进行聚类 (吴淑琼等 2021)。该方法主要用于呈现出一个类 (cluster) 内部成员的最大相似性以及类与类之间的差异性, 从而揭示数据的内部结构和趋势情况 (Divjak & Gries 2006)。

### 3 结果分析

#### 3.1 视觉动词“看”整体使用情况

通过对视觉动词“看”各项语义的数据进行人工标记与定量分析，最终得到关于视觉动词“看”的各义项语料占比和聚类树形图（图1）。

“观察并分析；认为”是最常出现的语义，占比约28%，同时“眼睛感受外界事物”和“观察、判断并得出结论”这两个义项同样占比大于20%，分别为26%和23%。“看望，拜访”“诊治”“安排，准备（旧）”这三个义项占比较小，其中“安排，准备（旧）”这一语义占比最少，在2,177条语料中只有2条，占比为0.09%，约为0。部分原因在于该义项在现代汉语的使用频率低。由此可以推断，视觉动词“看”所引申得到的认知义在使用上占主导地位，而视觉义“眼睛感受外界事物”，使用频率较高。

图1可以清楚地反映视觉动词“看”内部的语义关系。我们可以看到“看”的语义主要分为两大类，一类为“看”通过认知隐喻引申得到的“社交活动义”（位于图2的右侧），另一类为“看”的认知义与视觉义的结合（位于图2的左侧）。基于聚类分析“以类而聚”的原则，可以得出“看”的认知义占主导地位，且认知义的内部也分为不同的层次结构，在位于左侧的“看”中可以进一步分为两小类，一类义项为“取决于”，一类义项为“观察并分析”“观察并得出结论”“眼睛感受外界事物”“（被）视作，对待”，而且认知义“（被）视作，对待”和视觉义“眼睛感受外界事物”的结合最紧密。总的来看，视觉义“看”虽然占比较大，但使用频率仍低于认知义“看”，接下来我们会对每个“看”的语义特点，使用特征作进一步解释与分析，并尝试构建“看”的语义网络。

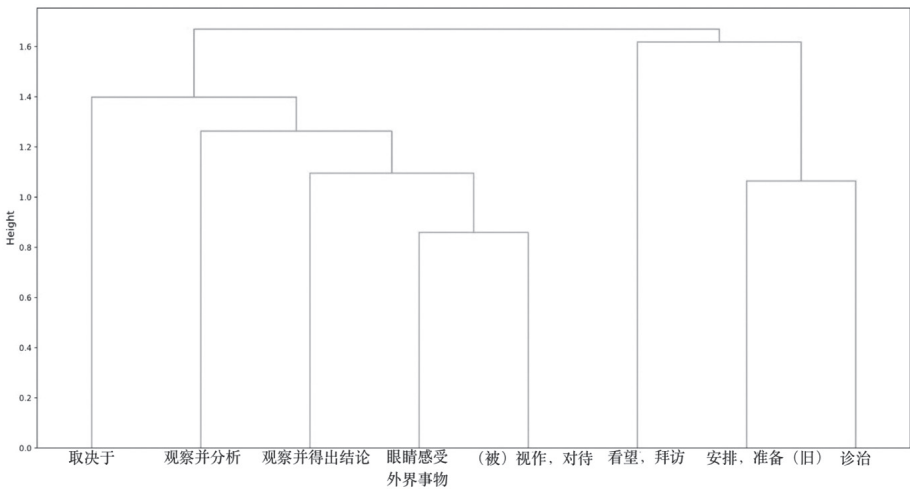


图1 “看”的义项聚类分析树形图

3.2 视觉动词“看”的原型义及多义语义网络

学界对判断多义词的原型义有许多不同的方法，其中最常见的判断标准是基于义项出现的“频率”，即越是接近词类本义的义项，发生的频率越高（Schmid 2000）。这种判断标准存在不足，因为原型义不仅与义项的使用频率有关，而且与该义项出现的语境有关。这一语言事实与本文使用的行为特征分析法基本原理高度契合，即行为特征是动词的句法和搭配模式以及各种模式的使用频率（Hanks 1996）。由此可知，如果要确认视觉动词“看”的原型义，我们应找出在各类语境下都能使用的义项，即涉及最多标识水平数的义项，而不是出现频率最高的义项。根据人工分析与标注的结果，我们可以得到“看”各个义项的标识码水平总数，见表5。

表5 视觉动词“看”各义项的标识码水平总数

义项	眼睛感 受外界 事物	看望， 拜访	安排，准 备（旧）	诊治	观察并 分析	被视作， 对待	观察并得 出结论	取决于
标识码 水平数	41	20	11	12	36	38	33	20

从表5中可以看出，视觉义“看”的标识码水平数最多，因此“眼睛感受外界事物”是视觉动词“看”的原型义。虽然认知义“看”的义项（如“被视作，对待”“观察并分析”等）标识码水平数也较高，但整体看来“眼睛感受外界事物”这一义项使用最灵活，符合判断多义词原型义的标准。

接下来我们将深入分析“看”的各个义项并基于所获取的研究数据构建“看”的相对完整的语义网络，根据图1的聚类分析结果，我们将“眼睛感受外界事物”作为原型义，则“看”的多义语义网络可以构建如图2。

由图2可以清楚地看出各个义项的关系与内部连接的情况。首先位于第一分支的义项主要分为两类，第一类“看望，拜访”，第二类“安排，准备（旧）”“诊治”，在这一部分语义中，视觉动词“看”由原型义“眼睛感受外界事物”这一视觉义拓展为社交活动义体现了从“自己看（某个东西）”到与他人建立联系的过程，而且根据统计数据结果可知在使用特点、具体搭配等方面，义项“安排，准备（旧）”“诊治”的关系更近。

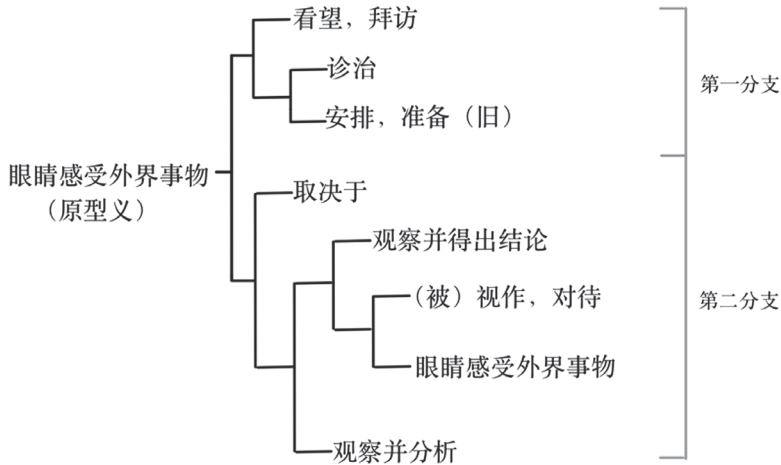


图2 视觉动词“看”的多义语义网络

而位于第二分支的义项之间连接相对复杂，但整体也可以看作两大类、4小类，第一大类（亦第一小类）“取决于”，第二大类“观察并得出结论”“（被）视作，对待”“眼睛感受外界事物”“观察并分析”。区分这两大类的主要依据在于“看”的主语类别，即第一大类（亦第一小类）义项“取决于”的主语一般为物，而第二大类义项的主语大多情况下都为人，因此具备主观判断的认知能力，如例（1）—（4）。

（1）儿童也未必一定会成为一个音乐家，除了遗传给他的可能性外，还要看他所处的社会条件，所受的教育和个人的努力如何来决定。

（2）一定的惩罚是否具有可接受性主要看其是否符合学生的年龄特征和现实社会角色。

（3）只有愚人才看不出这个明显的事实。

（4）后面我们将看到，道家讲“无为”的学说。

例（1）的主语为“儿童是否成为一个音乐家”这一情况，例（2）的主语为“一定的惩罚是否具有可接受性”这一说法，均为无生命主语。而在例（3）和例（4）中，主语分别是具有生命的“愚人”和“我们”，均为具有认知能力的主语，因此主语有无生命是区分右侧两大类义项的标准。

第二大类又可分为3个小类，分别为第二小类、第三小类和第四小类。第二小类是义项“观察并得出结论”，强调“观察—思考—得出相应结论”这一过程，



而且通常在“看”后伴随结果补语，凸显“思考之后作出的判断”，如例（5）—（6）。

（5）从上面的例子中可以**看出**，变换不是孤立地就一个结构本身来分析，而是借助于其他的结构来说明。

（6）这个不同，有些道家的人**看得**很清楚。他们用“忘”字表达其方法的诀窍，这是很有深意的。

句（5）能够体现通过认知判断得到的结论，即“变换不是孤立地就一个结构本身来分析，而是借助于其他的结构来说明”，句（6）也包含“观察—思考—得出相应结论”这一过程。

除了上述分析，从图2“看”的语义网络中我们还可以看出第二分支的这一部分义项反映了视觉动词“看”的多义复杂性，因为“看”的视觉义与认知义相关性高，在聚类树形图里联系紧密。最能体现这一特点的是第三小类，即“眼睛感受外界事物”与“（被）视作、对待”这两个义项，二者关系密切。即使前者为视觉义，后者为认知义，根据生活常识可以判断出后者的认知义是基于视觉义发展而来的，而且应该处于“看”的视觉义向认知义发展的开始位置。如例（7）—（10）。

（7）你**看**电影不啊？

（8）而且用这些不同的方音还能去读古书，用不着像**看**拼音文字写的古书那样，得先学古音。

（9）我们应该把转注问题**看**作文字学史上已经过时的一个问题，完全没有必要再去为它花费精力。

（10）口语被**看**成为不登大雅之堂的俚言俗语，不予重视。

例（7）和例（8）中的“看”为视觉动词，分别意为“观看”和“阅读”，例（9）和例（10）不仅带有“眼睛感受外界事物”的视觉含义，而且有“思考之后作出判断”的认知含义。值得注意的是，义项“（被）视作、对待”这类“看”常与“作”“成”等成分<sup>2</sup>构成一个整体，不可分割，而义项“眼睛感受外界事物”对应的“看”是及物动词，总是与宾语结合构成短语，例如“看电影”“看古书”。

第四小类包括义项“观察并分析”，这一义项与第二小类最突出的区别在于其一般不带有“思考之后作出的判断与选择”，仅包含“观察—思考”这一过程，这



一细微的区别可以通过如下两个例子体现。

(11) 成己的同时，一定要**看到**还要成人。

(12) 从表面上**看**中国哲学，不能说这些人说错了，因为从表面上**看**中国哲学，无论哪一家思想，都是或直接或间接地讲政治，说道德。

例(11)中包含“看”的结果和判断，即“还要成人”，而且这个句子中“看”与“到”相结合凸显结果补语，因此属于第二小类。例(12)中的“看”仅仅停留在“考虑、思考”的层面，即“从表面上来考虑/思考中国哲学”，而不作出进一步判断，因此属于第四小类。

## 4 结语

本文对认知语言学热点话题多义词的原型义与语义网络进行了探究，利用语料库的行为特征分析法，量化分析了视觉动词“看”的多义情况。本研究从CCL语料库中随机取得3,000条语料并根据一定标准选取了2,177条有效语料。研究结果显示：视觉动词“看”共有8个义项，“看”的原型义是视觉义“眼睛接触外界事物”，但“看”的视觉义与认知义联系非常紧密，而且二者的使用频率都很高。另外，基于本文的数据与分析，“看”的多义语义网络得以构建，不同义项之间的疏密关系也得以解释。

值得一提的是，本文运用的基于语料库的行为特征分析法在研究词类的多义现象方面具有一定的优势，因为国外已有许多利用该方法研究词汇语义的成果，但国内使用BP分析法的定量研究尚存不足，因此本研究对目前的现代汉语多义词研究进行了一定的补充。另外，不同于普通的语料库研究方法，基于语料库的行为特征分析法能够很好地揭示词类在具体语境中的使用情况与特点。因此，本研究不仅对汉语的多义词量化研究进行了一定的补充，而且拓展了行为特征分析法的适用范围，展示出其可以用于更广的研究范畴。

### 注释

- 1 感谢第五届认知语义学研讨会上的专家对本研究语料筛选提出的建议，故此处对本文的语料筛选标准做进一步说明。
- 2 本文认为此处的“看”与“成”“作”等成分构成一个词，其中“成”“作”有黏着语素的特点，其意思类似英语中的as。

## 参考文献

- DIVJAK D, GRIES S. Ways of trying in Russian: clustering behavioral profiles [J]. *Corpus Linguistics and Linguistic Theory*, 2006, 2(1): 23-60.
- GEERAERTS D. Vagueness's puzzles, polysemy's vagaries [J]. *Cognitive Linguistics*, 1993, 4(3): 223-272.
- HANKS P. Contextual dependency and lexical sets [J]. *International Journal of Corpus Linguistics*, 1996, 1(1): 75-98.
- ROSCH E. Natural categories [J]. *Cognitive Psychology*, 1973, 4(3): 328-350.
- SATO M. Effectiveness of acquiring of basic verbs by using core schema-based instruction [J]. *International Journal of Languages, Literature and Linguistics*, 2015, 1(1): 34-38.
- SCHMID H. English abstract nouns as conceptual shells: from corpus to cognition [M]. Berlin & New York: Mouton de Gruyter, 2000.
- STAMENKOVIC D. Metaphoric and extended uses of the hyponyms of the verbs look in English and gledati in Serbian [J]. *Facta Universitatis*, 2010, 8(1): 19-33.
- VIBERG A. The verbs of perception: a typological study [J]. *Linguistics*, 1983, 21(1): 123-162.
- 陈辰. “看”“看见”“看到”与“看完”的语义差异探究——认知构式语法视角[J]. *外国语文*, 2020 (4): 83-92.
- 陈梦. 与“看”有关的近义词语语义语用考察及对外汉语教学策略研究[D]. 昆明: 云南大学, 2015.
- 段颖玲. 论“看看”语义的古今演变[J]. *求索*, 2010 (4): 174-176.
- 江艳艳. 温度词“冷”的语义演变研究: 基于语料库的行为特征分析[J]. *外语研究*, 2022 (6): 33-41.
- 李福印. 认知语言学概论[M]. 北京: 北京大学出版社, 2008.
- 李仕春. 框架语义学视阈下的词义衍生研究——以多义词“看”为例[J]. *东北师大学报(哲学社会科学版)*, 2020 (1): 20-25.
- 欧德芬. “看”的语义网络[C]//澳门大学. 第十五届汉语词汇语义学国际研讨会论文集. 中国澳门: 澳门大学, 2014.
- 石莹莹. 单双音节动词“看”和“看见”的习得偏误研究[D]. 安阳: 安阳师范学院, 2017.
- 王文斌, 周慈波. 英汉“看”类动词的语义及词化对比分析[J]. *外语教学与研究*, 2004 (6): 412-419.
- 王宇. 对外汉语教学视角下“看”和“见”的语义研究[D]. 太原: 山西大学, 2019.
- 吴淑琼, 刘迪麟, 冉苒. 心理动词“想”的多义性: 基于语料库的行为特征分析[J]. *外语与外语教学*, 2021 (5): 1-13.

杨霞. “看+把+N/Pron+V/A+的”格式的语义语用分析[J]. 教学与管理, 2010 (3): 72-73.

张云. 视觉动词“看”、“见”使用情况历时、共时考察[D]. 武汉: 华中师范大学, 2009.

张子华. 从视觉动词的词义特征看古汉语动词“见”的语法化[J]. 榆林学院学报, 2018 (5): 86-90.

**通信地址:** 100083 北京市 北京语言大学语言学系 (董静雯)

300387 天津市 天津师范大学外国语学院 (李金妹)

400031 重庆市 四川外国语大学英语学院 (胡志勇)

# 不同水平中国英语学习者虚化动词搭配使用及迁移效应研究<sup>\*</sup>

北京航空航天大学 闫盛德 高霞

**提要：**本文基于EFCAMDAT语料库，结合定量与定性分析，研究不同水平中国英语学习者使用虚化动词make和take的搭配行为，探讨学习者虚化动词搭配能力发展，并与母语为西班牙语的英语学习者进行对比，探究母语迁移效应。研究发现：初级学习者使用虚化动词搭配意愿较低，搭配意义最为具体；中级学习者倾向于过度使用虚化动词搭配；高级学习者搭配丰富度和抽象度最高。中低水平学习者文本中出现虚化动词意义泛化和混用现象，或与搭配名词的多重语义特征相关。研究同时解释了母语的类型学差异影响迁移效应的不同机理，并从教材编写、课堂教学等方面提出了相应的教学建议。

**关键词：**虚化动词、搭配、二语教学、学习者语料库、母语迁移

## 1 引言

虚化动词 (delexicalized verb)，也称轻动词 (light verb)，指多词序列中语义弱化的动词，包括do、have、give、make、take等。这些动词作为主动词的意义较弱，须由共现的事件名词 (eventive nouns) 或名词类转 (type-shifting) 表达句子核心意义 (Pustejovsky 1995)。虚化动词与名词的搭配使用能赋予语言修饰更多灵活度，代替更正式的表达，在人际交流中传达更多意义 (谢家成 2010)。然而虚化动词搭配的语义理据较弱，容易受到词语同义现象影响 (Liao 2010)，易错性高，是二语学习的难点 (张爱朴 2014)。Lewis (1997) 提倡在二语教学中高度重视虚化动词用法及搭配，相关研究多对比英语学习者与本族语者的使用异同，探究学习者书面语中虚化动词搭配的使用特征 (Nesselhauf 2003; 方秀才 2015等)。现有研究鲜有关注学习者虚化动词搭配的动态发展，也极少讨论不同母语背景下习得的迁移效应。鉴于此，本研究基于不同水平英语学习者习作文本语料库，从

<sup>\*</sup> 高霞为本文通讯作者。

作者贡献：

闫盛德：选题构思、研究方法、数据收集、数据分析、讨论结论、初稿撰写、字数占比 (60%)、修改润色。

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

发展的角度探究中国学习者虚化动词搭配能力的动态变化,并与母语为西班牙语的英语学习者的虚化动词搭配使用进行对比,探究母语迁移影响。

## 2 学习者虚化动词搭配习得的相关研究

Sinclair (2004) 认为词汇在语境中的词义会受到共现词语的影响,丧失部分原有的语义并获得共现词语的部分意义,发生语义虚化 (delexicalization)。动词的虚化最为典型,现代英语本族语者在交际中倾向使用虚化结构来代替相应的动词搭配 (Carter & McCarthy 1997)。虚化动词通常与事件名词搭配,构成“虚化动词+名词”型式,譬如 make a decision、take action、give compliments 等 (Nesselhauf 2005)。“虚化动词+名词”型式是英语学习者最难掌握的搭配结构之一 (Allerton 2002; Liao 2010),探究学习者“虚化动词+名词”搭配的使用特征,提升学习者的英语搭配能力,一直是国内外学者关注的热点。

国外学界关于学习者使用虚化动词搭配的实证研究起始较早,多面向具有中高级二语水平、来自不同母语背景的英语学习者,大体上可分为以下两类。(1) 基于学习者和本族语者语料库的跨群体对比研究,探究学习者的虚化动词搭配特征 (如 Hasselgren 1994; Altenberg & Granger 2001; Kim & Yagong 2016)。研究发现学习者或过少使用虚化动词搭配 (Altenberg & Granger 2001; Wang 2016),或过度使用虚化动词搭配 (Altenberg & Cairns 1983; Hasselgren 1994; Nesselhauf 2003)。多数学习者的虚化动词搭配类型较为单一,不如本族语者丰富 (Altenberg & Granger 2001; Juknevičienė 2008; Wang 2016)。(2) 基于不同母语背景学习者语料库的跨群体对比研究,如法国和瑞典高级英语学习者 (Altenberg & Granger 2001)、中国和瑞典中高级学习者 (Wang 2016) 及西班牙和意大利高级学习者 (Vázquez 2018)。研究发现虚化动词搭配的问题用法与母语背景相关,不同母语背景的写作习惯、论证策略、文化差异等因素会影响学习者的搭配使用 (Wang 2016)。

国内学界相关实证研究亦主要面向中高水平英语学习者,揭示学习者使用虚化动词搭配时存在的问题。研究多基于高校英语学习者习作语料库,如对比非英语专业大学生 (邓耀臣、肖德法 2005; 缪海燕、孙蓝 2005; 桂诗春 2007; 王文宇、李小撒 2018)、英语专业大学生 (张淑静 2002; 王立非、张岩 2007; 刘国兵 2011; 张莎 2011; 朱慧敏、王俊菊 2019) 与本族语者文本中虚化动词搭配使用的异同,或对比研究英语专业与非英语专业学习者群体间的虚化动词搭配特征 (张文忠、杨士超 2009; 付思婧、陈月红 2022)。研究发现,中高英语水平中国学生能熟练使用的虚化动词搭配种类、数量均明显少于本族语者 (邓耀臣、肖德法 2005),倾向于过度使用中学阶段习得的、作为语块存储的高频搭配 (缪海燕、孙蓝 2005)。中国学习者不仅回避使用不确定的虚化动词搭配 (付思婧、陈月红

2022), 也会自创虚化动词搭配(王文宇、李小撒 2018)。

综上所述, 国内外研究多指向中高水平学习者群体, 初级英语学习者受到的关注远远不足, 且语言水平的区分多以大学学制和年级为标准, 判断指标较为模糊。甚少研究关注学习者动态发展, 探究语言水平是否或如何影响学习者虚化动词搭配的使用。国内几乎没有研究比较中国学习者与其他母语背景学习者, 从类型学角度探究虚化动词使用的母语迁移机理。鉴于此, 本研究基于英孚剑桥开放式语言数据库(EF-Cambridge Open Language Database, 简称EFCAMDAT)进行, 该库中学习者英语水平参照欧洲语言共同参考框架(Common European Framework of Reference for Languages, 简称CEFR)作了分级。本研究的主要目的是探究不同水平中国英语学习者习作中虚化动词的搭配使用情况和特征, 并与母语为西班牙语(以下简称西语)的学习者文本中虚化动词的搭配使用特征进行对比, 探究母语迁移影响。西语属罗曼语族, 与日耳曼语族的现代英语同属印欧语系, 而汉语普通话则属汉藏语系中的汉语族, 在类型学上西语比汉语与英语更为接近。在第二语言习得过程中, 类型学特征可导致显著的迁移效应(Håkansson *et al.* 2002)。西语中具有虚化动词用法的动词数量较多, 而汉语中能够充当虚化动词的实义动词数量极其有限(Alba-Salas 2002; Wang & Shaw 2008), 母语或会对两国学习者的虚化动词使用产生不同影响。本研究选取make与take两个典型虚化动词作为研究对象, 原因在于make与take构成的“虚化动词+名词”搭配种类繁多, 用法复杂, 学习者掌握难度较大(刘国兵 2011)。研究拟回答以下问题:

(1) 初级、中级、高级中国学习者的英语习作中虚化动词搭配使用是否存在差异? 是否存在典型发展特征?

(2) 中国学习者的英语习作中虚化动词搭配与西语母语学习者是否存在差异? 是否存在母语迁移效应?

### 3 研究方法

#### 3.1 研究语料

本研究所使用的学习者语料库EFCAMDAT由英国英孚教育公司和剑桥大学理论与应用语言学系合作建设, 总库容约为8,300万词, 包含17万名来自198个国家或地区学习者的约118万篇习作。EFCAMDAT将学习者的英语水平分为16级, 对应CEFR中的A1至C2等级。其中高水平学习者习作数量相对较少, C1、C2等级语料达不到学习者搭配研究要求的语料规模(Du *et al.* 2022), 故研究组选取EFCAMDAT中A1、A2、B1、B2等级, 学习者国籍为中国, 母语为汉语普通话的习作, 从A1至B2等级中各随机抽取23万余词的习作, 总计932,597词。考虑到高水平学习者使用的虚化动词搭配丰富度较高(朱慧敏、王俊菊 2019), 可为母语



迁移效应提供更多样化的证据支持，对照组语料库选取墨西哥国籍、西语母语学习者的B2等级习作，提取EFCAMDAT中的全部样本总计13万余词。研究语料详见表1。

表1 学习者语料库基本信息

学习者国籍	学习者母语	CEFR 等级	语料库名称	语料库库容 ( tokens )
中国	汉语	A1	CN_A1	232,492
		A2	CN_A2	230,347
		B1	CN_B1	231,057
		B2	CN_B2	238,701
墨西哥	西班牙语	B2	SP_B2	136,370

3.2 研究工具与方法

本研究选取在线语料库工具 Sketch Engine、USAS，编程工具Python，统计工具R等作为主要研究工具，具体步骤如下：

（1）将清理好的语料导入Sketch Engine赋码，采用CQL（corpus query language）检索make/take右跨距5以内的名词搭配，公式为[lemma = "make"][] {0,5}[tag = "N.\*"] 和 [lemma = "take"][] {0,5}[tag = "N.\*"]。检索结果通过Python提取词元合并，手工筛选属于虚化动词搭配的用法。选取英国国家语料库（British National Corpus，BNC）作为参照语料库计算t值，取临界值2作为判定构成搭配的标准。得到搭配词表后，计算频次与标准化频次，探究学习者使用make/take虚化动词搭配的总体特征。

（2）本文采用虚化动词搭配中的名词成分特征表征不同水平学习者的搭配特征（Uchida 2015；Du *et al.* 2022）。对名词成分进行语义域和复杂度的定量分析，使用USAS（UCREL Semantic Analysis System）工具进行搭配名词成分的语义域标注。USAS是兰卡斯特大学Paul Rayson团队开发的在线语料库工具，能够进行词性及语义的自动标注，对于英语文本的识别准确率可达92%（Piao *et al.* 2015），共包含21个顶层语义域（A-Z），如表2所示。将自动标注结果整理校验后，进行对应分析（correspondence analysis），探究各组学习者习作中虚化动词搭配名词语义域的区别与联系。

（3）选取搭配案例，定性分析不同等级学习者虚化动词搭配名词的语义域特征，发掘错误搭配原因，并通过母语与本族语的类型学一致性与异质性，解释不同水平中国学习者及西语母语学习者使用虚化动词搭配时的正向与负向迁移现象。

表2 USAS语义赋码集

编码	语义域名称
A	General & Abstract Terms 概括和抽象术语
B	The Body & the Individual 身体和个体
C	Arts & Crafts 艺术和工艺
E	Emotional Actions, States & Processes 情绪行为、状态和过程
F	Food & Farming 食物和耕种
G	Government & the Public Domain 政府和公共领域
H	Architecture, Buildings, Houses & the Home 建筑设计、建筑物、房子和家庭
I	Money & Commerce 财产和贸易
K	Entertainment, Sports & Games 娱乐、运动和游戏
L	Life & Living Things 生命和生物
M	Movement, Location, Travel & Transport 动作，位置，旅行和交通
N	Numbers & Measurement 数字和测量
O	Substances, Materials, Objects & Equipment 物质、材料、物体和设备
P	Education 教育
Q	Linguistic Actions, States & Processes 语言行为、状态和过程
S	Social Actions, States & Processes 社会行为、状态和过程
T	Time 时间
W	The World & Our Environment 世界和环境
X	Psychological Actions, States & Processes 心理行为、状态和过程
Y	Science & Technology 科学和技术
Z	Names & Grammatical Words 名称和语法词汇

4 结果与讨论

4.1 不同水平中国学习者文本中虚化动词搭配的发展特征

4.1.1 中国学习者文本中虚化动词搭配的定量分析

中国学习者文本中虚化动词搭配的总体特征见表3。从形符来看，A1水平学习者的虚化动词搭配频次极低，表明这一群体使用虚化动词搭配的意愿较低，而



B1水平学习者的使用频次则过高。随着学习者水平增长，虚化动词搭配使用频次整体呈上升趋势。从类符与类符/形符比来看，中国学习者虚化动词搭配的词汇丰富度随英语水平提高而逐渐上升。其中B1水平学习者的词汇丰富度与过高的使用频数并不相符，存在过度使用虚化动词搭配的倾向。这一发现与邓耀臣、肖德法（2005）的研究结果一致，但不支持朱慧敏、王俊菊（2019）的研究结果，即高年级中国学生少用have的虚化动词搭配，原因或在于have与make、take等不同虚化动词的使用特征存在差异。发现亦不支持Altenberg & Granger（2001）的研究结果，即法国和瑞典高级学习者过少使用make的虚化动词搭配，原因或在于不同母语背景影响目的语的方式和结果存在差异。

表3 中国学习者虚化动词搭配频次和标准化频次（每万词）

语料库	CN_A1	CN_A2	CN_B1	CN_B2
make/take 虚化动词 搭配频次（标准化 频次）形符	196（8.4）	974（42.3）	1,882（81.5）	1,248（52.3）
make/take 虚化动词 搭配频次（标准化 频次）类符	18（0.8）	75（3.3）	134（5.8）	179（7.5）
类符/形符比	9.2%	1.5%	7.1%	14.3%

使用USAS语义域标注不同水平学习者 make/take 虚化动词搭配名词情况，对应分析结果见表4。维度1、2对总惯量的累计贡献率为80.7%，可有效阐释各组语料库中虚化动词搭配名词语义域间的关系，建立二维坐标解释变量间的关系，如图1。维度1区分了初级与中高级水平学习者：A1与A2水平中国学习者位于负向，B1与B2水平学习者位于正向。结合维度2可确定与各水平组学习者相关性较强的USAS语义域。

表4 中国学习者虚化动词搭配名词 USAS 语义域的对应分析结果

维度	奇异值	惯量解释比例（%）	惯量累积比例（%）
维度1	0.093	53.5	53.5
维度2	0.047	27.2	80.7

注： $X^2 = 86.21$ ， $df = 57$ ， $P < 0.01$

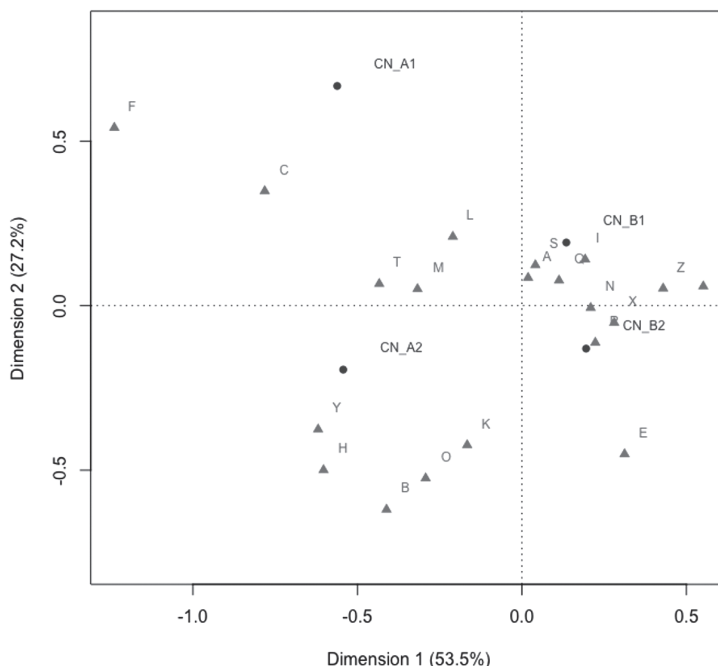


图1 不同水平中国学习者语料库与虚化动词搭配名词成分语义域的对对应分析

A1水平学习者文本中所使用的虚化动词搭配名词多为C (Arts & Crafts, 如take a party)、F (Food & Farming, 如take a drink) 语义域, A2水平多为H (Architecture, Buildings, Houses & the Home, 如make a living)、Y (Science & Technology, 如take a picture) 语义域。这些语义域下的名词意义较为简单具体, 多代指实体存在的物质 (Kurteš & Saville 2008)。B1水平多为A (General & Abstract Terms, 如make sure/a mistake, take advantage)、I (Money & Commerce, 如make money/profit, take charge)、Q (Linguistic Actions, States & Processes, 如make an apology/description)、S (Social Actions, States & Processes, 如make friends, take order) 语义域, B2水平多为E (Emotional Actions, States & Processes, 如make complaints, take courage/vengeance)、N (Numbers & Measurement, 如make a deal, take measures)、P (Education, 如take a test/a course/training)、X (Psychological Actions, States & Processes, 如make judgment/assessment) 语义域。中高水平中国学习者使用的E、Q、S、X语义域均包含行为、状态、过程, 抽象程度较高, 且涉及个体与社会多方面话题, 显著区别于初级学习者种类单一、意义具体的虚化动词搭配名词语义域。学习者习得语义域分类下的词汇比习得随机出现的词汇更有成效 (Lewis 1997), 进一步地针对各语义域内名词成分的梳理与教学, 将更有利于各水平学习者, 尤其是抽象搭配名词掌握不足的初级学习者的虚化动词搭配能力发展。

#### 4.1.2 中国学习者文本中虚化动词搭配的定性分析

在A2、B1水平中国学习者群体中，我们发现了同一名词成分的多重语义特征所导致的虚化动词搭配误用现象。例（1）—（4）是B1水平学习者误用的make a point虚化动词搭配实例。BNC语料库中make a point的检索行搭配名词point属于Q（Linguistic Actions, States & Processes）语义域，如例（2）中point表示“要点、重点”，例（3）中point表示“看法、观点”。而例（1）中B1水平中国学习者使用的搭配名词point属于K（Entertainment, Sports & Games）语义域，意为“（游戏、比赛的）分数”，这里学习者想表达的意思是“得分”。在BNC中，“得分”的搭配应为get a point，见例（4）。

（1）Use a frisbee to take two shots on each turn, and total of ten turns each knock down each pin, you will **make a point**. (CN\_B1)

（2）I just want to **make a point** on paragraph seven. (BNC)

（3）Colleagues, can I just **make a point** as well while we're progressing. (BNC)

（4）I was delighted to **get a point** after finding ourselves 2-0 down. (BNC)

例（5）—（7）中，A2水平中国学习者使用的make a break搭配，与本族语者语义亦不相同。BNC语料库中make a break的检索行搭配名词break属于A（General & Abstract Terms）语义域，意为“（状况的）改变、中断”，如例（6）。而例（5）中A2水平中国学习者使用的break搭配名词语义域为K（Entertainment, Sports & Games），意为“间歇、休息”。学习者想表达的“休息”语义，在BNC中搭配应为take a break，见例（7）。

（5）Sometimes I go on business trip, usually sit on the desk all the day and never **make a break**, but my job is exciting and rewarding, I love my job. (CN\_A2)

（6）At the time it seemed more truthful to **make a break**; then at least the position was defined. (BNC)

（7）When you **take a break** at lunch-time, have a look around the local shops. (BNC)

这类虚化动词搭配误用现象多出现于中低水平中国学习者群体。本研究认为误用原因有二。一是学习者泛化了虚化动词make的词义，扩大了make的使用范围，并混用了虚化动词make、take、get。这类虚化动词意义泛化（overgeneralization）和混用（mixed usage）现象在缪海燕、孙蓝（2005）和付思婧、陈月红（2022）的研究中也有提及。与其他虚化动词相比，中国学生对make一词的泛化最为严重，这可能与make在常见虚化动词中语义虚化抽象程度最高（付思婧、陈月红 2022）、语义最广泛、搭配能力最强（曾天娇、贾冠杰 2017）相关。二是由于虚化动词搭配名词point、break具有多重语义特征，学习者混淆了这

些搭配名词成分的语义域，所以模糊了 make a point、make a break 搭配的整体含义，将其使用在错误的语境下。值得注意的是，已有研究鲜少关注虚化动词搭配名词成分的多重语义对学习者的虚化动词搭配的影响。词语多义现象亦为影响学习者二语能力的重要因素（苗丽霞 2015），学习者在虚化动词搭配中表现出的非本族语特性，可能与名词成分不同义项的输入顺序和频率相关。如多义词 point 在 BNC 中顺序最先、频率最高的义项为“观点、要点”，而含有这一义项的 make a point 无论作为二语教学中的搭配实例显性输入，还是作为教师课堂指令隐性输入，都会影响学习者习得和产出的优先权，进而影响“（游戏、比赛的）分数”义项的搭配使用。

4.2 两国学习者文本中虚化动词搭配特征及母语迁移效应

4.2.1 两国学习者文本中虚化动词搭配的定量分析

B2 水平西语母语学习者文本中的虚化动词搭配与 B2 水平中国学习者的对比情况见表 5。西语母语学习者的虚化动词搭配频次略低于同水平中国学习者，但种类更丰富，使用更灵活。这一发现支持 Wang（2016）的观点，即母语与英语更为接近的瑞典学习者比中国学习者使用虚化动词搭配形式更丰富，符合母语差异预期的结果。

表 5 两国学习者虚化动词搭配频次和标准化频次（每万词）

语料库	CN_B2	SP_B2
make/take 虚化动词搭配频次 （标准化频次）形符	1,248（52.3）	602（44.1）
make/take 虚化动词搭配频次 （标准化频次）类符	179（7.5）	145（10.6）
类符/形符比	14.3%	24.1%

USAS 语义域对应分析见表 6，维度 1、2 对总惯量的累计贡献率为 78.9%，具有统计学意义，可视化结果如图 2。维度 1 将初级与中高级水平学习者区分于负向与正向。B2 水平西语母语学习者位于维度 1、2 正向区间内，与 B1 水平中国学习者搭配名词语义域的相关性最强，多使用抽象程度较高的 A（如 make contribution, take opportunity/refuge）、I（如 make property/a career/a fortune/arrangement/investment, take a loan）、Q（如 make application/an allegation/an apology/recommendations, take advice）、S（如 make sacrifice, take a miracle/control/responsibility）语义域。Peters（2016）指出名词长度与结构复杂度可作为表征学

习者搭配能力的指标。对比本节西语母语学习者与4.1.1中各水平中国学习者的虚化动词搭配实例，可发现前者倾向使用的搭配名词成分词长最长、结构最复杂，即西语母语学习者的虚化动词搭配能力最强，亦可佐证母语迁移效应假设。

表6 两国学习者虚化动词搭配名词 USAS 语义域的对应分析结果

维度	奇异值	惯量解释比例（%）	惯量累积比例（%）
维度1	0.095	55.5	55.5
维度2	0.040	23.4	78.9

注：X<sup>2</sup> = 334.11，df = 80，p < 0.01

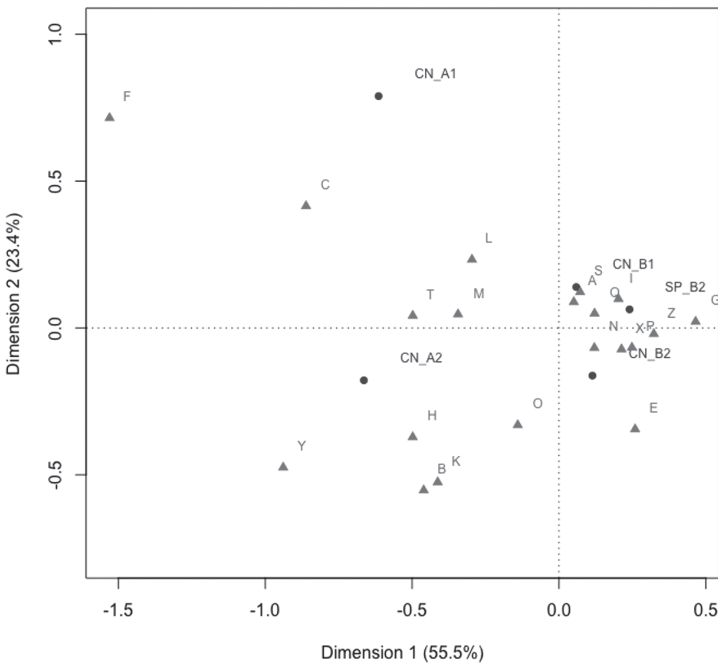


图2 两国学习者语料库虚化动词搭配名词成分语义域的对应分析

4.2.2 母语迁移效应的定性分析

Jarvis (2000) 归纳了三类判断母语迁移的语言证据，指出出现两类证据即可较为准确地支持母语迁移现象，得到研究者的广泛认可 (蔡金亭 2015)。本研究发现了其中两类证据：(1) 汉语/西语背景学习者母语和英语搭配的使用具有一致性；(2) 汉语和西语背景学习者英语搭配的使用具有异质性。

汉语中make的对应动词“做”、take的对应动词“拿”，以及两者在词典中

的其他释义动作，多作为实义动词出现（Wang & Shaw 2008）。而西语中make的对应动词hacer、take的对应动词tomar，均同时具有实义与虚化动词用法。在类型学差异影响下，两国学习者的英语搭配既与母语存在一致性，如正向迁移例（8）—（11）、负向迁移例（12）—（15），又各自具有异质性，如正向迁移例（16）—（19）。

（8）As this house oozes with magic, marble filing willing to **make the sacrifice of** restoration to buy it.（CN\_B2）

（9）He decided to stay and act as a coach, and **made a great contribution** to the China's spaceflight.（CN\_B2）

（10）I believe this is the best way to get stability and **making a good investment** in the long run.（SP\_B2）

（11）If I **take vacations** I feel confidence that she is going to work as if I was there.（SP\_B2）

在例（8）和（9）中，make sacrifice和make contribution在汉语内分别有“做（作）出牺牲”和“做（作）出贡献”的对应搭配，学习者会将汉语中相似的表达方式和搭配规则迁移到英语中。这里依照李晓倩（2015），在《现代汉语大词典》中，“做”和“作”均有“从事某种活动或动作”的语义，我们不区分二者差异。在例（10）和（11）中，make an investment和take vacations分别对应西语中的hacer una inversión和tomar vacaciones。这里英语的虚化动词搭配不仅在意义上，也在结构上与西语一一对应，正迁移效应显著。

（12）He has **made a unbreakable emotional relation** with this amazing man-like species and eventually helped them defeated the army invader.（CN\_B2）

（13）I promise that I will never **make such stupid fault** again.（CN\_B2）

（14）Lisbeth played by Rooney Mara is an outlaw feminist heroine who enjoyed **make justice** by her own hands against the men who principally abuse of women.（SP\_B2）

（15）When I am **making the exercises** and I find unknown words I write them down and later I check into my dictionary.（SP\_B2）

英语中并没有例(12)中 **make a relation** 的虚化动词搭配, 汉语中却有相应说法“搞关系”。中国学习者可能泛化了 **make** 的词义, 并将汉语用法迁移到英语内, 造成搭配错误。例(13)中错误搭配 **make (a) fault** 的正确表达应为 **make a mistake**。虽然 **fault** 与 **mistake** 在汉语内都有“错误”的语义, 但英语内只有 **mistake** 能与 **make** 构成虚化动词搭配, **fault** 则不可。中国学习者受母语影响, 在区分二者的搭配结构时, 不能像本族语者一样自如。在例(14)和(15)中, **make justice** 和 **make the exercises** 也是错误的虚化动词搭配, 是西语中虚化动词搭配 **hacer justiciar** 和 **hacer ejercicio** 的意义与形式的直接对应, 母语成了错误搭配的来源。

(16) He will not **take a decision** on his own. (BNC)

(17) We'll need to take some action and **make a decision**. (BNC)

(18) Sometimes he hesitates to **take a decision**, but once he does, he takes the best decision for the group. (SP\_B2)

(19) I always help my boss **make a decision**, always talk to people and always help people. (CN\_A2)

例(16)和(17)中, 英语内 **decision** 分别可与虚化动词 **take** 和 **make** 构成搭配。本研究发​​现西语母语学习者使用 **take a decision** 的数量较多, 而中国学习者几乎没有使用虚化动词 **take** 与 **decision** 搭配, 只有 **make a decision** 的用法, 如例(18)和(19)。造成差异的原因在于, 汉语对应搭配“做(作)决定”中, “做(作)”与英语中虚化动词 **make** 的语义一致, 因而 **make a decision** 更符合汉语母语迁移的搭配习惯; 而西语对应搭配是 **tomar una decisión**, 虚化动词 **tomar** 对应 **take**, 西语母语者倾向于使用 **take a decision** 作为 **tomar una decisión** 的逐词翻译对应。两国学习者均将母语中的虚化动词搭配结构直接迁移至英语, 造成了搭配习惯差异。

## 5 结语

本研究基于 EFCAMDAT 语料库, 探究不同水平中国英语学习者 **make** 与 **take** 虚化动词搭配特征, 并对照西语母语的英语学习者, 解释母语迁移效应。研究发现: 低水平中国英语学习者使用虚化动词搭配较为单一和具体, 且积极性不高; 中等水平学习者存在过度使用倾向; 高水平学习者使用虚化动词搭配语义丰富, 抽象度最高。定性分析显示搭配名词的多重语义特征易导致中低水平学习者将虚化动词意义泛化和混用, 本研究解释了母语类型学差异对正向和负向迁移的不同机理。



本研究对二语教育教学有一定启示。第二语言习得是重度输入导向 (Dietrich *et al.* 1995) 和输入驱动 (Goldschneider & DeKeyser 2001) 的, 教材是英语习得环境中目的语的主要输入方式之一 (Uchida 2015)。针对低水平中国英语学习者虚化动词搭配使用频率和意愿较低, 中等水平学习者使用过度, 中低水平学习者虚化动词意义泛化和混用等问题, 首先应从教材层面, 为学习者提供准确充足的目的语输入, 如从英语本族语国家最新的大型语料库中选取素材, 呈现最典型的本族语虚化动词搭配用法, 以强化使用规范。其次在课堂教学中, 教师可引导低水平中国学习者正确使用虚化动词搭配, 鼓励其积极性。对于中高水平学习者, 教师可增强其对虚化动词搭配的敏感性, 以进一步提升虚化动词搭配的灵活性和丰富度。同时, 二语教师也有必要拓展词汇搭配的教学模式, 引入数据驱动方法, 指导学习者使用权威英语语料库检索搭配用法, 引导学习者从真实具体的语言现象中发现语言事实, 总结词汇在语境中的使用规律, 掌握语篇中词语的共现特征。针对母语迁移现象, 教师可通过将英语与汉语对比教学等方法, 让学习者感知两种语言在使用上的差异, 最大限度地促进母语正迁移, 排除母语负迁移效应, 使学习者的语言产出更加接近本族语者的使用模式。

#### 参考文献

- ALBA-SALAS J. Light verb constructions in Romance: a syntactic analysis [D]. Cornell: Cornell University, 2002.
- ALLERTON D. Stretched verb constructions in English [M]. London: Routledge, 2002.
- ALTENBERG B, GRANGER S. The grammatical and lexical patterning of MAKE in native and non-native student writing [J]. *Applied Linguistics*, 2001, 22(2): 173-195.
- ALTENBERG E, CAIRNS H. The effects of phonotactic constraints on lexical processing in bilingual and monolingual subjects [J]. *Journal of Verbal Learning and Verbal Behavior*, 1983, 22(2): 174-188.
- CARTER R, MCCARTHY M. Exploring spoken English [M]. Cambridge: Cambridge University Press, 1997.
- DIETRICH R, KLEIN W, NOYAU C. The acquisition of temporality in a second language [M]. Amsterdam: John Benjamins, 1995.
- DU X, AFZAAL M, AL FADDA H. Collocation use in EFL learners' writing across multiple language proficiencies: a corpus-driven study [J]. *Frontiers in Psychology*, 2022, 13: 1-10.
- GOLDSCHNEIDER J, DEKEYSER R. Explaining the “natural order of L2 morpheme acquisition” in English: a meta-analysis of multiple determinants [J]. *Language Learning*, 2001, 51(1): 1-50.
- HÅKANSSON G, PIENEMANN M, SAYEHLI S. Transfer and typological proximity in

- the context of second language processing [J]. *Second Language Research*, 2002, 18(3): 250-273.
- HASSELGREN A. Lexical teddy bears and advanced learners: a study into the ways Norwegian students cope with English vocabulary [J]. *International Journal of Applied Linguistics*, 1994, 4(2): 237-258.
- JARVIS S. Methodological rigor in the study of transfer: identifying L1 influence in them interlanguage lexicon [J]. *Language Learning*, 2000, 50(2): 245-309.
- JUKNEVIČIENĖ R. Collocations with high-frequency verbs in learner English: Lithuanian learners vs. native speakers [J]. *KALBOTYRA*, 2008, 59(3): 119-127.
- KIM H, YANGON R. Effects of verb semantics and proficiency in second language use of constructional knowledge [J]. *The Modern Language Journal*, 2016, 100(3): 716-731.
- KURTEŠ S, SAVILLE N. The English profile programme—an overview [J]. *Research Notes*, 2008, 33: 2-4.
- LEWIS M. Implementing the lexical approach: putting theory into practice [M]. Hove: Language Teaching Publications, 1997.
- LIAO E. An investigation of crosslinguistic transfer in EFL learners' phraseology [D]. San Diego: Alliant International University, 2010.
- NESSELHAUF N. The use of collocations by advanced learners of English and some implications for teaching [J]. *Applied Linguistics*, 2003, 24(2): 223-242.
- NESSELHAUF N. Collocations in a learner corpus [M]. Amsterdam: John Benjamins, 2005.
- PETERS E. The learning burden of collocations: the role of interlexical and intralexical factors [J]. *Language Teaching Research*, 2016, 20(1): 113-138.
- PIAO S, BIANCHI F, DAYRELL C, et al. Development of the multilingual semantic annotation system [R]. Presented at the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Denver, Colorado. 2015.
- PUSTEJOVSKY J. The generative lexicon [M]. Cambridge, MA.: The MIT Press, 1995.
- SINCLAIR J. Trust the text: Language, corpus and discourse [M]. London: Routledge, 2004.
- UCHIDA S. Kihon doushi no korokeshonn nanido sokutei [The measurement of the difficulty level of collocation composed of basic verbs: an investigation of a teaching material corpus based on CEFR levels] [J]. *Gengo Shori Gakkai Nenji Daikai Happyou Ronbunshu*, 2015, 21: 880-883.
- VÁZQUEZ A. The use of MAKE and TAKE by Spanish and Italian learners of English: a corpus study [D]. Stockholm: Stockholm University, 2018.
- WANG Y. The idiom principle and L1 influence: a contrastive learner-corpus study of delexical verb + noun collocations [M]. Amsterdam: John Benjamins, 2016.

- WANG Y, SHAW P. Transfer and universality: collocation use in advanced Chinese and Swedish learner English [J]. ICAME Journal, 2008, 32: 201-232.
- 蔡金亭. 在二语产出中判断母语迁移的比较—归纳方法框架[J]. 解放军外国语学院学报, 2015 (5): 56-65.
- 邓耀臣, 肖德法. 中国大学生英语虚化动词搭配型式研究[J]. 外语与外语教学, 2005 (7): 7-10.
- 方秀才. 基于语料库的中国EFL学习者动词语义虚化用法研究——以BE和HAVE为例[J]. 中国外语, 2015 (5): 68-76.
- 付思婧, 陈月红. 基于语料库的中国学生习得英语虚化动词搭配语的研究[J]. 语言教育, 2022 (1): 52-65.
- 桂诗春. 不确定性判断和中国英语学习者的虚化动词习得[J]. 外语教学与研究, 2007 (1): 3-12.
- 李晓倩. 基于语料库的莎士比亚戏剧汉译本中虚化动词的应用研究——以“做”和“作”为例[J]. 当代外语研究, 2015 (3): 26-30.
- 刘国兵. 中国英语学习者虚化动词TAKE词汇型式研究[J]. 当代外语研究, 2011 (10): 14-19.
- 苗丽霞. 中国学习者英语多义词语义知识的使用和发展[J]. 现代外语, 2015 (2): 217-226.
- 缪海燕, 孙蓝. 非词汇化高频动词搭配的组块效应——一项基于语料库的研究[J]. 解放军外国语学院学报, 2005 (3): 40-43.
- 王立非, 张岩. 大学生英语议论文中高频动词使用的语料库研究[J]. 外语教学与研究, 2007 (2): 110-116.
- 王文宇, 李小撒. 高水平二语学习者在写作任务中的动名搭配使用研究[J]. 解放军外国语学院学报, 2018 (1): 90-98.
- 谢家成. 基于语料库的英语教材虚化动词对比研究——以虚化动词“do”为例[J]. 外语教学理论与实践, 2010 (3): 13-17.
- 曾天娇, 贾冠杰. 英语轻动词与动转名词搭配特征研究——基于“词汇体”分布的实证研究[J]. 外国语, 2017 (6): 20-28.
- 张爱朴. 英语虚化动词组构使用的多维分析[J]. 北京第二外国语学院学报, 2014 (8): 14-19.
- 张莎. 高频动词经验类型及其搭配型式的语料库研究——以make和find为例[J]. 外语学刊, 2011 (3): 36-41.
- 张淑静. 中国英语专业学生MAKE的使用特点调查报告[J]. 解放军外国语学院学报, 2002 (4): 58-63.
- 张文忠, 杨士超. 中国学习者英语语料库中动名搭配错误研究[J]. 解放军外国语学院学报, 2009 (2): 39-44.

朱慧敏, 王俊菊. 二语写作中虚化动词“HAVE”组构型式发展特征研究[J]. 外国语文, 2019 (6): 132-141.

通信地址: 100191 北京市 北京航空航天大学外国语学院

# 中国抗疫推文特征对海外受众行为参与度的影响研究<sup>\*</sup>

对外经济贸易大学 江进林 王佳慧

**提要：**本研究选取我国主流媒体在Twitter上发布的2,500条中国抗疫推文，分析推文的关系行为、信息参与质量以及语言特征对海外受众行为参与度的影响。研究表明，在宏观层面，披露疫苗研发与接种、宣传对外援助与合作、宣传国产疫苗与中药疗效的推文，其海外受众行为参与度更高；在中观层面，推文的信息交互性和多元性与海外受众行为参与度呈现显著正相关关系；在微观层面，现在时态词、副词、二分法思维词、社会指称词等语言特征与海外受众行为参与度显著相关。多元线性回归分析结果显示，三个层面的推文特征能够解释海外受众行为参与度25.2%的变异。本研究可为中国故事的内容设置、对外传播策略和写作风格提供启示。

**关键词：**抗疫推文、特征、海外受众、行为参与度、影响

## 1 引言

如今，社交媒体的蓬勃发展为我国对外讲述中国故事、构建中国的国际话语体系提供了新渠道。尽管我国媒体积极深耕于海外社交平台，但对外传播仍面临“有理说不出”“说了传不开”的窘境（许静、刘煦尧 2017）。在学术界，中国故事的对外传播存在实证研究不足、故事主题单一等问题，对外传播效果的影响因素探究则更为鲜见。基于以上不足，本研究结合语言学和传播学的研究视角，从宏观层面的关系行为、中观层面的信息参与质量、微观层面的语言特征三个维度分析我国抗疫推文的特征，并探究文本特征对海外受众行为参与度的影响，为提升中国故事的对外传播效果提供借鉴。

<sup>\*</sup> 本文系北京市社会科学基金一般项目“中国抗疫故事在海外社交媒体上的传播效果与文本特征影响研究”（20YYB008）的阶段性成果。江进林为本文通讯作者。

作者贡献：

江进林：选题构思、研究方法、讨论结论、初稿撰写、字数占比（60%）、修改润色。

王佳慧：数据收集、数据分析、初稿撰写、字数占比（40%）。

## 2 文献综述

传播过程包括五个基本构成要素：谁（who）、对谁（to whom）、通过什么渠道（in which channel）、说什么（what）以及取得什么效果（with what effect），即“5W模式”（Lasswell 2013）。这个模式简明而清晰，是传播过程模式中的经典。

“取得什么效果”是“5W模式”中的最后一环。近年来，国外越来越关注社交媒体的传播效果，主要研究包括以下两个方面。第一，社交媒体在健康传播、公共生活参与等方面的效果。例如，Cohen *et al.*（2019）通过实验发现Instagram上身材健美的女性图片能够增强女性的愉悦感和对自我外形的欣赏。Boulianne（2019）考察全球多个国家和地区的133个横向研究，探究了社交媒体的使用与公共和政治生活参与之间的关系。研究结果表明，社交媒体能够促进公共生活的参与和政治意见的表达，但其影响力的大小取决于政治环境。第二，社交媒体传播效果的影响因素研究。例如，Nekmat *et al.*（2019）通过实验，发现信源（source）是否可信（如亲朋/机构）对受众参与行为（如点赞）的影响很大。Heiss *et al.*（2019）则通过内容分析，对信息本身和推送者简介中的各种指标进行量化，考察了它们对政治人物参与行为的影响。

国内近年来也开始关注微信和微博等社交媒体的传播效果，研究对象包括科学传播、健康传播、政务传播等（匡文波、武晓立2019；张放等2020），不过对海外社交媒体的研究非常匮乏。此外，评估传播效果也成为对外讲述中国故事的一块短板。相关研究多为思辨性探讨和政策建议，实证研究的话题主要限于脱贫攻坚、“一带一路”、中国文化等，且媒介多为电视纪录片，主要探讨中国故事的对外传播策略与模式（李文英2018；陈旭鑫、李露2021）。关注中国故事在海外社交媒体上传播效果的研究不到5项（徐翔2016；许静、刘煦尧2017；郑越、陆浩2018）。例如，徐翔（2016）考察了中国文化在YouTube上的传播效果与推文的内容丰富度、内容倾向性、内容话题度、内容类型、传播形式、传播持续度、传者影响力、传者扩散力的关系，发现内容话题度与传者影响力对海外传播效果具有明显影响。郑越、陆浩（2018）选取《人民日报》、央视、新华社在Facebook上发布的282篇“一带一路”推文，通过推文形式（文本/图片/视频等）、有无互动等文本特征，以及点赞量、分享量等指标，探究了推文的对外传播效果。这种从文本分析与受众行为交互的角度探讨传播效果的方法值得借鉴。许静、刘煦尧（2017）总结了以央视Facebook账号为代表的新媒体的运营策略和品牌化策略所产生的传播效应，并以网络数据分析为基础进行受众研究，发现推文的阅读和点赞多，评论和转发少，建议媒体与受众加强互动，培养用户群网络社区。

总之，中国故事的对外传播效果存在实证研究不足、故事主题单一、对海外社交媒体的重视不够等问题，且对外传播效果的影响因素探究非常鲜见，分析的推文数量也较少。语料库研究方法能够为新闻传播学研究提供新的思路。例如，



可以围绕某个主题构建较大容量的真实文本库，并从微观层面（如词汇—语法、语篇、语用等）对文本进行标注、检索和统计分析。该方法近年来逐步运用于分析媒体语料，如对涉华报道的批评话语分析（Cheng & Lam 2013；孙有中、江璐 2017）。鉴于中国故事对外传播研究的不足，本研究结合传播学与语言学领域的研究视角，从多个维度挖掘中国抗疫推文的文本特征，并探究其对海外传播效果的影响。

3 推文特征的多维分析框架

本研究提出推文特征的多维分析框架，包括宏观层面的关系行为、中观层面的信息参与质量、微观层面的语言特征。

3.1 宏观层面的关系行为分析

关系行为起源于言语行为理论，是媒体语境中的言语行为，指任何以言语建立并维持关系可持续发展的言语行为（Feng & Ren 2019）。Li & Wu（2018）在研究国际化品牌如何利用语言资源维持与受众的关系时，将推文按信息类型分为宣传与建构人际关系两类。宣传类包括推广企业产品与服务等，人际关系建构类则以问候、指令、表达、分享为主。本文参考Li & Wu（2018）的分类，构建了中国抗疫推文的关系行为标注方案，并对推文进行标注，详见表1。

表1 中国抗疫推文的关系行为标注方案

关系行为	描述	解释	示例
披露类	披露疫情形势	披露病毒变异、感染与恢复病例等信息	The Chinese mainland on Sunday reported 164 locally transmitted confirmed COVID-19 cases, including 81 in Sichuan.
	披露疫苗研发与接种	披露国产疫苗研发、国民疫苗接种信息	More than 161.12 million doses of #COVID19 vaccines had been administered across China as of Friday, the National Health said.
宣传类	宣传对外援助与合作	宣传中国在疫苗、医疗物资、抗疫经验等方面对他国提供援助或进行合作	From sharing experience on pandemic control to carrying out vaccine cooperation; promoting fair distribution, China’s support for other countries; efforts to safeguard people’s right to life; health has been widely recognized.

（待续）



(续表)

关系行为	描述	解释	示例
宣传类	宣传疫情防控举措与成果	宣传中国的疫情防控举措与效果	Expert: Dynamic zero-COVID strategy key to curbing Omicron spread in China for the time being.
	宣传国产疫苗与中药疗效	宣传国产疫苗与中药的科学性和安全性等优势	Chinese COVID-19 vaccines have performed very well in preventing deaths and severe cases, and reducing hospitalization rates.
	表达态度	表达中国对抗疫防疫的立场、态度与观点	China stands ready to strengthen cooperation with Portugal and other European countries on vaccines and drugs, and jointly oppose “vaccine nationalism” and “political virus”.
表达类	表达建议	表达中国对抗疫防疫的建议、呼吁等	China calls on capable countries to provide #COVID19 vaccines to developing countries and promote the fair distribution and use of vaccines globally, said Chinese Foreign Ministry spokesperson on Tuesday.
	表达情感	表达中国对抗疫防疫的情感, 如希望、乐观、鼓舞、感谢、关心等	May the year 2022 mark an end to the COVID-19 pandemic.
需求类	寻求互动	鼓励受众与推文进行互动, 如浏览、点赞、分享和评论等	Fighting the pandemic, sharing technological know-how, reducing poverty... Find out what African students in China have to say about China-Africa ties. #GLOBALink.
分享类	分享疫后生活	分享后疫情时代中国百姓的生活样貌	With the effective control of COVID-19, summer night economy comes back to life in many parts of China. Late into the night, an ancient street in Wuhu City is bustling.
	科普防疫知识	普及病毒变体、疫苗接种等相关知识和应对技巧	#Omicron is not a ‘big flu’, and it can bite, especially for people without strong immunity, Zhang Wenhong, leader of the expert team.
	分享抗疫故事	分享百姓抗疫故事	COVID-19 Diary: Over 200 residents form emergency volunteer team to help community fight against epidemic in Beijing.

3.2 中观层面的信息参与质量分析

传播内容的信息参与质量会塑造媒体用户的感知体验和态度，从而影响其传播效果（Zhang & Du 2020）。其中，传播内容的交互性提示和媒体丰富度能有效促进用户的点赞、评论以及转发行为（Moran *et al.* 2020）。交互性提示指促使用户进行回应或行为反馈的文本信息，不同层级的交互性提示需要用户投入不同程度的认知努力和时间精力。媒体丰富度也称为多元性（variety），包括推文内容的广度（内容产生的感官刺激）和深度（内容的呈现形式）（Zhang & Du 2020）。本研究采用Moran *et al.*（2020）关于信息参与质量的分析框架，从交互性和多元性两方面对中国抗疫推文进行编码，详见表2。

表2 中国抗疫推文的信息参与质量编码方案

类型	描述	赋值
交互性	零级：不包含与受众互动的信息提示	0
	低级：号召受众点击或点赞，如“点击这里”“了解更多”	1
	中级：鼓励受众参与讨论，如“提出问题”“告诉我们您的想法”	2
	高级：同时出现低级与中级交互性的信息提示	3
多元性	零级：纯文本信息	0
	低级：包含超链接的信息	1
	中级：包含图片的信息	2
	高级：包含视频的信息	3

3.3 微观层面的语言特征分析

微观层面的语言特征主要指词汇—语法特征。本研究使用LIWC工具提取推文的语言特征，该工具由Boyd等人开发，可基于词量丰富的核心词典，对推文、博文等社交媒体文本进行词语计量分析（Boyd *et al.* 2022）。近年来，诸多学者使用LIWC对Reddit、Facebook、Twitter上的短文本进行特征提取和量化分析（Vaičiukynaitė & Gatautis 2018；Deng *et al.* 2021）。

4 研究设计

4.1 研究问题

本研究旨在回答以下问题：

(1) 我国抗疫推文表达了哪些关系行为? 关系行为是否影响海外受众的行为参与度?

(2) 我国抗疫推文的信息参与质量如何? 信息参与质量是否影响海外受众的行为参与度?

(3) 我国抗疫推文的语言表达有何特征? 语言特征是否影响海外受众的行为参与度?

(4) 以上三个层面的文本特征对海外受众行为参与度是否具有显著预测力? 如有, 预测力有多大?

## 4.2 语料收集

本研究选取中央电视台、《人民日报》《中国日报》、新华社和中国网 (China.org.cn) 五家媒体的 Twitter 账号, 以 COVID-19、coronavirus、pandemic、variants 和 vaccine 等为关键词, 检索它们在 2021 年 4 月 1 日至 2022 年 6 月 1 日期间发布的推文, 得到 6,386 条。然后通过人工筛选, 剔除与中国抗疫无关的推文, 得到 4,195 条。最后随机选取 2,500 条, 建成中国抗疫推文语料库, 其形符数为 86,644 词, 类符数为 10,733 词。

## 4.3 研究工具

本研究使用 LIWC 工具提取推文的语言特征。LIWC-22 共提取 109 项指标, 包括 4 项叙事风格指标 (分析性思维、影响力、权威性、情绪基调)、3 项一般描述性指标 (每句平均词数、超过 7 字母词数、核心词汇数)、20 项语言学维度指标、36 项心理维度指标、3 项文化类别指标、5 项生活类别指标、8 项生理类别指标、6 项状态类别指标、4 项动机类别指标、10 项感知类别指标、6 项标点符号指标和 4 项对话类别指标 (Boyd *et al.* 2022)。其中, 语言学维度指标包括功能词、代词、副词、连词等, 心理维度又进一步划分为情感过程、社会过程、认知过程和驱动 4 个子范畴。

## 4.4 数据分析

本研究考察三个层面的文本特征对海外受众参与度的影响。用户参与度体现受众对内容的参与意愿与行为反馈, 是预测社交媒体传播效果的一个关键因素 (Leek *et al.* 2019)。用户参与度包括认知、情感和行为三个维度 (Brodie *et al.* 2011), 其中行为参与度表现为点赞、评论与转发行为, 受到更多重视 (Dolan *et al.* 2019)。本研究参考 Bonsón & Ratkai (2013) 关于企业利益相关者的社交媒体参与度测量公式<sup>1</sup>, 计算中国抗疫推文的海外受众行为参与度。由于中国主流媒体 Twitter 账号的关注人数众多, 为方便统计, 本研究将该公式中的  $10^3$  修改为  $10^8$ 。

本研究使用单因素方差分析，考察中国抗疫推文的关系行为与海外受众行为参与度的关系；采用Spearman相关分析，从交互性和多元性两方面考察推文信息参与质量与海外受众参与度的关系；采用Pearson相关分析，考察推文的语言特征与海外受众参与度的关系。最后对三个层面的文本特征与海外受众参与度进行多元线性回归分析，探讨推文文本特征对海外受众行为参与度的预测力。

5 结果与讨论

5.1 推文关系行为与海外受众行为参与度的关系

如表3所示，中国抗疫推文的关系行为包括披露、宣传、需求、表达与分享5大类、12个主题。其中，披露类包括披露疫情形势、疫苗研发与接种；宣传类包括宣传对外援助与合作、疫情防控举措与成果、国产疫苗与中药疗效；需求类包括寻求互动；表达类包括表达情感、态度与建议；分享类包括分享疫后生活、科普防疫知识、分享抗疫故事，可见中国抗疫推文的主题比较多元。其中，宣传、披露类推文明显多于其他推文，而需求类与分享类相对较少。并且，宣传对外援助与合作、国产疫苗与中药疗效推文的受众参与度（M = 47.84, 47.04）最高，而分享抗疫故事、科普防疫知识推文的受众参与度（M = 22.69, 29.16）最低。

表3 中国抗疫推文的关系行为及其受众行为参与度

关系行为	描述	推文		受众行为参与度	
		N	百分比	均值	标准差
披露类	披露疫情形势	57	2.28%	31.04	30.756
	披露疫苗研发与接种	571	22.84%	42.79	25.601
宣传类	宣传对外援助与合作	986	39.44%	47.84	42.206
	宣传疫情防控举措与成果	238	9.52%	36.56	56.740
	宣传国产疫苗与中药疗效	162	6.48%	47.04	31.699
需求类	寻求互动	26	1.04%	42.31	29.038
表达类	表达情感	39	1.56%	40.44	49.716
	表达态度	196	7.84%	35.64	33.533
	表达建议	77	3.08%	36.52	31.833
分享类	分享疫后生活	90	3.60%	31.79	39.822
	分享抗疫故事	13	0.52%	22.69	14.343
	科普防疫知识	45	1.80%	29.16	21.325

单因素方差分析结果显示,不同关系行为推文的海外受众行为参与度具有显著差异( $F = 5.096, df = 11, P = .000$ )。Games-Howell事后检验结果(见表4<sup>2</sup>)显示,披露疫苗研发与接种、宣传对外援助与合作、宣传国产疫苗与中药疗效的推文,其受众参与度显著高于分享抗疫故事、科普防疫知识、分享疫后生活、表达态度、披露疫情形势的推文。

表4 事后多重比较结果

关系行为(I)	关系行为(J)	均值差(I-J)	显著性
披露疫苗研发与接种	分享抗疫故事	20.094**	.009
	科普防疫知识	13.631**	.008
宣传对外援助与合作	披露疫情形势	16.807*	.010
	表达态度	12.204**	.001
	分享疫后生活	16.053*	.020
	分享抗疫故事	25.149**	.001
	科普防疫知识	18.686**	.000
宣传国产疫苗与中药疗效	披露疫情形势	16.002*	.049
	表达态度	11.399*	.049
	分享抗疫故事	24.345**	.001
	科普防疫知识	17.881**	.001

5.2 信息参与质量与海外受众行为参与度的关系

本研究从交互性与多元性两方面考察推文的信息参与质量。表5显示,中国抗疫推文的交互性分布不均( $\chi^2 = 4558.902, df = 3, P = .000$ ),其中83.24%的推文呈零级交互性,即不包含任何促使海外受众作出行为反馈的提示,仅有5.36%的推文具有高级交互性,即号召受众进一步点击、获取更多信息,或鼓励受众进行反馈。中国抗疫推文的多元性同样分布不均( $\chi^2 = 3599.568, df = 3, P = .000$ )。但与交互性不同,76.32%的推文呈中级多元性,即并非单一的文字形式,而是带有图片。

本研究进一步对推文信息参与质量与海外受众行为参与度进行Spearman相关分析。结果表明,推文的交互性与受众行为参与度具有中度相关关系( $\rho = 0.448, P = .000$ ),推文的多元性与受众行为参与度具有显著的弱相关关系( $\rho = 0.143, P = .000$ ),表明推文的互动性语言提示与多元性内容会提升受众的参与意愿。

表5 中国抗疫推文信息参与质量的卡方拟合检验结果

信息参与质量		实际推文数	期望推文数	卡方值	自由度	显著性
交互性 (N = 2500)	零级	2,081	625	4558.902	3	.000
	初级	249	625			
	中级	36	625			
	高级	134	625			
多元性 (N = 2500)	零级	6	625	3599.568	3	.000
	初级	289	625			
	中级	1,908	625			
	高级	297	625			

5.3 语言特征与海外受众行为参与度的关系

本研究对LIWC提取的109个语言特征与海外受众行为参与度进行Pearson相关分析。表6显示<sup>3</sup>，44个语言特征与受众行为参与度显著相关。相对而言，现在时态词、副词、二分法思维词、情绪基调和社会指称词与受众行为参与度的相关度较高（ $r = 0.154, 0.136, 0.123, 0.111, 0.106$ ， $P$ 均为0.000），这是因为现在时态词顺应受众求新求近的心理需求（黄碧蓉 2009），有助于受众直观感知中国抗疫故事；二分法思维词（如everything、always）和情绪基调词有助于增强受众的正面认知（Pezzuti *et al.* 2021）；副词和社会化指称词则有助于营造社区归属感、缩短与受众的交际距离（Zappavigna & Dreyfus 2022），从而提升受众对中国抗疫故事的参与度。

表6 中国抗疫推文的语言特征与受众行为参与度的相关关系

维度	子维度	语言特征	受众行为参与度	
			皮尔逊相关	显著性
描述性维度	叙事风格	分析性思维	-.120**	0.000
		情绪基调	.111**	0.000
		权威性	.094**	0.000
		影响力	.074**	0.000

（待续）

(续表)

维度	子维度	语言特征	受众行为参与度			
			皮尔逊相关	显著性		
语言学维度	语言元素	副词	.136**	0.000		
		代词	.110**	0.000		
		数词	-.103**	0.000		
		人称代词	.097**	0.000		
		功能词	.091**	0.000		
		非人称代词	.072**	0.000		
		动词	.070**	0.000		
		概数词	-.062**	0.002		
		否定词	.048*	0.016		
		二分法思维词	.123**	0.000		
		洞察词	.088**	0.000		
		认知过程	认知历程	.087**	0.000	
		差距词	.087**	0.000		
		因果词	.047*	0.020		
心理学维度	情感过程	消极情感词	-.119**	0.000		
		焦虑词	.066**	0.001		
		正向情绪词	.059**	0.003		
		负向情绪词	-.043*	0.031		
		社会指称词	.106**	0.000		
		道德词	.060**	0.003		
		社会过程	社会行为词	.047*	0.018	
		亲社会行为词	.041*	0.040		
		文化类别	种族词	-0.039	0.050	
		生活类别	家庭词	.061**	0.002	
			工作词	.051*	0.010	
		个人化维度	生理类别	疾病词	-.064**	0.001
				健康词	-.062**	0.002
			状态类别	需求词	.044*	0.029
获得词	-.040*			0.044		
动机类别	吸引词		.095**	0.000		

(待续)



(续表)

维度	子维度	语言特征	受众行为参与度	
			皮尔逊相关	显著性
个人化维度	感知类别	现在时态词	.154**	0.000
		过去时态词	-.095**	0.000
		移动词	.094**	0.000
		空间词	.085**	0.000
		将来时态词	.078**	0.000
	对话类别	注意词	.072**	0.000
		网络语言	-.128**	0.000
		句号	-.111**	0.000
	标点符号	其他符号	-.090**	0.000
		问号	.052**	0.010

此外，网络语言词、分析性思维词和消极情绪与受众行为参与度具有显著负相关关系（ $r = -0.128, -0.120, -0.119$ ， $P$ 均为0.000）。其中，网络语言词（如haha、lol）体现出社交媒体语言的新奇性特征，具有较强的娱乐性和趣味性；分析性思维词则与叙述风格密切相关，分析性思维值越低，叙事越平实亲切。负相关系数表明，中国抗疫推文的泛娱乐化倾向越弱、叙事风格越平实亲近（如使用第一、二人称代词）、传达的情绪越积极，海外受众的参与行为可能越多。

5.4 文本特征对海外受众行为参与度的预测作用

本研究以三个层面的文本特征为自变量、海外受众参与度为因变量，进行多元线性回归分析。根据Field（2009）的做法，本研究将连续变量（推文语言特征）和有序变量（推文交互性、多元性）直接用作自变量，将包含多个水平的分类变量（关系行为）设置为哑变量（dummy variable）后再用作自变量。

逐步回归分析结果（见表7）显示，回归模型调整后的 $R^2$ 为0.252，可见文本交互性、二分法思维词、将来时态词、副词、焦虑词以及注意词6个文本特征能够解释受众行为参与度25.2%的变异。各自变量的方差膨胀因子（VIF）均小于10、条件指数（CI）均小于5，表明模型不存在共线性问题。标准化回归方程为：海外受众行为参与度 =  $0.47 \times \text{交互性} + 0.06 \times \text{二分法思维词} + 0.056 \times \text{将来时态词} + 0.049 \times \text{副词} + 0.041 \times \text{焦虑词} + 0.040 \times \text{注意词}$ 。在自变量中，文本交互性特征对受众参与度的预测作用最大（ $\beta = 0.470, P = .000$ ），其次是二分法思维词（ $\beta = 0.060, P = .001$ ），接着是将来时态词（ $\beta = 0.056, P = .001$ ），可见当推文根

据疫情发展作出判断、预测与规划，且推文的交互性越强、态度越明确时，海外受众的参与行为越多。

表7 文本特征与受众行为参与度的回归分析结果

变量	R	修正后 R <sup>2</sup>	F	显著性	标准化系数		t	显著性	共线 性参 数	共线 性诊 断
					标准误	β			VIF	CI
因变量 海外受众行为参与度		.504	.252	141.12	.000					
	(常数项)					.062	13.688	.000		1
	交互性					.069	.470	26.687	.000	1.033 1.477
	二分法思维					.062	.060	3.444	.001	1.026 1.538
	将来时态词					.037	.056	3.181	.001	1.020 1.567
	副词					.030	.049	2.772	.006	1.042 1.655
	焦虑词					.205	.041	2.367	.018	1.007 1.758
自变量	注意词					.079	.040	2.274	.023	1.016 2.091

6 结语

本研究通过分析中国抗疫推文三个层面的文本特征，探究其对海外受众行为参与度的影响。研究表明：在宏观层面，抗疫推文的传播主题多样化，披露疫苗研发与接种、宣传对外援助与合作等关系行为的海外受众行为参与度高于其他推文，而分享、表达类等推文的内容和形式尚待优化。在中观层面，推文的交互性和多元性特征均与海外受众行为参与度显著相关，但约80%的推文缺乏互动性表达。在微观层面，有44个语言特征与海外行为参与度显著相关。回归分析结果表明，文本交互性、二分法思维词、将来时态词、副词、焦虑词以及注意词6个变量能够解释海外受众参与度25.2%的变异。

本研究对于中国故事的对外传播具有重要启示。第一，在宏观内容设置方面，我国媒体应抓住重大事件下蕴藏的宣传机遇，主动设置内容多元的传播主题，积极回应各方关切，使海外受众愿意看、愿意听。第二，在传播策略方面，中国故事的对外传播应考虑海外受众的语言习惯和接受度，优化分享类和表达态度类推文的内容和形式，如激活对话意识，利用互动性语言与海外受众建立联系，通过

优质配图、视频等强化受众的视觉体验，并对受众的评论等参与行为给予及时反馈。第三，中国故事的写作风格应符合相应传播议题的内容，采取平实、亲切的叙事视角，提升受众对传播内容的信任和情感认同，为传播中国故事营造人性化、有温度的话语交际氛围，在潜移默化中激发海外受众对中国故事的关注度和参与度。

本研究也存在一定的不足，如仅考察海外受众的行为参与度，忽略了受众的情感态度。今后的研究可通过收集受众评论，挖掘其情感倾向，以更全面地揭示推文文本特征对海外传播效果的影响。

### 注释

- 1 社交媒体参与度 = (点赞数 ÷ 推文发布数 ÷ 账号关注人数) × 10<sup>3</sup> + (评论数 ÷ 推文发布数 ÷ 账号关注人数) × 10<sup>3</sup> + (转发数 ÷ 推文发布数 ÷ 账号关注人数) × 10<sup>3</sup>
- 2 由于数据繁杂，表4仅呈现具有显著差异的比较结果。本文的\*表示在.05的水平上具有显著性，\*\*表示在.01的水平上具有显著性。
- 3 由于数据繁杂，表6仅呈现具有显著相关关系的数据结果。

### 参考文献

- BONSÓN E, RATKAI M. A set of metrics to assess stakeholder engagement and social legitimacy on a corporate Facebook page [J]. *Online Information Review*, 2013, 37(5): 787-803.
- BOULIANNE S. Revolution in the making? Social media effects across the globe [J]. *Information Communication & Society*, 2019, 22(1): 39-54.
- BOYD R, ASHOKKUMAR A, SERAJ S, et al. The development and psychometric properties of LIWC-22 [R]. Austin: University of Texas at Austin, 2022. <https://www.liwc.app>.
- BRODIE R, HOLLEBEEK L, JURIC B, et al. Customer engagement: conceptual domain, fundamental propositions, and implications for research [J]. *Journal of Service Research*, 2011, 14(3): 252-271.
- CHENG W, LAM P. Western perceptions of Hong Kong ten years on: a corpus-driven critical discourse study [J]. *Applied Linguistics*, 2013, 34(2): 173-190.
- COHEN R, FARDOULY J, NEWTON-JOHN T, et al. BoPo on Instagram: an experimental investigation of the effects of viewing body positive content on young women's mood and body image [J]. *New Media & Society*, 2019, 21(7): 1546-1564.
- DENG Q, WANG Y, ROD M, et al. Speak to head and heart: the effects of linguistic features on B2B brand engagement on social media [J]. *Industrial Marketing Management*,

- 2021, 99: 1-15.
- DOLAN R, CONDUIT J, FRETHEY-BENTHAM C, et al. Social media engagement behavior: a framework for engaging customers through social media content [J]. *European Journal of Marketing*, 2019, 53(10): 2213-2243.
- FENG W, REN W. “This is the destiny, darling”: relational acts in Chinese management responses to online consumer reviews [J]. *Discourse, Context & Media*, 2019, 28: 52-59.
- FIELD A. *Discovering statistics using SPSS (3rd edition)* [M]. Los Angeles, CA: Sage Publications, 2009.
- HEISS R, SCHMUCK D, MATTHES J. What drives interaction in political actors’ Facebook posts? Profile and content predictors of user engagement and political actors’ reactions [J]. *Information, Communication & Society*, 2019, 22(10): 1497-1513.
- LASSWELL H. *The structure and function of social communication* [M]. Beijing: Communication University of China Press, 2013.
- LEEK S, HOUGHTON D, CANNING L. Twitter and behavioral engagement in the healthcare sector: an examination of product and service companies [J]. *Industrial Marketing Management*, 2019, 81: 115-129.
- LI C, WU D. Facework by global brands across Twitter and Weibo [J]. *Discourse, Context & Media*, 2018, 26: 32-42.
- MORAN G, MUZELLEC L, JOHNSON D. Message content features and social media engagement: evidence from the media industry [J]. *Journal of Product & Brand Management*, 2020, 29(5): 533-545.
- NEKMAT E, GOWER K, ZHOU S, et al. Connective-collective action on social media: moderated mediation of cognitive elaboration and perceived source credibility on personalness of source [J]. *Communication Research*, 2019, 46(1): 62-87.
- PEZZUTI T, LEONHARDT J, WARREN C. Certainty in language increases consumer engagement on social media [J]. *Journal of Interactive Marketing*, 2021, 53(1): 32-46.
- VAIČIUKYNAITĖ E, GATAUTIS R. How hotel companies can foster customer sociability behaviour on Facebook [J]. *Journal of Business Economics and Management*, 2018, 19(4): 630-647.
- ZAPPAVIGNA M, DREYFUS S. “In these pandemic times”: the role of temporal meanings in ambient affiliation about COVID-19 on Twitter [J]. *Discourse, Context & Media*, 2022, 47: 1-11.
- ZHANG J, DU M. Utilization and effectiveness of social media message strategy: how B2B brands differ from B2C brands [J]. *Journal of Business & Industrial Marketing*, 2020, 35(4): 721-740.

- 陈旭鑫, 李露. 讲好中国战“疫”故事凸显制度优势和治理效能——《同心战“疫”》叙事分析[J]. 电视研究, 2021 (1): 72-74.
- 黄碧蓉. 新闻标题现在时态生成动因及构建机制释解[J]. 外语教学, 2009 (5): 33-36.
- 匡文波, 武晓立. 基于微信公众号的健康传播效果评价指标体系研究[J]. 国际新闻界, 2019 (1): 153-176.
- 李文英. “讲好中国故事”与“一带一路”题材纪录片叙事探析[J]. 电视研究, 2018 (4): 89-91.
- 孙有中, 江璐. 澳大利亚主流媒体中的“一带一路”[J]. 现代传播, 2017 (4): 37-41.
- 徐翔. 中国文化在视频自媒体的传播效果及其影响因素分析——基于YouTube的样本挖掘与实证研究[J]. 北京邮电大学学报(社会科学版), 2016 (5): 1-7.
- 许静, 刘煦尧. 以海外社交媒体策略传播讲好中国故事[J]. 中国出版, 2017 (18): 7-11.
- 张放, 杨颖, 吴林蔚. 政务微信“软文”化传播效果的实验研究[J]. 新闻界, 2020 (1): 59-73.
- 郑越, 陆浩. 讲好海外社交媒体上的中国故事——以我国三家主流媒体“一带一路”Facebook报道为例[J]. 电视研究, 2018 (9): 7-9.

通信地址: 100029 北京市 对外经济贸易大学英语学院

# 多模态话语分析视阈下的新闻价值研究<sup>\*</sup>

集美大学 韩存新 上海大学 赵宇飞

**提要：**近年来，新闻价值的话语建构观日益受到学界关注。本文基于自建《中国日报》新冠疫情新闻报道多模态语料库，运用语料库辅助的多模态话语分析法，参照新闻价值话语分析框架，对新冠疫情新闻报道中包含的新闻价值及其建构的语言和视觉手段进行了分析。研究表明，中国在重大突发公共事件的对外传播中十分重视新闻话语主体的权威性和客观性，以及话语文本的真实性和科学性，同时这些新闻价值的构建除了向世界展现了一个积极、正面、高效、负责的中国国家形象以外，还增强了中华民族内部抗击疫情的凝聚力和决心。本研究展示的语料库辅助多模态话语分析法对未来的新闻价值话语研究有一定的方法论借鉴意义。

**关键词：**新闻价值、新冠疫情、多模态、语料库、话语分析

## 1 引言

2020年，新冠肺炎疫情突然暴发并在全球范围内迅速蔓延。面对这一特别重大突发公共卫生事件，中国第一时间采取措施进行疫情防控。同时，中国一贯秉持人类命运共同体理念，及时通过主流媒体向国际社会披露疫情相关信息。新闻价值指的是“事实内含的能够在多大程度上引起受众普遍关注的素质（要素）”（童兵、陈绚 2014：8）。一直以来，新闻价值都是新闻传播领域的重要研究对象。但不可否认，新闻信息传递的主要载体是语言，而“现有研究忽略了新闻价值研究的语言学视角，无法解释新闻价值是如何在新闻话语中建构的问题”（郇昌鹏 2016：45）。澳大利亚两位学者Bednarek和Caple于2012年提出新闻价值的话语观认为，新闻价值是记者通过新闻话语重构的价值（Bednarek & Caple 2012；Caple & Bednarek 2013）。相关研究不仅归纳了负面性、时效性、接近性和精英性等11种新闻价值，还构建了相应的语言和视觉资源目录，从而搭建起新闻价值

<sup>\*</sup> 本文系福建省社会科学基金项目“多模态视阈下中国商务新闻价值话语研究”（FJ2022BF023）的阶段性成果。韩存新为本文通讯作者。

作者贡献：

韩存新：选题构思、研究方法、讨论结论、初稿撰写、字数占比（60%）、修改润色。

赵宇飞：数据收集、数据分析、初稿撰写、字数占比（40%）。



分析的多模态话语分析框架 (Bednarek & Caple 2017)。“*China Daily*是我国唯一有效进入西方主流社会、国外媒体转载率最高的中国报纸,具有一定的权威性与公信力”(邓斯佳 2015: 113)。本研究旨在运用语料库辅助的多模态话语分析法,参照Bednarek和Caple提出的新闻价值话语分析框架,调查《中国日报》关于新冠肺炎疫情的新闻报道中具体建构了哪些新闻价值以及建构的手段和目的。

## 2 文献回顾

### 2.1 新闻价值

国内外新闻价值的研究基本上都是围绕新闻价值的定义、要素、选择、评价、演变等展开理论探讨和分析。特别是关于新闻价值定义的争论持续至今。虽说国内对新闻价值的定义众多,但比较有代表性的观点主要有5种:素质说、标准说、功能说、效果说和关系说(李春邦 1983;陈韵昭、吴文虎 1984;何光先 1985;雷跃捷 1992;刘建明 2002)。国际上对新闻价值的认识大致可以分为3种:社会观、认知观和物质观(Galtung & Ruge 1965; Golding & Elliot 1979; Palmer 2000)。之所以产生不同的看法,源于人们对新闻价值概念的不同理解和认识(童兵、陈绚 2014)。新闻价值的构成要素是过去新闻价值研究讨论最多的问题(杨保军 2003)。“被美国新闻界公认的新闻价值五要素是时效性、重要性、接近性、显著性、趣味性”(童兵、陈绚 2014: 8)。受其启发,国内学者也提出了若干新闻要素,比如:时效性、新鲜感、时宜性等(王新友 1980)。美国传播学者Shoemaker认为,媒介内容并不完全是对社会的真实反映,媒介在积极建构事实(包括扭曲事实)(转引自陈力丹 2000)。这与Bednarek和Caple的新闻价值话语建构观是一致的。此外,图像也是传递新闻信息的重要载体。新闻传播学中的新闻价值研究往往重文轻图(Caple & Bednarek 2013)。

总体上,国内新闻价值研究多集中在新闻传播领域,少有语言学视角的跨界思考,而且研究内容多聚焦于对新闻价值的定义、要素、运用、效果的理论探讨上,缺乏对新闻价值建构的语言手段进行实证研究。绝大多数新闻价值研究仅讨论文本,忽视了图像在新闻价值创造中的参与、补充作用,导致了现有新闻价值研究在模态上的失衡和不足。

### 2.2 新冠疫情相关新闻报道的研究

来自不同领域(如新闻传播、语言学、文学)的研究人员对新冠肺炎疫情相关的新闻报道进行了多维剖析。其中一些偏向研究媒体中的人物形象。例如,报道中出现的女性工作者形象(赵雅馨 2020)。有的研究则从新闻传播角度出发,探讨哪些策略可以让新闻报道客观有效(罗强、张瀚祥 2020)。有的则分析疫情



相关新闻报道的特征和模式（Wen *et al.* 2021；许茜 2020）。仅少数研究属于语言学领域，涉及话语分析、认知语言学、评价理论等几种研究视角。例如，赵心宇（2021）从文本实践、话语实践和社会实践三个向度对《人民日报》中的疫情相关新闻报道进行了研究，发现《人民日报》作为主流媒体承担着信息传播、舆论引导和安抚民众的重要责任。Zheng（2020）则尝试运用概念隐喻理论来分析疫情相关新闻报道的语言框架。得出的结论是，新闻报道有助于引导舆论、吸引国际受众并为中国赢得国际认可。Fan（2020）运用Martin的评价理论分析了《中国日报》的十篇新闻报道，揭示了隐藏在客观话语中的意识形态。

综上，虽然从新闻传播领域到语言学领域，已有大量关于新冠肺炎疫情的新闻报道研究，但从新闻价值理论视角出发的研究尚未见到。本研究基于自建语料库对《中国日报》中与新冠疫情相关的新闻报道进行多模态分析，一方面揭示新闻报道中建构的新闻价值及其建构方式和目的，另一方面也揭示新闻文本与图像在新闻价值构建方面的模态组合关系。

### 3 研究设计

#### 3.1 语料库及分析工具

本文收集了《中国日报》官网2020年1月至2月有关新冠肺炎疫情的新闻报道及插图。首先，在高级搜索中输入关键词novel coronavirus以及它的一些语言变体如Corona-virus、Covid-19、COVID-19等，将检索日期限定为“2020-01-20至2020-02-20”，选择重复数据删除。其次，整理和剔除无效的或不相关的新闻，确保筛选出的新闻均为中国日报最初发表。最终整理出了126篇有效新闻，涉及的频道包括社论、世界观、企业、教育、评论等。接下来，将采集到的新闻文本按日期命名，整理为纯文本文件，以便通过语料库软件进行检索和分析。建成的新冠疫情新闻报道多模态语料库包含语言文本共计6,115类符，57,499形符。新闻图像则按照报道时间命名排序后专门构建新闻图像语料库，共收集新闻图像101张，其中摄影图像83张，卡通漫画图像14张，统计图1张，其他图像3张。将数据采集的时间跨度设置为“2020-01-20至2020-02-20”，是因为2020年1月20日是一个重要转折点，这一天国家领导人就疫情作出重大批示，钟南山院士在央视《新闻1+1》节目发声，证实了疫情可以人传人。从这一天开始，媒体对疫情的报道也进入了白热化阶段（栾轶玫、张雅琦 2020）。直到2020年2月20日，疫情蔓延势头初步得到遏制，相关新闻报道数量才趋于稳定。本研究使用的语料库检索和分析工具为AntConc 3.5.9w。

#### 3.2 研究问题

本文旨在回答以下三个问题：

- (1)《中国日报》关于新冠肺炎疫情的新闻报道建构了哪些新闻价值?
- (2) 这些新闻价值是通过哪些语言手段和视觉手段建构的?
- (3) 建构这些新闻价值的目的和功能是什么?

### 3.3 研究步骤

本研究中的文本分析主要有以下三个步骤。首先,分别使用AntConc 3.5.9w的Wordlist功能和Keywords功能生成语料库的高频词表和关键词表。通过分析两表中的重合词汇识别出语料库的主题关键词及其反映的主要新闻价值。然后,使用语料库软件特有的KWIC功能对搜索词的索引行进行分析。最后,通过文件查看浏览功能查看搜索词出现的原文,进一步扩展语境来分析新闻价值建构的语言手段和功能。图像分析则采取人工分析的方式,定性分析每张新闻图像中包含的新闻价值并用Excel软件记录和统计。

## 4 结果与分析

### 4.1 总体趋势

借助AntConc 3.5.9w的词表分析,获得一个排除了虚词的新冠疫情语料库高频词表。前30位高频词如表1所示。

表1 新冠疫情报道文本语料库前30高频实词

排序	词语	频次	排序	词语	频次
1	China	655	16	virus	267
2	that	514	17	Chinese	264
3	is	513	18	novel	247
4	said	485	19	health	221
5	's	426	20	Wuhan	196
6	has	386	21	cases	180
7	have	340	22	medical	159
8	coronavirus	335	23	control	150
9	it	315	24	Hubei	133
10	epidemic	304	25	measures	132
11	will	304	26	province	131
12	Be	303	27	new	130
13	outbreak	300	28	country	128
14	people	287	29	government	118
15	are	271	30	spread	106

不难发现,表1中的主导词类为名词,有China、coronavirus、epidemic、outbreak、people、virus、health、Wuhan、cases、Hubei、measures、province、country和government。其次是动词并且大部分是助动词和BE动词。形容词和代词较少。根据outbreak、coronavirus、epidemic、virus、novel等高频词来判断,新闻主题无疑是新冠病毒。从关注区域来看,《中国日报》的新闻报道主要聚焦于中国本土,尤其是湖北武汉,这可以通过排名最高的词China以及其他高频词Chinese、Hubei、Wuhan看出。《中国日报》在抗疫报道中始终强调人民的重要性,这可以从高频词people中窥见一斑。此外,从will一词可以看出《中国日报》尽量采用积极以及充满希望的表达,以期向全世界人民传递战胜新型冠状病毒的决心和信心。

AntConc 3.5.9w的Keywords功能可以将一个语料库与另一个参照语料库进行对比。通过比较两个词表中的词语频数自动生成一个关键词列表,其中包含了出现频率比预期高的所有词语(Baker 2006)。关键词的显著性通过软件自动计算出的关键性(Keyness)来衡量。本研究用来计算关键性的统计手段是对数似然率(Loglikelihood,简称LL),当 $LL > 6.63$ 时,说明该词语的关键性具有统计学上的显著意义( $P < 0.01$ )。关键词词表倾向于呈现3种类型的词语,即专有名词、主题词和语法词(Scott 1999)。其中主题词通常是实词,反映文本的主题特征。

表2 新冠疫情报道语料库中的前30位实词关键词

排序	关键词	频次	关键性	排序	关键词	频次	关键性
1	China	654	3833.657	16	medical	159	458.866
2	coronavirus	329	1928.552	17	Tuesday	74	433.778
3	epidemic	302	1682.894	18	confirmed	98	408.025
4	outbreak	299	1559.827	19	WHO	68	398.606
5	Chinese	264	1547.531	20	control	150	370.891
6	virus	261	1148.099	21	prevention	88	361.231
7	Wuhan	194	1137.201	22	spread	106	337.813
8	novel	247	901.053	23	Monday	57	334.126
9	Hubei	133	779.627	24	infected	65	315.699
10	province	131	668.495	25	fight	90	314.406
11	cases	180	579.660	26	national	53	310.679
12	measures	132	548.603	27	Beijing	52	304.817
13	Health	86	504.120	28	Wang	49	287.231
14	pneumonia	91	489.071	29	commission	46	269.646
15	said	485	483.325	30	He	46	269.646

比较表1和表2可以发现，两表重合的词汇有15个，分别是China、coronavirus、epidemic、Wuhan、virus、Chinese、novel、Hubei、health、outbreak、province、medical、measures、control、spread。这些词语可以看作是语料库中的“主题关键词”，也是主要的新闻价值指示词（News Value Pointer）。不难看出，China、Chinese、Wuhan、Hubei这四个词指向的是疫情暴发的区域，建构的是接近性新闻价值。Health通常用于National Health Commission词组当中，指代国家卫健委。Medical常常与workers、teams、staff等词连用表示医疗专家和医护人员。二者共同建构机构和人员的精英性。coronavirus、epidemic、virus、novel、outbreak、spread指代的是病毒的暴发和传播，具有建构负面性新闻价值的潜势。最后，本研究采集的语料都是疫情爆发高峰期的语料，本身就建构了时效性。这一点从高频词中的Tuesday和关键词词表中的Monday这些时间名词上也可以得到印证。综上，数据显示语料库中主要建构的是接近性、精英性、负面性和时效性这四种新闻价值（见表3）。接下来，我们将详细分析这四种新闻价值在新闻报道中的语言建构方式。

表3 新闻价值的类型及定义（Bednarek & Caple 2017: 55）

新闻价值	定义
接近性 (proximity)	该事件被话语建构为地理上或文化上接近（相对于发表地点或目标受众）
精英性 (eliteness)	该事件被话语建构为具有高地位或高名望（包括但不限于所涉及的人、国家或机构）
负面性 (negativity)	该事件被话语建构为负面的，比如，灾难、冲突、争议、犯罪行为
时效性 (timeliness)	该事件被话语建构为相对于发表时间比较及时，比如，新、最近、正在进行、将要发生，或者其他与当前形势或时间相关的（当前的或季节性的）

4.2 新闻报道文本的新闻价值分析

4.2.1 接近性

《中国日报》新冠疫情报道中建构的首要新闻价值为接近性，因为新冠疫情最先在中国武汉暴发，在地理上与目标受众接近。高频专有名词China、Wuhan等也反映出这一特征。通过查看China、Wuhan出现的语境，我们发现大部分索引行都是在描述疫情发生地、流行病学调查、病情溯源、中国中央政府以及武汉地方政府采取的疫情应对措施等（参考以下索引行）。

1. The central Chinese city of Wuhan, where the first case of novel coronavirus
2. Among the newly-infected patients in Wuhan, 66 are male and 70 are female.
3. the virus has been identified outside Wuhan in China, said local health authorities.
4. That the new cases include people outside Wuhan, capital of Hubei province,
5. All three patients had traveled to Wuhan recently. The patient in Shenzhen
6. a regular news conference on Monday that China had informed the WHO,
7. in close communication with them. China has lost no time in sharing the DNA
8. ith the WHO, Geng added. Wuhan has taken measures to manage and control people
9. the coronavirus after visiting relatives in Wuhan on Dec 29. The spread of the coronavirus
10. (2019-nCoV) infection has been confirmed dead in Wuhan, capital of Central China's Hubei

除了直接使用地名来建构接近性以外，我们还发现排在第5位的“’s”也跟接近性有关。据调查，它总是跟地区、机构、组织等名词连用，以所有格的形式来指示接近性（见图1）。例如，Alibaba’s logistics unit、each area’s prevention work、Beijing’s Xicheng District、capital’s down town area和China’s fight。Cotter（1999：168）认为，接近性是新闻的两个最重要的定义性特征之一，因为语言不仅报道新闻，而且还指明它对社区的影响。疫情发生之初，主流媒体的当务之急就是向读者及时通报疫情的相关情况，包括发生地、流行病学调查、受影响地区、当地政府应对措施等。自然而然，国内的一些组织、机构、地区的专有名词就会被大量使用，从而使新闻语篇中的接近性新闻价值异常突出。接近性新闻价值在语篇中的大量构建，反映了地理地域信息的密集发布，从而彰显了中国政府在疫情防控方面高质高效、反应迅速、认真负责的国家形象。

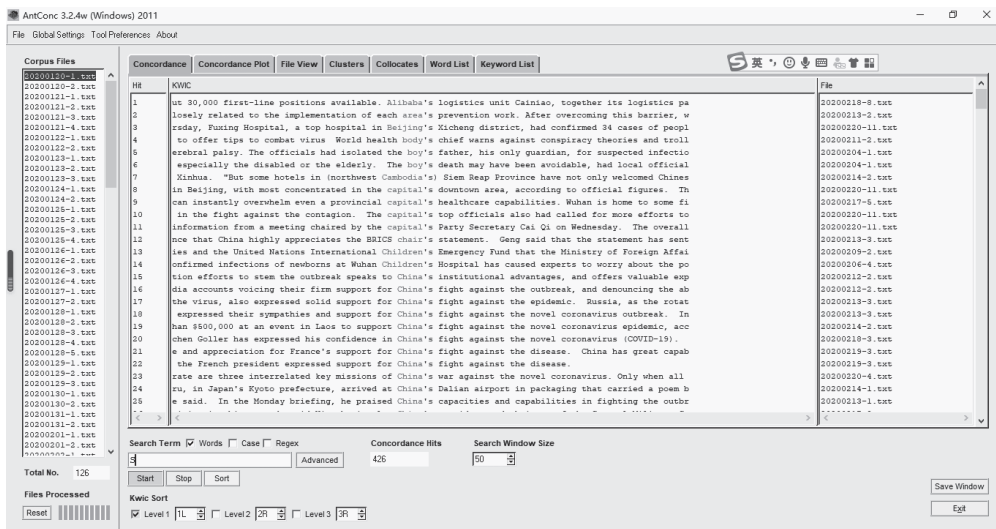


图1 高频词“s”的索引行

#### 4.2.2 精英性

精英性可以通过参与某一新闻事件的著名人士、著名机构或有影响力的国家来建构 (Bednarek & Caple 2014: 156)。高频词that的新闻价值取向不明显, 因此我们提取了that的前10位搭配词, 发现除了功能词the、of、and以外, said排在第4位, 而said本身也是高频词, 说明二者具有紧密的联系, 经常在语境中共现。对于调查新闻价值来说, 引用动词将是一个非常有用的起点 (Potts *et al.* 2015: 157)。作为最常用的引用动词, said的主语通常都是一些人称代词、人物名、机构。为了了解引用的来源, 我们提取了said左/右1距位的搭配词, 发现经常位于said前面用于介绍引用源的词语有he (85)<sup>1</sup>、commission (26)、Xi (14)、Wang (13)、Zhong (11)、Li (10)、it (10)、Tedros (8)、Zhang (6), 而经常位于said后面介绍引用源的词语有Wang (3)、Liu (3)、Zhang (2)、Zhong (2)、Xi (1)。其中第三人称代词he是频数最高的词语。我们通过索引行扩展语境分析, 结合上下文, 对它在具体语境中的指称进行了统计分析 (见表4)。

表4显示he的指称大概可以分为9大类, 其中频数较高的是中国领导人、外国领导人及官员、卫健委官员以及国内外医学人员。结合与said搭配的其他词语的频数, 我们发现总频数较高的分别是卫健委 (35)、习近平 (26)、钟南山 (22) 和Tedros (19)。精英人士通常会被引用来建构精英性这一新闻价值, 如总统、官员、高管和指代领导阶层的专有名词。上述分析表明, 《中国日报》主要是通过引用一些权威机构、政治领袖以及权威医学专家来建构新闻报道中的精英性。卫健委是国家负责组织指导突发公共卫生事件预防控制工作的官方机构。卫健委的政策发布和数据披露代表着国家意志, 行使着国家话语权,



表4 代词he在语境中的指称统计

类别	引用源	频数
中国领导人	习近平(11)、王毅(1)	12
外国领导人及官员	Putin(2)、Mcron(2)、其他(9)	13
世卫官员	Tedros(11)、Ryan(3)	14
卫健委官员	国家卫健委(7)、湖北卫健委(2)	9
中国疾控中心及官员	Zeng Guang	4
国内外医学人员	钟南山(9)、其他(4)	13
境内外商人及组织	Dell、Liu Xiaohu等	9
外国友好人士及组织	Firestein、Ilia等	5
其他人员	Li、Liu Hai、Hu Peng等	6
总计		85

具有高度的权威性和指导性。引用卫健委发布的信息,有利于回应民众了解疫情相关信息的迫切要求,也有利于宣传国家的卫生政策,指导民众的疫情防控实践。引用领导人讲话是新闻媒体惯用的语言手段之一,使得大众媒体拥有了官方性,因此也更容易被大众所接受(石岩 2020: 6)。而引用世卫官员和权威医学专家的话语,则有利于突出中国抗疫为世界做出的贡献,并为中国争取积极的国际评价。总之,精英性新闻价值的建构有助于提升新闻的权威性、可靠性和科学性,让受众能够更好地消除误解,理解政策并增强信心,起到正视听、安民心的舆论引导作用和在国际舞台上提升国家形象的作用。

#### 4.2.3 负面性

负面性一直以来被认为是“基本新闻价值”(Bell 1991: 156)。负面性新闻价值的建构可以采取很多方式。一方面可以使用负面性的评价语言,比如: terrible news、a tragedy、gaffe和wannabe等。另一方面可以提及消极的情绪,比如: distraught、worried、breaking our hearts、shock和disappointment等。另外还可以使用包括“灾难词汇”在内的消极词汇,比如: damaged、killed、deaths、bodies、crime、the IRA、destruction、confusion、offence和road closures等(Bednarek & Caple 2014: 155)。在15个关键主题词当中就有8个与新冠疫情相关,分别是 coronavirus、epidemic、virus、novel、outbreak、measures、control、spread。这些名词本身就参与建构了负面性。进一步调查这些名词在-/+5跨距范围内的搭配词后,我们发现在它们的周围聚集了很多具有战斗意义的搭配词,比如: curbed、tackle、control、struggle、combat、brunt、battle、won、defeat和fight。语境中似乎将疫情



比喻成一场可怕的战争，而我们的目的就是要控制它、解决它、战胜它并赢得这场战争（参考下文索引行）。可见，这些关键主题词的使用语境也充满了负面性。

1. Authorities must put people's health first in resolute fight against coronavirus

2. infectious diseases to effectively fight against the pneumonia caused by the novel

3. China's state-owned enterprises mobilized more resources to combat the outbreak of

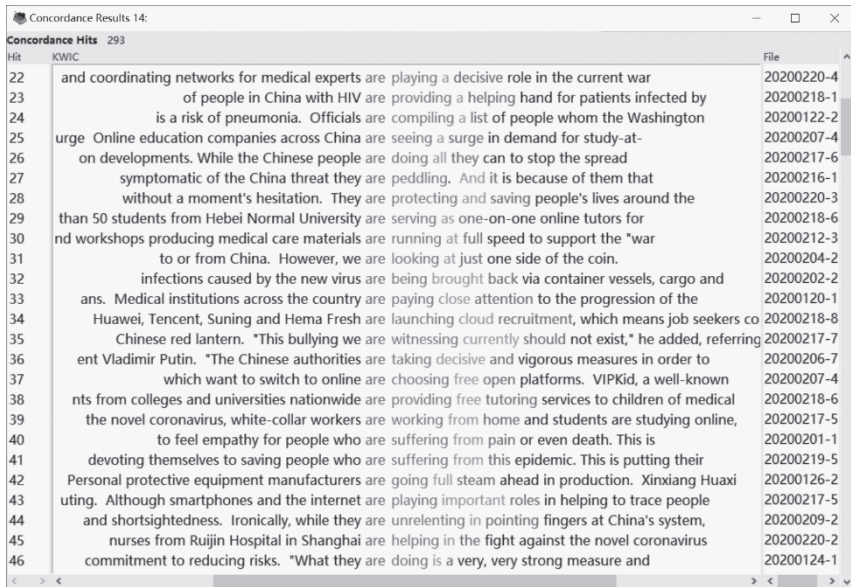
4. 1 billion yuan (\$144.2 million) on Thursday to support the battle against the new virus.

5. China has full capability and confidence in winning the battle against the epidemic.

如果在报道中只是一味地赞美而不批评，那么就会给外国受众一种不客观、不真实的印象（曹碧波 2010），而负面性价值的建构则有效平衡了媒体的报道风格，有助于提升媒体的客观性和权威性。另外，中华文化中有多难兴邦的精神传统，中华民族具有坚强不屈的民族精神。《中国日报》中对负面性新闻价值的凸显不仅有利于提升海内外中华民族同胞在灾难面前自强不息的凝聚力，还能激起中华儿女发奋图强，众志成城战胜疫情的决心。

#### 4.2.4 时效性

时效性既可以通过直接使用时间名词如today、yesterday来建构，也可以通过动词的时态来建构，例如：residents have described the horrific moments（Bednarek & Caple 2014: 155）。表1中的前20位高频词中，Be动词及助动词就占了6个，分别是be、is、are、has、have、will。这些动词能够用于描述时效性所包含的最近、正在进行、将要发生、对现在有影响或季节性的事件。我们以正在进行的事情为例。在搜索栏中输入“\*ing”，然后点击高级，勾选Use Context Words and Horizons，输入“are”“is”“have”等词，就可以检索到be doing和has/have been doing的所有索引行（见图2）。



Hit	KWIC	File
22	and coordinating networks for medical experts are playing a decisive role in the current war	20200220-4
23	of people in China with HIV are providing a helping hand for patients infected by	20200218-1
24	is a risk of pneumonia. Officials are compiling a list of people whom the Washington	20200122-2
25	urge Online education companies across China are seeing a surge in demand for study-at-	20200207-4
26	on developments. While the Chinese people are doing all they can to stop the spread	20200217-6
27	symptomatic of the China threat they are peddling. And it is because of them that	20200216-1
28	without a moment's hesitation. They are protecting and saving people's lives around the	20200220-3
29	than 50 students from Hebei Normal University are serving as one-on-one online tutors for	20200218-6
30	nd workshops producing medical care materials are running at full speed to support the "war	20200212-3
31	to or from China. However, we are looking at just one side of the coin.	20200204-2
32	infections caused by the new virus are being brought back via container vessels, cargo and	20200202-2
33	ans. Medical institutions across the country are paying close attention to the progression of the	20200120-1
34	Huawei, Tencent, Suning and Hema Fresh are launching cloud recruitment, which means job seekers co	20200218-8
35	Chinese red lantern. "This bullying we are witnessing currently should not exist," he added, referring	20200217-7
36	ent Vladimir Putin. "The Chinese authorities are taking decisive and vigorous measures in order to	20200206-7
37	which want to switch to online are choosing free open platforms. VIPKid, a well-known	20200207-4
38	nts from colleges and universities nationwide are providing free tutoring services to children of medical	20200218-6
39	the novel coronavirus, white-collar workers are working from home and students are studying online,	20200217-5
40	to feel empathy for people who are suffering from pain or even death. This is	20200201-1
41	devoting themselves to saving people who are suffering from this epidemic. This is putting their	20200219-5
42	Personal protective equipment manufacturers are going full steam ahead in production. Xinxiang Huaxi	20200126-2
43	uting. Although smartphones and the internet are playing important roles in helping to trace people	20200217-5
44	and shortsightedness. Ironically, while they are unrelenting in pointing fingers at China's system,	20200209-2
45	nurses from Ruijin Hospital in Shanghai are helping in the fight against the novel coronavirus	20200220-2
46	commitment to reducing risks. "What they are doing is a very, very strong measure and	20200124-1

图2 “-ing” 的索引行

(1) Third, the central government is mobilizing more resources from all over China to reinforce Wuhan in the war against the epidemic.

(2) Being nurses we have learned so much seeing news on how healthy professional working staffs, including nurses and doctors, have been devoting themselves to saving people who are suffering from this epidemic.

例（1）中，政府采取的防疫措施用现在进行时来表达，很好地体现了“时效性”。而例（2）则体现了医务工作者一直以来的辛勤付出和不懈努力，有利于激发读者内心的感激和感动。时效性是新闻的核心价值之一。它的大量建构不仅可以彰显新闻中“新”的价值，而且还有利于吸引读者的注意力，将信息及时、快速、准确地传播出去，主动展现积极、正面、高效的中国形象。

4.3 新闻报道图像的新闻价值分析

参照Caple *et al.*（2020）构建的新闻价值分析视觉资源目录并结合视觉语法（Kress & van Leeuwen 1996/2006），我们定性分析了101张新冠疫情新闻图像中蕴含的新闻价值及其语言、视觉建构手段。Kress和van Leeuwen基于系统功能语言学的三大元功能构建了一个图像分析的三维框架，认为图像同时具有再现意义、互动意义和构图意义，分别对应系统功能语言学中的概念功能、人际功能和语篇功能。图像中人物、地点和事件之间的交际关系或概念关系由图像的再现意义来

表征（Kress & van Leeuwen 1996/2006：114-115）。互动意义体现观看者与图像参与者之间的特定关系，主要通过（目光）接触、距离和视角三方面的共同作用，构建出观看者与再现内容之间复杂、微妙的关系。构图意义包括三个方面：信息值、显著性和框架（潘艳艳 2019：79）。

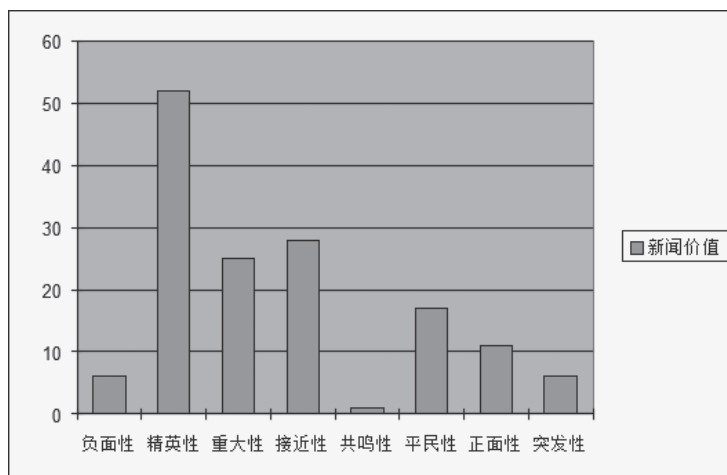


图3 新冠疫情新闻图像中的新闻价值分布

如图3所示，新闻图像中构建的三种主要新闻价值依次为精英性、重大性和接近性（占总量的72%），特别是精英性（36%）新闻价值的分布较广。其次是平民性和正面性新闻价值。突发性、负面性以及共鸣性新闻价值的出现频率较低。83张摄影图像中绝大多数是人物摄影，只有8张是非人物摄影。有关精英人物和机构的图片多达44张。其中出现频率最高的还是包括钟南山院士在内的广大医务人员，共有22张。钟南山院士以及医务人员在图像中的频繁出现能起到防疫教育和展示防疫举措的作用。从图像的再现意义角度来看，这些图像几乎再现了抗疫医生工作的全流程，包括：整装待发、穿戴医服医具、抢救病人、检查病人、检验医学样本、查看医学报告、照顾病人、整理医疗物资、欢送康复病人等等，展现了广大抗疫医务人员积极工作、顽强拼搏的职业精神，以及国难当前、勇于冲锋的英雄精神。

在互动意义层面，大部分图像中的医疗人员与观众之间没有直接的目光接触，属于提供类图像。图像仅提供信息，没有邀请观众与图像中的参与者建立某种关系。从镜头距离来看，呈现钟南山院士时用近景镜头和特写镜头，起到邀请观众与图中角色互动的效果，加深了彼此之间的情感。“一般近景镜头或特写镜头揭示一种亲密的个人关系和特定的情感”（潘艳艳 2019：79）。而呈现医务人员时则主要使用中景镜头，有利于“读者与图中角色进行带适当距离的社会互动”

(Kress & van Leeuwen 1996/2006: 124)。另外,从镜头视角来看,呈现钟南山院士和医务人员的图片摄影多采用平角镜头,显示了观众与图中人物的一种平等的权力关系,更有利于观众对图中人物的认可。

图像的构图意义包括信息值、显著性和框架三个方面。医务人员通常都被放置在图像中信息值最高的中心位置,只有少量图像将病人放在中心位置,但如此却可以体现医生救死扶伤的高尚医德。几乎所有图像都将医务人员放置在前景位置,只有个别送别康复病人的图像将部分医务人员设为背景。前景化的安排凸显了医务人员的地位,增强了人们对医务人员的认同感和信任感。框架选择上,大部分图像都是以医务人员为主体的一元整体框架,但有些图像采用的则是医生病人共存的二元框架,体现了医生与病人之间肝胆相照,共克时艰的关系。

重大性新闻价值主要体现了疫情暴发对人民生活造成的广泛影响。新闻图像主要使用了俯拍的镜头角度和远景的镜头距离。俯拍角度主要用于体现环境的宽广和规模,强调环境、空间和人物在其中的位置,有一种宏观表达的意义(张菁、关玲 2013: 32-33)。比如,呈现大量的私家车滞留,大量的人员离开疫区。同时也显示了政府对疫情的迅速反应,包括派出大量医务人员、军人,生产大量口罩、大量机械同时施工抢建医院(如图1)。图像中出现的大量货币、蔬菜、快递、包裹传递出政府竭力保障民生的信息(如图2)。这些图像主要使用的是水平视角和近景距离,拉近了读者与图中事物的社会距离。接近性新闻价值主要通过呈现城市的交通收费站、建筑、超市等生活设施来建构,主要使用平角和远景镜头,全方位展现疫情下的人民生活状态(如图3)。



图1



图2



图3

新闻图像的主体是摄影图像(82%),只有少量是卡通图像。卡通图像主要构建的是以医学人员为代表的精英性以及以病毒为代表的负面性新闻价值。这与摄影图像中主要构建的精英性、接近性和重大性新闻价值有所差异。这是因为卡通相比摄影可以比较生动地刻画新冠病毒的形象,有利于民众对新冠病毒的直观认识。另外,卡通图像的运用还有利于宣传“万众一心抗击新冠疫情”等抽象意

识，增强人民抗击疫情的决心和凝聚力。这是写实的新闻摄影图像难以实现的（如图4、图5）。

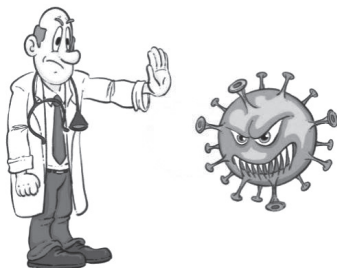


图4

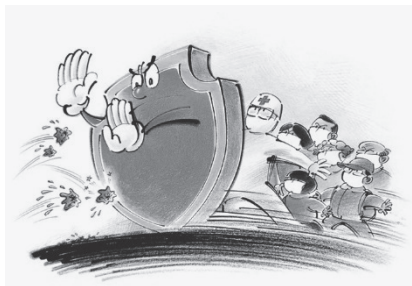


图5

## 5 结语

本文在新闻价值话语分析的框架下，运用多模态话语分析法对疫情高峰期《中国日报》疫情相关新闻报道中的新闻价值进行了分析。研究发现，《中国日报》有关新冠疫情的新闻报道中主要建构了四种新闻价值，即接近性、精英性、时效性和负面性。首先，《中国日报》在疫情发生之初，向读者及时通报了疫情的相关情况，China、Wuhan等相关地名的大量出现，建构了新闻中的接近性价值，展现了中国负责任的国际形象。其次，通过引用卫健委、国家领导人、世卫组织官员、医学专家等权威机构和人士的话语，新闻的权威性、可靠性和科学性得到了提升，同时也建构了新闻价值中的精英性。同样，不同动词时态的使用，集中反映了新闻价值中的时效性，也有利于建构积极、正面、高效的政府形象。最后，大量的新冠疫情词汇以及在这些词汇周围聚集的战斗词汇，建构了疫情新闻中的负面性新闻价值。然而，负面性新闻价值的建构不仅有利于提升报道的客观性，还有助于增强中华民族内部的凝聚力，发动广大人民共同战胜疫情。新闻图像与新闻报道在新闻价值建构中相互配合，相互补充。新闻摄影图像主要建构了精英性、重大性和接近性三种新闻价值，集中展现了中国政府在抗击突发疫情方面的坚强领导，中国政府与世界卫生组织在抗疫工作中的紧密合作，以及各级政府应对新冠疫情方面所采取的有力措施。新闻卡通图像则直观地构建了新冠病毒的卡通形象，并象征性刻画了“万众一心”的抗疫精神，达到了新闻摄影图像无法达到的传播效果。本研究使用的语料库辅助多模态话语分析法对未来的新闻价值研究具有一定的方法论借鉴意义。



## 注释

- 1 括号里的数字为该词出现的频数。

## 参考文献

- BAKER P. Using corpora in discourse analysis [M]. London: Continuum, 2006.
- BEDNAREK M, CAPLE H. “Value added”: language, image and news values [J]. *Discourse, Context & Media*, 2012, 1 (2-3): 103-113.
- BEDNAREK M, CAPLE H. Why do news values matter? Towards a new methodological framework for analyzing news discourse in critical discourse analysis and beyond [J]. *Discourse & Society*, 2014, 25(2): 135-158.
- BEDNAREK M, CAPLE H. The discourse of news values: how news organizations create newsworthiness [M]. New York: Oxford University Press, 2017.
- BELL A. The language of news media [M]. Oxford: Blackwell, 1991.
- CAPLE H, BEDNAREK M. Delving into the discourse: approaches to news values in journalism studies and beyond [R]. University of Oxford: Reuters Institute for the Study of Journalism. Oxford, 2013.
- CAPLE H, HUAN C, BEDNAREK M. Multi-modal news analysis across cultures [M]. Cambridge: Cambridge University Press, 2020.
- COTTER, C. Language and the news media: five facts about the Fourth Estate [C]// WHEELER R. The workings of language: from prescriptions to perspectives. Westport. CT.: Praeger, 1999: 165-179.
- FAN, C. An analysis of English news reports from the perspective of graduation [J]. *Theory and Practice in Language Studies*, 2020, 10(12): 1634-1639.
- GALTUNG J, RUGE M. The structure of foreign news [J]. *Journal of Peace Research*, 1965, 1: 64-90.
- GOLDING P, ELLIOT P. Making the news [M]. London: Longman, 1979.
- KRESS G, VAN LEEUWEN T. Reading images: the grammar of visual design [M]. London/ New York: Routledge, 1996/2006.
- PALMER J. Spinning into control: news values and source strategies [M]. London: Leicester University Press, 2000.
- POTTS A, BEDNAREK M, CAPLE H. How can computer-based methods help researchers to investigate news values in large datasets? A corpus linguistic study of the construction of newsworthiness in the reporting on Hurricane Katrina [J]. *Discourse & Communication*, 2015, 9(2): 149-172.
- SCOTT M. Wordsmith 3.0 [M]. Oxford: Oxford University Press, 1999.
- WEN F, YE H, WANG Y, et al. Icing on the cake: “Amplification Effect” of innovative

- information form in news reports about COVID-19 [J]. *Frontiers in Psychology*, 2021, 12: Article No. 600523.
- ZHENG S. The communication power of Chinese Novel Coronavirus Pneumonia (COVID-19) news reports in light of the framing theory [J]. *Theory and Practice in Language Studies*, 2020, 10(11): 1467-1470.
- 曹碧波. 媒体如何利用突发事件报道塑造国家形象[J]. 中国出版, 2010 (9): 31-33.
- 陈力丹. 美国传播学者休梅克女士谈影响传播内容的诸因素[J]. 国际新闻界, 2000 (5): 79.
- 陈韵昭, 吴文虎. 我们对新闻价值的基本观点[J]. 复旦学报(社会科学版), 1984 (3): 73-79.
- 邓斯佳. 中美关于我国突发事件编译比较——以CNN与《中国日报》关于昆明火车站“3·01”暴恐事件的新闻报道为例[J]. 解放军外国语学院学报, 2015 (3): 112-118.
- 何光先. 新闻价值的讨论引起重视——近几年来讨论的主要内容: 什么是新闻价值? 决定新闻价值的要素是什么? 如何增强新闻的社会效果? [J]. 新闻界, 1985 (5): 15-16.
- 郇昌鹏. 新闻价值研究的话语分析视角[J]. 当代外语研究, 2016 (5): 45-51.
- 雷跃捷. 新闻价值定义再探[J]. 现代传播, 1992 (1): 15-23.
- 李春邦. 新闻价值的含义[J]. 广西大学学报(哲学社会科学版), 1983 (1): 99-102.
- 刘建明. 创立现代新闻价值理论[J]. 新闻爱好者, 2002 (12): 10-14.
- 栾轶玫, 张雅琦. 新冠肺炎疫情报道中的信息呈现与媒体表现[J]. 新闻战线, 2020 (3): 12-15.
- 罗强, 张瀚祥. 以新媒体视角做好疫情防控阻击战报道——上游新闻客户端“新冠肺炎疫情主题报道”浅析[J]. 传媒评论, 2020 (2): 30-32.
- 潘艳艳. 多模态视阈下的国家安全话语分析——以中美警察形象宣传片的对比分析为例[J]. 外国语文, 2019 (1): 78-87.
- 石岩. 《人民日报》抗“疫”报道的话语分析[J]. 采写编, 2020 (4): 4-6.
- 童兵, 陈绚. 新闻传播学大辞典[Z]. 北京: 中国大百科全书出版社, 2014.
- 王新友. 关于新闻价值问题[J]. 现代传播, 1980 (4): 12-16.
- 许茜. 怎么用数据新闻进行灾难性报道——以2019新冠肺炎疫情事件为例[J]. 中国传媒科技, 2020 (2): 12-14.
- 杨保军. 新闻文本的价值属性[J]. 当代传播, 2003 (6): 19-20.
- 张菁, 关玲. 影视视听语言[M]. 北京: 中国传媒大学出版社, 2013.
- 赵心宇. 《人民日报》新冠肺炎疫情报道的话语分析[J]. 传媒论坛, 2021 (3): 29-31.



赵雅馨.《中国妇女报》新冠肺炎疫情防控报道中女性工作者的媒介形象研究[J].  
声屏世界, 2020 (3): 82-83.

通信地址: 361021 福建省厦门市 集美大学外国语学院 (韩存新)  
200436 上海市 上海大学外国语学院 (赵宇飞)

# 基于语料库的中美媒体关于TikTok新闻报道的批评话语分析<sup>\*</sup>

东北大学 黄馨 大连海事大学 罗卫华

**提要：**本文运用批评话语分析和语料库语言学相结合的研究方法，以《人民日报》（海外版）和《纽约时报》上关于TikTok的新闻为语料，采用费尔克劳夫三维模型，从描述、阐释和解释三个维度对文本进行分析。研究发现：在语言使用和互文性层面，两家媒体均突出了TikTok的受欢迎程度，但中方媒体更强调其安全性，美方媒体则表现其安全隐患。两家媒体背后反映的意识形态差异可追溯到中美两国具体的政治、经济、社会和文化语境。本研究对科技类话题的新闻话语分析具有一定启示作用。

**关键词：**批评话语分析、意识形态、科技新闻语料、费尔克劳夫三维模型

## 1 引言

TikTok（抖音海外版）是一款由中国企业字节跳动设计的短视频软件。自海外发行以来，这款软件深受用户尤其是青少年群体的喜爱，却在美国引发了当地科技公司甚至官方的“恐慌”。2020年7月7日，美国国务卿蓬佩奥透露，特朗普政府正在考虑禁止中国社交软件的使用，其中包括TikTok。此后，该软件就一直深陷中美政治经济争端的漩涡中。2021年6月，现任总统拜登撤销前总统针对TikTok的禁令，但同时要求美国商务部对“外国对手”掌握的应用程序进行评估，并根据结果“酌情采取行动”。TikTok风波仍未完全平息。本研究聚焦争端较为集中的年份（2020—2022年），选取《人民日报》（海外版）和《纽约时报》上关于TikTok的报道自建对比语料库，运用费尔克劳夫三维理论模型对语料进行分析，比较其语言使用和互文性特征，探究背后的意识形态问题，并结合政治、经济、社会和文化语境，解释意识形态差异的可能成因。研究以日常生活短视频软

<sup>\*</sup> 本文系教育部哲学社会科学重大攻关项目“古汉语英译大辞典编纂与数据库建设研究”（21JZD049）的阶段性成果。罗卫华为本文通讯作者。

作者贡献：

黄馨：选题构思、研究方法、数据收集、数据分析、讨论结论、初稿撰写、字数占比（75%）。

罗卫华：选题构思、研究方法、讨论结论、字数占比（25%）、修改润色。

件TikTok为切入点,探讨当下愈演愈烈的中美科技竞争和意识形态问题,既可帮助提升读者对科技话题新闻文本的批判意识,也可以加深读者对中美文化、中美关系问题的认识。

## 2 文献综述

批评话语分析(Critical Discourse Analysis,简称CDA)是一种致力于探讨语言、权力与意识形态之间关系的研究框架(Fairclough 1995),旨在分析语篇如何从社会结构和权力关系中产生,又是如何为之服务。“批评”就是揭示社会生活中习以为常的权力关系和意识形态(田海龙 2006)。Fairclough(1995)指出批评话语分析领域的“话语”是社会关系和过程中的语言使用,包括文本中出现的语言形式。话语可以看作是社会实践的一种形式,反映着社会现实。据此,批评话语分析的研究常将话语置于广阔的社会语境中,以揭露意识形态意义。起初,由于主观性过强和分析文本数量有限,该分析方式屡遭质疑,直到Hardt-Mautner(1995)首次提出将语料库方法应用到批评话语分析上,分析的客观性、文本代表性和结果可视化才具备更多保障。唐丽萍(2011)认为语料库语言学在批评话语分析中的作为空间主要在于从上下文和互文语境中对语篇成品进行分析,对大量已经实现了的意义表达方式进行批量处理,以及在低级阶进行词汇语法分析。受社会意识形态与价值取向等诸多因素影响,一般情况下,新闻语篇中会隐含意识形态方面的内容,往往导致误读,而批评话语分析能够帮助揭示这些内容,因此被广泛应用于新闻语篇分析中,分析时应注意语篇的准确性、侧重性、互文性和语言“为什么”表达的问题(单胜江 2011)。纵观以往借助语料库方法对新闻文本进行的批评话语分析研究,政治经济类话题居多(Baker & McEnery 2005; 钱毓芳 2010; 邵斌、回志明 2014; 熊文新 2022)。而TikTok涉及更多科技议题,与意识形态的关系具有间接性和隐蔽性,本文可以对这方面的研究进行一定补充。此外,中美之间有关TikTok的争端尚未完全化解,本研究或许可以为争端的解决提供一定启发。

## 3 研究设计

### 3.1 理论框架

Fairclough(1995)将话语视作文本(text)、话语实践(discourse practice)和社会文化实践(sociocultural practice)三个维度构成的统一体,并考察了三个维度之间的联系。文本分析考察的是语言形式,包括词汇、语法和文本结构方面的特征。话语实践探究文本如何产生并得到阐释,涵盖生产(production)、传播

(distribution) 和接受 (consumption) 三个过程。社会文化实践则反映文本生成过程中深层次的社会、政治、经济和文化因素, 关注话语与斗争过程和权力之间的关系。每个维度的分析对应一个层面: 对文本的语言学描述 (description)、对文本与话语实践之间关系的阐述 (interpretation) 以及对话语实践与社会文化实践之间关系的解释 (explanation)。其中描述层面刻画文本表层特征, 解释层面揭示深层意识形态和权力关系, 阐释层面则是两者的中介。

### 3.2 研究语料及问题

本研究基于语料库的方法, 借助语料库工具 AntConc 4.1.2 和 UAM Corpus Tool 6, 将定量分析与定性分析相结合。笔者以 “TikTok” 为关键词, 筛选《人民日报》(海外版) 和《纽约时报》2020年7月至2022年7月关联性较强的有效报道。检索日期基本涵盖特朗普政府从发布 TikTok 禁令到败选的时间, 其中特朗普和拜登政府的任期几乎各占一半。鉴于《纽约时报》符合要求的文本篇幅和数量远大于《人民日报》(海外版), 为保持对比语料库容量对等, 本研究进一步对前者文章进行删减, 最终建立《人民日报》(海外版) (CDY) 和《纽约时报》(NYT) 两个小型语料库。其中, CDY 语料库包含 54 篇文章, 共计 25,969 词, NYT 语料库包括 22 篇文章, 共计 26,753 词。本研究拟回答以下三个问题:

(1) CDY 语料库和 NYT 语料库中的文本在文本语言特征 “描述” 上有何特征和差异?

(2) CDY 语料库和 NYT 语料库中的文本在互文性 “阐释” 上有何特征和差异?

(3) CDY 语料库和 NYT 语料库中的文本背后反映了怎样的意识形态? 这些意识形态又是如何通过语境 “解释” 成因的?

## 4 研究结果和讨论

### 4.1 文本语言特征描述

本节借助软件 AntConc 4.1.2 对语料主题词、索引行和情态动词进行检索和分析, 对文本进行语言特征描述。

#### 4.1.1 主题词分析

主题词指在和参照语料库比较时统计出的具有特殊词频的词 (钱毓芳 2010)。相较于高频词分析, 主题词分析能够凸显那些在参照语料库常规标准下具有显著特征的词, 揭示隐含的意识形态。而且主题性绝对值越大, 主题词显著性越高。本研究选用的参照语料库是布朗语料库的分支 Crown 语料库, 共计 1,020,323 词,

涵盖美式英语新闻文本。经统计，CDY语料库主题词共获232个，NYT语料库208个。限于篇幅，笔者仅列举显著性较高的前30个主题词（见表1），并已删去无太多实义的功能词。为便于比较，本研究根据语义范畴将主题词进行分类，并划分两个语料库共用和特有的主题词（见表2）。

表1 部分CDY和NYT语料库主题词

序号	CDY 语料库		NYT 语料库	
	主题词	主题性	主题词	主题性
1	TikTok	2,450.953	TikTok	2,969.001
2	app	868.884	app	825.452
3	Chinese	706.626	videos	814.176
4	Bytedance	488.049	Trump	517.734
5	China	436.885	Bytedance	513.574
6	users	434.530	Chinese	460.356
7	Trump	375.933	users	411.264
8	ban	366.410	said	404.079
9	video	289.307	content	303.508
10	said	251.793	Oracle	293.414

表2 CDY和NYT语料库主题词分类

语义范畴	共有主题词	CDY 语料库专有主题词	NYT 语料库专有主题词
政治类	China, Chinese, Trump, administration	Ban, UK, India, Indian	Biden
商贸类	Company, Bytedance	Commerce	Oracle
媒体类	Said, media	...	...
感受评价类	Security	...	Concerns, transparency, comment
软件技术类	App, TikTok, Instagram, video(s), platform(s)	Tech, douyin, WeChat, online, livestreaming	Data, algorithm, Youtube, Facebook, résumés
社会影响类	Users, content	Social, popular, global, covid	Viral, creators, followers, cartel, lawmakers

表2显示,在政治层面,共有主题词China、Chinese、Trump和administration表明两个语料库文本都比较关注TikTok的“中国”背景和“特朗普政府”发挥的作用。不同的是,CDY语料库直接聚焦于“禁令”,并将视野拓展到“英国”和“印度”的相关政策,如印度的TikTok禁令,以强调这些政令的不合理性。而NYT语料库还突出“拜登”上台后持续监测TikTok的决定,可见《纽约时报》重视传递美国官方声音。在商贸上,两者均凸显多方“公司”的动向,并重点关注TikTok的母公司“字节跳动”。此外,CDY语料库重视TikTok与“贸易”的联系,如促进电商的发展。NYT语料库则更多关注美国公司“甲骨文”收购TikTok的进展,侧重美方利益。在媒体类中,media点出TikTok的社交媒体属性,said表示新闻文本多引用他者话语,避免单一的主观陈述。涉及情感态度的主题词较少。总体上,双方都关心TikTok能否给用户带来“安全感”。结合语境具体来看,CDY语料库文本多次声明TikTok的安全性,而NYT语料库则对此屡屡表示“担忧”,对其“透明度”经常持消极态度。通过索引行检索,本文发现comment在NYT语料库中出现了33次,其中16次与短语decline to搭配,10次与did not immediately respond to a request for搭配。据此,NYT语料库着重表现双方在诸多问题上“不予评价”的谨慎或模糊态度。

在科技应用方面,通用主题词多反映TikTok的基本属性,或将其与美国同类软件Instagram进行比较。CDY语料库强调“技术”一词,意味着此事谈论的不仅是一个软件的归属问题,更是中美之间的科技竞争。由于美方的技术霸权,“微信”也面临着同样处境。此外,该语料库特别强调TikTok的“联网”技术和“直播”功能。而NYT语料库文本显示美方热衷于获取TikTok的“数据”和“算法”,还关注美国社交软件“脸书”和“油管”,并与TikTok进行比较,凸显后者的些微不足。Résumés一词指向TikTok可帮助用户投放视频简历找工作的功能。在社会影响上,两个语料库均强调“用户”体验和视频“内容”。CDY语料库彰显TikTok“广受欢迎”的“全球”影响力,尤其在“新冠”疫情期间,发挥着“社交”桥梁作用。NYT语料库中,followers出现23次,其中有20次与十万级以上数字搭配,它和creators、viral一起体现了该软件视频传播快、流量高、粉丝多的特点。但cartel一词反映了TikTok可能为非法组织如贩毒集团利用的弊端。最后,lawmakers体现了美国司法体系在这场风波中所起的作用。

综上,两个语料库文本的主题词均突出了TikTok的政治经济关联性和受欢迎程度;此外,CDY语料库侧重国际视野、具体功能和全球影响,NYT语料库则侧重美方视角、核心技术和主观感受。

#### 4.1.2 情态分析

Fairclough (1992a)认为情态可以反映话语制造者确信或疏离某个观点的程度。情态可通过情态副词、情态形容词等多种途径体现。限于篇幅,本文仅考察

文本中情态动词的特征。Halliday（2004）依据说话者对观点的确信程度将情态动词分为高值、中值和低值情态动词。本研究利用AntConc 4.1.2软件的KWIC功能统计两个语料库的情态动词数量，其分布情况如下（见表3）。同时，情态动词比较情况见图1。

表3 CDY和NYT语料库情态动词分布

值	情态动词	CDY 语料库			NYT 语料库		
		数量	总数	占比	数量	总数	占比
高值	<i>Must</i>	6	22	8.0%	8	30	9.7%
	<i>Has/had/</i>						
	<i>have to</i>	8			19		
	<i>Can't</i>	5			2		
	<i>Couldn't</i>	3			1		
	<i>Will</i>	74			28		
	<i>Would</i>	50			85		
中值	<i>Should</i>	25	165	60.0%	18	151	48.9%
	<i>Is/was/are/</i>						
	<i>were to</i>	9			8		
	<i>Won't</i>	3			3		
	<i>Wouldn't</i>	2			9		
	<i>Shouldn't</i>	2			0		
	<i>Can</i>	46			42		
低值	<i>May</i>	14	88	32.0%	13	128	41.4%
	<i>Could</i>	23			56		
	<i>Might</i>	5			17		
	共计	275		100.0%	309		100.0%

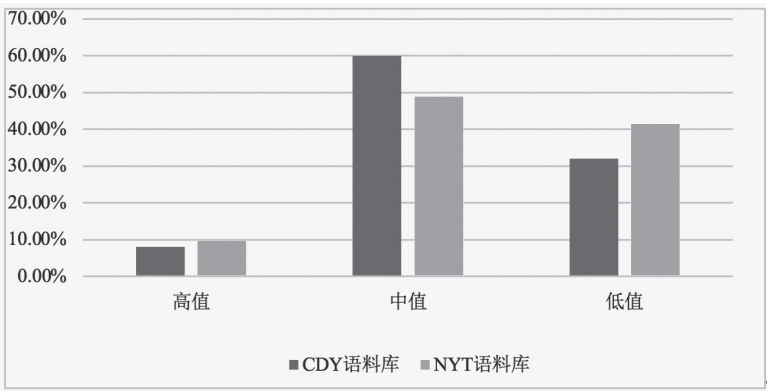


图1 CDY和NYT语料库情态动词比较情况



根据图1,两个语料库中的中值情态动词占比均最大。通常情况下,中值情态动词能够避免流露高值情态动词绝对化或低值情态动词不确定的语气,传递客观适中的态度。这说明两家媒体均尽可能采用客观口吻进行表达。CDY语料库的中值情态动词占比高于NYT语料库,后者的低值情态动词占比更高,高值情态动词占比也略高于前者。据此,总体而言,CDY语料库新闻文本的客观性要强于NYT语料库。为便于分析,笔者从语料库中抽取以下例句。

(1) While there should be guardrails for Sino-US relations, they should not be fences built by the US side with the aim of containing China. (CDY)

(2) Mr. Biden has said America must be tougher toward Beijing... (NYT)

(3) TikTok and its owner, the Chinese social media giant ByteDance, have been in the cross hairs of the Trump administration, which is concerned that the app could help the Communist Party of China obtain Americans' private information. (NYT)

例(1)在对中美关系友好发展发出呼吁时,用中值情态动词“should”表明打破“藩篱”是客观上双方应尽的义务,同时流露出礼貌协商的语气,更易于听者接受。例(2)则通过高值情态动词“must”彰显美方的强硬对华政策。最后,例(3)通过低值情态动词“could”映射美方在TikTok是否构成威胁问题上模棱两可的态度。

综上,两个语料库文本均倾向使用中值情态动词体现新闻的客观立场。而NYT语料库文本还常运用高值情态动词表明美方对华态度的强硬,此外,对TikTok安全问题的模糊态度也可透过低值情态动词看出。

## 4.2 互文性阐释

“互文性”强调意义产生于语篇之间的相互联系,任何语篇中都充满其他语篇的片段和他者的声音(辛斌 2021)。Fairclough (1992b: 269)指出“在分析作为话语实践的话语事件时互文性是个重要概念。它为了解(由语篇的异质性、意义、形式和风格实现的)话语事件的复杂性提供了途径。”由于标注工作难度较大,本研究仅从两个语料库中抽取样本,借助软件UAM Corpus Tool 6对样本转述来源和方式进行标注。CDY语料库样本共计9,283词,转述来源数目184条,转述方式180条;NYT语料库样本共计9,145词,转述来源条目163条,转述方式170条。

### 4.2.1 转述来源

转述来源可以反映新闻渠道的可信度,按照具体程度可以分为具体、略具体

和不具体来源（张健 1994）。具体来源提供了说话者的详细信息，如姓名、职业、阶层等。略具体来源仅透露模糊的信息，如“一些专家”。最后，不具体来源完全忽略或隐匿说话者信息，如“据说”。经统计，两个语料库中转述来源的分布情况如下（见表4）。

表4 CDY & NYT语料库样本转述来源分布

类型	CDY 语料库		NYT 语料库	
	数量	占比	数量	占比
具体	157	85.3%	111	68.1%
略具体	18	9.8%	34	20.9%
不具体	9	4.9%	18	11.0%
共计	184	100.0%	163	100.0%

表4中，CDY语料库的具体来源占比高于NYT语料库，而后者的其他两类来源占比均高于前者。总体来看，CDY语料库新闻文本相较于NYT语料库更具真实性和透明度。鉴于具体来源涵盖甚广，本研究依据转述来源性质，将其分成以下五大类。（1）官方来源：包括国家领导人，如“特朗普”；代表政府机关的专有名词，如“白宫”；以及政府公布的声明和文件等。（2）专业来源：包括为TikTok事件发声的专业人士、机构组织和专门文件等。（3）企业来源：包括企业的名称、发言人、文件和产品等。（4）媒体来源：包括新闻机构的网站、报纸、报道和数据等。（5）公众来源：包括如用户、目击证人等无特殊背景的大众。

两个语料库文本都倾向采用官方和企业来源，也从侧面反映了政府和科技公司在这场争端中拥有很高的话语权。此外，专业来源在CDY语料库中的占比远大于NYT语料库，而后者的公众来源占比更高，初步反映了CDY语料库的新闻来源更具权威性。具体来说，CDY语料库的官方来源可大致分为两类：一类是美国政府的言论，多数映射美方政治霸权，另一类是中国政府反击的声音。而NYT语料库中几乎没有中方发言，只有美方声音，多数为对TikTok安全问题的质疑和指控。媒体来源情况类似，CDY语料库文本敢于引用中媒以外的一些正义之词（见例4），而NYT语料库只引用美媒和同立场的西方媒体发言（见例4和例5）。这在一定程度上降低了后者的说服力。

（4）The Washington Post obtained internal details from Targeted Victory that outlined a campaign to undermine TikTok. (CDY)

(5) The New Yorker has called the invasion the world's "first TikTok war."  
(NYT)

研究发现,略具体来源多是显示职业的复数名词,如judges、analysts和experts等。究其原因,一是作者本身对来源不甚清楚;二是详写必要性不大;三是作者刻意模糊化处理。第三种通常最能反映作者的主观意识。如例(6),allies一词一般带有政治军事色彩,比如二战时的同盟国。CDY语料库避免列出具体国家,一方面默认读者知晓,另一方面也避开了直接点明政治立场的敏感词汇。例(7)中,employees是TikTok公司员工,他们不想因发表对自己现在或前公司的不利言论而惹上麻烦。报道者采取这种方式,可能是出于保护隐私考虑,也可能是在避免后续的直接对质。最后,两个语料库均使用不具体来源陈述众所周知的事实。但在NYT语料库中,由于说话者身份敏感,作者无法透露任何信息的情况时有发生见例(8)。这虽然符合新闻记者未经允许不得公开受访人信息的职业道德,侧面说明受访者不必为自身言论担责,却也在一定程度上损害了新闻的公信力。

(6) With targets that include not only competitors but also allies, the United States is a true pursuer of large-scale, indiscriminate wiretapping and secret theft. Even its allies find this "unacceptable". (CDY)

(7) Three current and former employees expressed concerns about the Chinese-owned app's safeguards for preteen children. (NYT)

(8) But while the White House did not extend the deadline for a deal again, it also plans to take no immediate action in response to the lapsed cutoff, said the person with knowledge of the discussions, who was not authorized to speak publicly. (NYT)

#### 4.2.2 转述方式

新闻的转述方式有多种分类,限于篇幅,本文仅涉及3种转述方式:直接转述、间接转述和混合转述。混合转述指的是前两种转述方式的混合,即在间接转述的话语中包含双引号标注的直接引用部分。间接转述中,转述者可能对原话进行修改加工,因此其还原度要低于直接转述,混合转述的客观性介于前两者之间。混合转述能够在读者未来得及反应时混淆叙述者与原说话人之间的话语边界(Leech & Short 1981),从而加强佐证。据统计,两个语料库中的样本转述方式分布见表5。

表5 CDY & NYT 语料库样本转述方式分布

类型	CDY 语料库		NYT 语料库	
	数量	占比	数量	占比
直接	50	27.8%	46	27.0%
间接	100	55.5%	106	62.4%
混合	30	16.7%	18	10.6%
共计	180	100%	170	100%

表5显示，两者直接转述占比基本一致，NYT语料库的样本间接转述比例高于CDY语料库，而后的混合转述高于前者。总体而言，CDY语料库的文本真实还原度要高于NYT语料库。

在直接转述中，两个语料库均热衷于直接引用专家或TikTok用户的发言，保证文本的专业性和通俗化。其中，两者的用户言论基本呈正向，这印证了TikTok在顾客群体中的受欢迎程度不容忽视。CDY语料库中，专家发言多是对和平解决争端的建议，而NYT语料库却更偏向引用美国专家对TikTok安全性的质疑和担忧。

间接转述是作者灌输自身观点最常见的方式，而且这种灌输有时难以察觉。本研究从两个语料库中提取一对内容接近的间接引语，见例（9）和例（10），均涉及特朗普在2020年8月份针对TikTok发布的政令。

（9）Trump ordered last August that the TikTok app, which lets users share video clips and is especially popular with young people, be sold to an American firm or face a ban in the US, but that sale effort failed. (CDY)

（10）It said TikTok could maintain U.S. operations only if it sold itself to a U.S. company and shed all Chinese-based infrastructure and ties. (NYT)

CDY语料库的引语插入了定语从句，表明TikTok是一款很受欢迎的短视频分享软件，暗示美方觊觎该软件的原因。or连接美方给予的两个选择，实际上TikTok毫无选择权可言，因为没有一个是公正合理的。but又将笔锋一转，直指特朗普的计谋不会得逞。而NYT的转述开头并未指出TikTok拒绝出售的严重后果，而是反向着笔，说明出售的好处。said相较于ordered，语气大幅度减弱，给人可以商量的余地。但only if又强硬表明TikTok继续在美运行是有硬性条件的。此外，shed做“摆脱”义时，后面多搭配贬义表达，这里也在暗示美方针对的是TikTok的中国背景。综上，作者或多或少会在间接引用过程中加入自己的主观意识。

混合型转述中，两个语料库文本多会直接引用主观性的关键信息，间接引用

部分多在客观叙述事实。但媒体为避免直接担责，也会将敏感话语借他人之口表述，同时加强自身观点。例（11）双引号中的poison chalice一般指迷人却有害的事物，带有贬义。结合前文，可知作者已表示担忧，其观点可借名人比尔·盖茨之口加强，又能规避直接责任。

（11）Yet the concerns are that some under-13 users may lie to get around the age restrictions, and that the platform is not obtaining the required consent from those users' guardians. Bill Gates, Microsoft's co-founder, recently told Wired's magazine that TikTok was a "poison chalice" for any buyer, referring to its complexity. (NYT)

### 4.3 语境因素解释

通过前文的文本描述和话语实践，可知两家媒体均在新闻报道中映射了中美双方意识形态。CDY语料库文本致力于透过TikTok软件传递开放包容、推崇公平正义的中国态度，以及坚决抗争不公的魄力。NYT语料库文本一方面积极宣传TikTok在广大美国用户中的热度和实用性，另一方面又质疑甚至无端指控其安全问题，借“国家安全”名义希望美方接管，体现了美方为自身利益大搞科技霸权和强权政治的处世之道。本节将从两国政治、经济、社会、文化语境解释意识形态差异的可能成因。

#### 4.3.1 政治经济语境

作为宣传的重要手段，新闻媒体难以不受政权影响。《人民日报》（海外版）肩负着向海外传递中国声音的使命。《纽约时报》是美国国内三大报刊之一，国务院、国会、各国大使馆和社会团体都依赖它来建立普遍性的参考框架，社论常反映美国国务院的观点（耿芳 2012）。

2020—2022年，TikTok事件前后历经两任美国总统。特朗普政府秉持着“美国优先”的外交原则，不惜一切代价维持美国“世界霸主”地位，将中国在各个领域的崛起视为“威胁”。TikTok在美国热度上升的同时，也在逐步占据当地市场份额。据彭博社报道，TikTok在2022年的广告收入额仅次于美国巨头视频软件YouTube。此外，该社预测2024年两者将持平。TikTok能够创收高额利润，主要依靠其智能推荐算法，这也是此次争端中双方谈判交易的关注重点。因此，尽管只是商业公平竞争，美国政府为了美国的利益也会竭力遏制中国在全球的数字经济发展，维持其在科技和经济领域的领先地位。此外，笔者认为，作为政客，特朗普在一定程度上也有个人想法。他颁布禁令时正值美国大选，一方面由于TikTok吸引了大量美国底层民众，他莫名恐惧中国政府会通过TikTok干涉选民决



定；另一方面他希望通过强硬对华政策赢得美国反华势力选民的支持，将民众视线从他应对新冠疫情的失败中转移出来。但他的禁令由于证据不足屡遭法院驳回。直到大选结束，拜登胜出，拜登政府废除前总统的禁令，转而采用一种迂回的方式，要求商务部审查包括中国在内的外国竞争对手所掌控的软件是否会对美国国家安全构成威胁，再根据审查结果采取“合适”措施。实际上这与特朗普政府的行为殊途同归。

中国始终坚持独立自主的外交政策，不会以牺牲他国利益为代价来谋发展，也绝不会放弃自身合法利益。一方面，CDY语料库的新闻文本致力于传播TikTok的全球影响，塑造友好和平、追求互利共赢的中国形象，如其一直强调TikTok的“分享”观念，与开放共享的中国发展理念趋同。另一方面，它多次申明中方严拒不公要求。2020年8月28日，中国商务部、科技部调整发布《中国禁止出口限制出口技术目录》，其中便限制了TikTok算法的出口。

#### 4.3.2 社会文化语境

文化是一个群体共享、学习并代代相传的信仰、习俗、价值观、行为、制度和沟通方式总和（Davis 2001）。不同的文化土壤能够孕育出不同的社会风气。

在美国，金融垄断资本主义长期占据市场，金融寡头在相对自由松散的政策环境下掌握着国民经济命脉。实际上，纵观古今，资本主义国家的政治经济策略主要代表并维护资本家利益。由于美国政府对“反垄断法”的忽视，当地巨头公司通过各种手段垄断自家产品市场，如谷歌收购YouTube，企图垄断短视频和网络广告交易市场。TikTok的美国竞争对手多次尝试收购未果，就开始抹黑，如《华盛顿邮报》就报道过Meta公司（原名Facebook）雇佣咨询公司，计划引起公众对TikTok的不满。尽管屡遭挫折，该软件依旧在美收获众多用户。《人民日报》（海外版）的一篇报道对此评价道：“实用主义最终战胜了意识形态”。美国实用主义哲学“要求一切从实际出发，而不是从理论和逻辑出发，把实际效果看作是检验一切理论和学说的标准，其目的在于应付生活环境，解决人们在现实中所遇到的问题”（许一多、周俊峰 2002：51）。当多数美国客户在TikTok上获得实际好处时，便不怎么关心还未实现的“国家安全威胁论”。此外，美国司法体系也在此次事件中发挥着重要作用。美国实行三权分立，立法、行政和司法权相互制约制衡。因此无论是特朗普还是拜登，都无法凌驾于法律之上，随意禁用一款能为美国社会创造利益的软件，这也是字节跳动公司可以抓住的机会。

此次事件中，透过中国企业文化和产品设计理念，中国社会价值观的冰山一角也得以瞥见。面对打压，TikTok的母公司字节跳动主动采取一系列公开有效的措施，彰显了中国企业的诚意、骨气和担当。如该公司在当地调整人员结构，雇佣更多美国员工，在新加坡建立数据储存库等。协商不成，它才与美国政府对簿公堂。此外，TikTok的设计软件也彰显了开放包容的价值观。这款软件给予不同

种族背景的人发声机会，任何有创造力的人都可能通过努力名利双收，任何人都可以找到符合自己喜好的内容。

## 5 结论

本研究选取《人民日报》（海外版）和《纽约时报》上2020—2022年关于TikTok的报道自建语料库，以费尔克劳夫三维模型为理论框架，从描述、阐释、解释三个维度对新闻语料进行批评话语分析。研究发现：在文本描述层面，CDY语料库和NYT语料库的主题词均凸显了TikTok的用户吸引力和政治经济关联性，但侧重各有不同，前者强调国际视野、具体功能和全球影响，后者则突出美方视角、核心技术和主观感受。此外，CDY语料库采用了更多中值情态动词，传递出客观适中的态度；而NYT语料库则包含更多高值和低值情态动词，语气更为绝对或随意。在互文性阐释层面，CDY语料库运用更多具体新闻来源和直接转述方式，加强了文本真实性；而NYT语料库使用更多模糊报道渠道和间接引用方法，使得文本透明度降低。语境解释上，《人民日报》（海外版）文字背后反映了中方开放包容、推崇公平正义的态度，以及面对不公的抗争意识和魄力，可以溯源到中国政府对外经济贸易政策、中国企业文化和TikTok设计理念。《纽约时报》则映射了美方为自身利益大搞科技霸权和强权政治的处世之道，以及美国社会的逐利意识，也可从美国政客对华策略、垄断资本主义、实用主义思想和司法体系中找到一些根源。

本研究也存在不足：所用语料库规模不大，缺乏对两任美国总统在任期间的历时对比研究，对费尔克劳夫三维模型第一阶段文本描述的研究仅涉及词汇层面，不够全面。期待后续相关学者能够进行补充拓展，以丰富科技类话题的新闻话语分析研究。

### 参考文献

- BAKER P, MCENERY T. A corpus-based approach to discourses of refugees and asylum seekers in UN and newspaper text [J]. *Language and Politics*, 2005, 4(2): 197-226.
- DAVIS L. *Doing culture: cross-cultural communication in action* [M]. Beijing: Foreign Language Teaching and Research Press, 2001.
- FAIRCLOUGH N. *Discourse and social change* [M]. London: Polity, 1992a.
- FAIRCLOUGH N. Intertextuality in critical discourse analysis [J]. *Linguistics and Education*, 1992b, 4(3-4): 269-293.
- FAIRCLOUGH N. *Critical discourse analysis: the critical study of language* [M]. London: Longman, 1995.



HALLIDAY M. An introduction to functional grammar [M]. New York: Oxford University Press, 2004.

HARDT-MAUTNER G. Only connect: critical discourse analysis and corpus linguistics [M]. Lancaster: University of Lancaster, 1995.

LEECH G, SHORT M. Style in fiction [M]. London: Longman, 1981.

耿芳.《纽约时报》涉华报道的意识形态倾向——批评话语分析的视角[J]. 现代传播, 2012 (7): 155-156.

钱毓芳. 语料库与批判话语分析[J]. 外语教学与研究, 2010 (3): 198-202.

单胜江. 新闻语篇的批评性话语分析[J]. 外语学刊, 2011 (6): 78-81.

邵斌, 回志明. 西方媒体视野里的“中国梦”——一项基于语料库的批评话语分析[J]. 外语研究, 2014 (6): 28-33.

唐丽萍. 语料库语言学在批评话语分析中的作为空间[J]. 外国语, 2011 (4): 43-49.

田海龙. 语篇研究的批评视角: 从批评语言学到批评话语分析[J]. 山东外语教学, 2006 (2): 40-47.

辛斌. 批评话语研究中的互文性分析[J]. 外语与外语教学, 2021 (3): 1-12.

熊文新. 新闻报道主观性的语言学透视——一种结合语料库驱动和批评话语分析的方法[J]. 现代传播, 2022 (5): 22-32.

许一多, 周俊峰. 实用主义: 美国的外交哲学[J]. 长白学刊, 2002 (5): 51-54.

张健. 新闻英语文体与范文评析[M]. 上海: 上海外语教育出版社, 1994.

**通信地址:** 110057 辽宁省沈阳市 东北大学外国语学院 (黄馨)  
116033 辽宁省大连市 大连海事大学外国语学院 (罗卫华)

# 基于CIA模型的列举类词习得研究<sup>\*</sup>

山东大学(威海) 李艳娇 李 齐 庄会彬

**提要:** 基于中国英语学习者语料库和英语本族语者语料库, 本文从频率、共现、位置三方面对高频使用的列举类词语 such as 与 for example 进行定量和定性分析。研究发现, 中国英语学习者对 such as 和 for example 的使用存在以下问题: such as 和 for example 使用偏少, 但考试环境下 for example 过度使用; 在共现上, such as 和 for example 均过多引导多个列举子项, 左侧与句号共现过多, 与后助式列举类词共现过多; 句首使用过多。这些问题是母语负迁移、词典及教材例句选取不当、辨析不够全面等多种原因共同作用的结果。基于此, 本文得到如下启示: (1) 教师应提高语料库素养, 引导学生归纳 such as 和 for example 的用法, 并进行预警, 以减少母语负迁移的影响; (2) 词典及教辅资料应根据英语母语者使用习惯进行编写, 增加辨析与举例, 并提供必要的信息。

**关键词:** such as、for example、语料库

## 1 引言

中介语对比分析模型(Contrastive Interlanguage Analysis, 简称CIA)由Granger提出, 主要涉及母语和中介语、不同中介语的两种对比(Granger 1998a), 能够更好地揭示学习者的语言特征。随着大规模二语语料库的开发, 该模型实现了对比分析理论与语料库语言学的结合, 已成为二语习得研究的重要研究范式(邢红兵、辛鑫 2013)。其研究信度与效度因使用大量真实语言数据而大幅提高, 对促进二语教学有重要作用(卫乃兴、陆军 2018)。因此, 本文将利用语料库对比分析中国英语学习者和英语本族语者的语言特点, 进行CIA范式“最为基础, 也最为经典的对比分析活动”(卫乃兴、陆军 2018: 47), 这对深入认识中国英语学习者的表达特点有重要意义。

<sup>\*</sup> 本文系国家社科基金青年项目“汉语会话行为标注及自动识别研究”(20CYY021)的阶段性成果。李艳娇为本文通讯作者。

作者贡献:

李艳娇: 研究方法、讨论结论、初稿撰写、字数占比(60%)、修改润色。

李 齐: 数据收集、数据分析、字数占比(40%)。

庄会彬: 选题构思。

中介语对比分析模型的研究对象包括词语、词语组合、词语搭配等，共选理论渗透其中。共选是语料库语言学领域重要的理论阐述（Sinclair 1991, 1996, 2004），指“语言交际过程中形式与形式、形式与意义的共选”（卫乃兴 2012: 1）。其主导思想为语言形式的选择受到周围环境的制约，包括词汇与词汇、词汇与语法、由词汇和语法构成的型式（*pattern*）与意义之间的共选，遵循从词项到语境再从语境到词项的“词项—语境法”，并以语料库的KWIC词语索引为技术手段展示可观察的、可量化的实例。本文将基于语料库对节点词语使用特征进行观察、统计及分析，是共现理论在具体研究中的体现。

列举类词使用频率较高，涉及范围较广，用法较为灵活（徐敏 2010）。英语中这类词语涉及 *such as*、*for example*、*for instance*、*and so on* 和 *and so forth* 等，其中 *such as* 和 *for example* 使用频次最高（Hyland 2007; Triki 2021）。国外学者侧重利用语料库对列举类词进行多维度考察，如 Triki（2021）基于语料库考察列举类词在不同学科书面语文本中的分布情况，以反映不同学科的语言表达特点；Paquot（2008）利用学习者和母语者语料库，对比分析了荷兰语、法语、德语、波兰语、西班牙语五种母语背景学习者的多词组合使用情况，研究表明各母语背景学习者都过多使用了 *for example*，主要受母语迁移、教材引导及学习者使用心理的影响。此外，Rodríguez-Abruñeiras（2015, 2021）从历时和共时两个角度，对例证标记（如 *including*、*included*、*for example*、*for instance*）进行了基于语料库的描述与分析，重点关注语法化过程。国内对这些英语列举类词语关注不多，成果寥寥，多以教学规定语法的口吻描述其用法（戴卫红 2001；黄龙旺、龚汉忠 2008；梁石 2012；陈仁楨 2014；金子铭 2018；马懿 2021），缺乏研究性成果，尤其缺乏面向二语习得的定量研究（郭书彩等 2015）。

对英语学习者的书面表达来说，举例是一个非常重要且经常使用的修辞功能。在外语教学中，我们发现中国英语学习者对列举类词语的掌握并不理想（黄龙旺、龚汉忠 2008；郭书彩等 2015），这为本文的研究留下了巨大空间。鉴于此，本文以列举类词中使用频率最高的 *such as* 和 *for example* 为研究对象，通过语料库的对比分析揭示中国英语学习者的使用特点，为国内的外语教学提出合理化建议。

## 2 研究设计

### 2.1 研究问题

本文试图回答以下两个问题：第一，中国英语学习者与英语本族语者对 *such as* 及 *for example* 的使用是否存在差异？如果存在差异，导致这些差异的原因是什么？第二，国内的英语教学应该如何相应改进？

2.2 语料选取

本研究选取的语料为NESSIE语料库（Native English Speakers Similarly or Identically-prompted Essays, Version 2）中英语本族语者依照中国大学英语四六级、英语专业四八级考试作文题目所撰写的文章NESSIE\_C，以及TECCL（Ten-thousand English Compositions of Chinese Learners, Version 1.1，中国学生万篇英语作文语料库）大学部分的语料TECCL\_C和CLEC（中国学习者英语语料库）中的子语料库ST3、ST4（CLEC\_C）。NESSIE语料库v2由许家金教授在2013年建成，共包含781个样本，共计321,768形符，其文本主要是英语本族语者依照中国大学英语四六级、英语专业四八级考试作文题目所撰写的作文，也有部分语料来自BAWE、MICUSP等英美高校学生（含非母语者）撰写的作文。中国学生万篇英语作文语料库（TECCL）收录9,864篇作文，1,990,258形符，来自2011—2015年的课堂限时作文，课后家庭作业，期中期末考试作文，课堂演讲稿，以及小组协作作文等，涉及不同作文题目一千多个，因此该语料库具有“语料新”“题目多”“学段宽”“地域广”“任务活”（薛熙哲 2015）等特点。中国学习者英语语料库（CLEC）包括ST2、ST3、ST4、ST5和ST6五个子库，语料分别来源于中国中学生，大学英语四、六级考试，以及英语专业低、高年级五种难度等级的英语学习者作文，包含考试作文及少量自由作文（日记、读书笔记、不限定题目的作文），整个语料库共1,070,602形符，以广泛代表中国学习者的书面语学习状况（杨惠中等 2005）。本文的语料分布见表1。

表1 本文的语料分布

语料	构成		总计
NESSIE_C	CETtopicsbyNS	96,579	205,240
	TEMtopicsbyNS	108,661	
TECCL_C	university	1,530,408	1,530,408
CLEC_C	ST3	209,043	421,898
	ST4	212,855	

本文选取以上3个语料库，主要考虑以下因素。内容接近，三者都是高校学生的作文，且话题相近，对比性强。“CLEC\_C”和“NESSIE\_C”都来自大学英语四六级的作文，前者由中国学习者完成，后者由英语本族语学生完成，在话题和内容上可直接对比。此外，“TECCL\_C”的作文“属于英语课程体系内的学业任务”，与“高风险的标准化考试作文”不完全一致，本文采纳了该部分数据，一方面是作文题目相近，另一方面，也想考察考试环境与非考试环境对学习者的语言使用的影响。

2.3 研究步骤

本文采用中介语对比分析的语料库研究方法。具体研究步骤如下：第1步，分别在三个语料库中提取such as和for example的原始频数，转换成标准频率，即每百万词中该词出现的次数，运用对数似然率计算器（Log-likelihood Ratio Calculator）<sup>1</sup>检验频数之间是否存在显著差异性。其中，NESSIE\_C及TECCL\_C中的语料由北京外国语大学多语种在线检索平台CQPweb检索获得，CLEC\_C中的语料由AntConc（4.0.11）检索《中国英语学习者语料库》（桂诗春、杨惠中 2002）附带光盘进行获得。第2步，将such as和for example的例句复制到表格中，利用Antconc（4.0.11）对其列举子项的数量及形式、与标点符号及后助式列举类词的共现情况、句中位置等进行检索及人工标注。第3步，基于CIA模型将英语本族语者和学习者的数据分别进行对比分析，得出中国英语学习者在使用过程中存在的问题，同时分析原因，以期对外语教学有所启示。

3 英语本族语者与中国英语学习者使用情况对比

本节主要通过英语母语者和中国学习者的数据对比，突出中国学习者such as和for example的使用特点，具体涉及总体频次、共现情况、句中位置等信息。

3.1 总体频次

表2展示了such as和for example在三个语料库中的总体分布情况。通过观察可知，与本族语者语料库（NESSIE\_C）相比，两个学习者语料库表现出了不同的分布规律。第一，TECCL\_C中中国学习者such as和for example使用偏少，且such as较母语者差异显著（利用对数似然比进行检验，P值为0，详见表3）；CLEC\_C中中国学习者such as的使用跟母语者基本一致，但for example明显高于母语者，差异显著，存在过度使用的现象。第二，英语本族语者such as的频数高于for example，TECCL\_C中保持一致，而CLEC\_C中for example的频数高于such as。

表2 such as和for example出现频次

	出现频次	NESSIE_C	TECCL_C	CLEC_C
such as	原始频次	187	972	361
	标准频次（每百万词） <sup>2</sup>	911	635	886
for example	原始频次	97	620	641
	标准频次	473	405	1519

表3 such as 及 for example 的差异显著性检验

		TECCL_C	CLEC_C
such as	Log-likelihood <sup>3</sup>	18.81	0.48
	Sig. <sup>4</sup>	0 *** -	0.487 -
for example	Log-likelihood	1.92	150.56
	Sig.	0.166 -	0 *** +

两个中国学习者语料库中 such as 和 for example 的使用频数存在差异，主要原因是，TECCL\_C 的作为学业任务，学习者有更多时间可以查字典、请教别人，因此可以选择更为丰富的表达方式，such as 和 for example 的使用频数少于母语者，也少于 CLEC\_C 四六级考试环境的中国学习者。CLEC\_C 为四六级试卷作文，受制于写作时间及避免出错的需求，学习者头脑中优先调取较有把握的 such as 及 for example。而相比 such as，for example 作为插入语，语法化程度及自主性更高 (Rodríguez-Abruñeiras 2015)，在写作时使用更加自由灵活，加入并不影响句子结构，不必考虑列举子项必须为名词性形式等 (黄龙旺、龚汉忠 2008)，因此考试环境下学习者会更青睐使用 for example，造成 for example 使用频率高于 such as 的情况。

3.2 共现情况

共现关系的考察主要涉及列举子项的形式和数量、共现的标点符号、共现的后助式列举类词等，主要通过索引行来观察，以充分反映母语者与学习者的语言使用差异。

3.2.1 列举子项形式及数量

such as 及 for example 引导的列举子项形式分为短语和句子，判断标准为是否含有谓语动词。列举子项的数量分为一项和多项。见图1和图2。

图1表明，英语本族语者 such as 后的列举子项主要为短语，for example 后主要为句子，两者比例都在 97% 左右。中国学习者基本保持一致，但比例都稍稍低于本族语者，考试环境下则更为接近本族语者的比例。说明学习者在考试环境中对 such as 和 for example 后列举子项的形式更为注意。

图2显示本族语者 such as 之后可以出现一个或多个列举子项，二者比例约为 4 : 6，而学习者 such as 之后一个和多个列举子项的比例约为 2 : 8；本族语者 for example 后面主要为一个列举子项，一个与多个列举子项的比例为 9 : 1，而中国学习者的这个比例为 8 : 2，即中国学习者均过多使用多个列举子项。其原因可能是，过去的描述让学习者误以为 such as 和 for example 引导一个与多个列举子项



的数据分布存在一定互补关系,“for example 表示‘例如’时,一般只以同类事物或人中的‘一个’为例”,“such as 表示‘例如’时,通常要以同类事物或人中的‘多个(两个或两个以上)’为例”(马懿 2021: 95),而通过对本族语者的数据统计我们发现,这里描述的“一般”和“通常”比例是完全不同的。

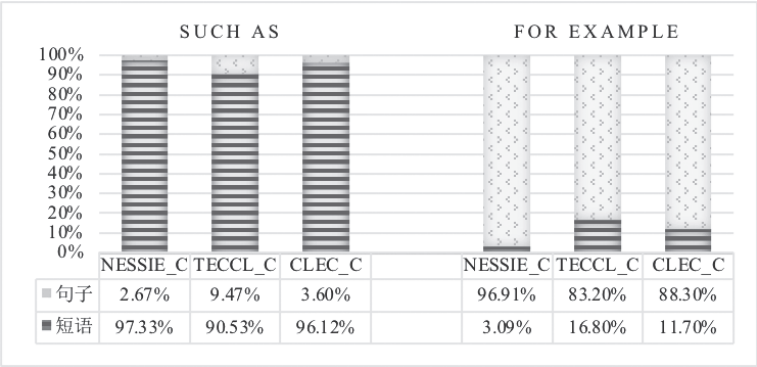


图1 列举子项形式<sup>5</sup>

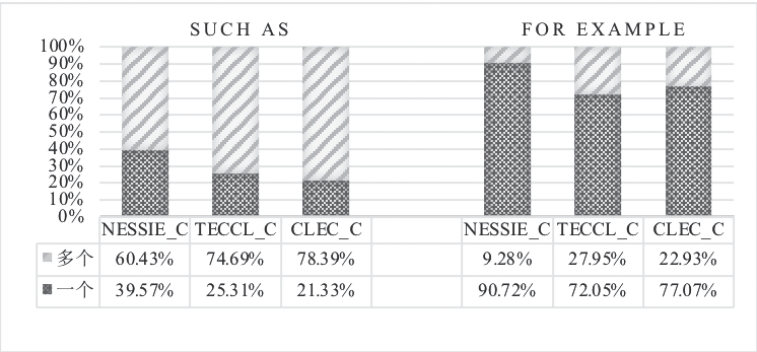


图2 列举子项数量

综合图1、图2的信息,中国学习者 such as 和 for example 之后列举子项的形式和数量跟母语者基本保持一致,但是由于过去的语法描述较为模糊,不能给学习者呈现具体的数据,导致学习者在比例的把握上有所偏差。值得注意的是,考试环境并没有加剧这种偏差。

3.2.2 标点符号

与标点符号共现是指 such as 和 for example 左右两侧紧邻标点符号的情况。根据语料, such as 和 for example 两侧出现的标点符号包括逗号、冒号、句号及其他。我们首先观察 such as 及 for example 左侧的标点情况(见图3)。

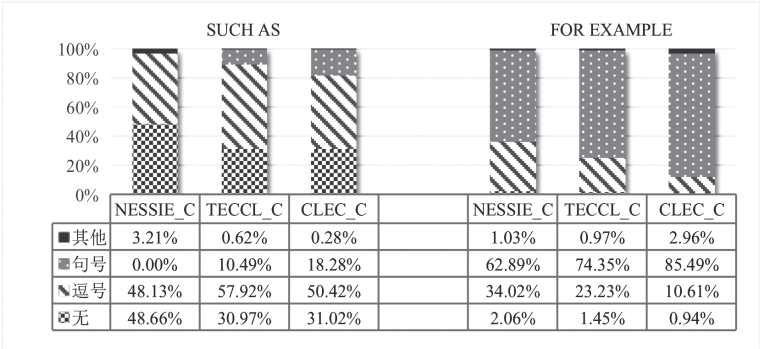


图3 左侧标点<sup>6</sup>

分析可知，英语本族语者所用such as左侧不与句号共现，一半左右不使用标点符号，而for example左侧主要与句号共现，占到60%以上。而中国英语学习者不同，两个语料库的数据一致显示：such as左侧不使用标点符号只占30%，10%以上与句号共现，存在与句号共现过多的问题；for example左侧标点符号则更倾向于使用句号，占到70%以上，与句号共现过多。学习者such as和for example左侧都与句号共现过多，并且在考试环境中比例会进一步提高。

图4则反映了such as及for example右侧标点使用情况。图4表明，英语本族语者98.40%的such as右侧没有标点，而超过90%的for example右侧有逗号出现。结合图3左侧的标点情况，说明such as与列举项联系较为紧密，两侧基本没有标点符号，for example使用时一般通过停顿将其与列举项隔开。学习者的右侧标点符号分布基本能够与母语者保持一致。需要注意的是，本族语者such as和for example右侧都没有冒号出现，而中国学习者在两个语料库中都有少量实例。

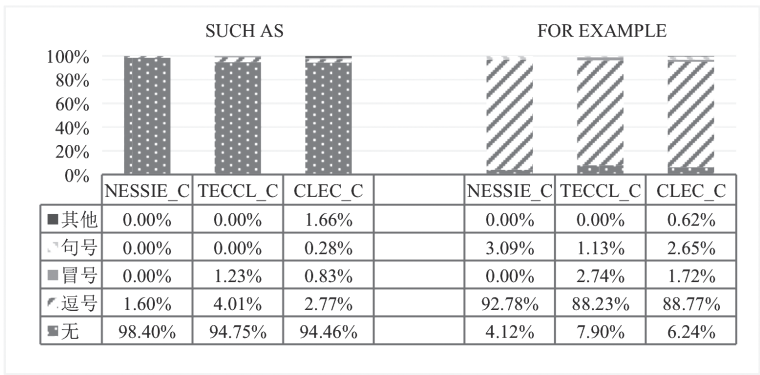


图4 右侧标点

3.2.3 后助式列举类词

后助式列举类词指除了语气式列举类词外，用于列举子项后面表示列举关系的列举类词（徐敏 2010），例如and so on，and so forth，etc和and others等。根据语料，本族语者such as和for example列举子项之后很少出现这些后助式列举类词，但中国学习者会使用，详见表4。

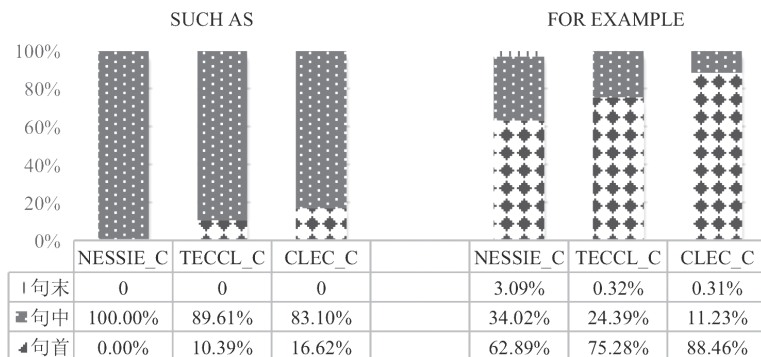
表4 such as和for example与后助式列举类词共现对比

共现的后助式 列举类词		NESSIE_C		TECCL_C		CLEC_C	
		原始 频次	标准 频次	原始 频次	标准 频次	原始 频次	标准 频次
such as	and so on	0	0	191	125	79	187
	etc.	2	10	41	27	47	111
	and so forth	0	0	3	2	5	12
	among (many)						
	others/and other						
	things/and the like/ and others	1	5	3	2	0	0
	总计	3	15	238	156	131	310
for example	and so on	0	0	30	20	24	57
	etc.	0	0	8	5	13	31
	and so forth	0	0	1	1	3	7
	among (many)						
	others/and other	0	0	0	0	1	2
	things/and the like/ and others						
	总计	0	0	39	26	41	97

我们发现学习者such as与后助式列举类词连用的频数多于for example，并且考试环境中（CLEC\_C）这种连用更为频繁。戴卫红（2001）认为such as后可与and so on连用，黄龙旺、龚汉忠（2008）则指出such as及for example不可与and so on或etc.连用，可以推断，such as和for example此类偏误可能与前人对用法描述的不一致有关，后文将加以验证。

3.3 句中位置

句中位置指句首、句中或句末。句首标志为单词首字母大写，或分号之后；句末的标志为其后紧邻句号、问号或感叹号；其余看作句中位置。such as和for example在句中位置的分布信息见图5。

图5 位置分布<sup>7</sup>

由图5可知,本族语者such as只在“句中”出现,而中国学习者such as“句首”出现比例为10%以上,考试环境下比例更高,为16.62%。本族语者for example主要出现在“句首”位置,占60%左右,“句中”位置占34%;而中国学习者for example“句首”出现比例过高,均在70%以上,考试环境下这一比例更是高达88%。这说明中国学习者基本掌握了such as和for example在句中的位置,但存在“句首”比例过高的问题,且考试环境下这一问题会进一步加重。

综上,通过数据对比,我们发现中国学习者主要存在如下问题。第一,从使用频次来看,中国学习者such as和for example使用偏少,但考试环境下for example的使用明显高于母语者,存在过度使用的现象。第二,从共现情况来看,在列举子项数量上,such as和for example均过多引导多个列举子项;在标点符号上,such as和for example左侧都与句号共现过多,而二者右侧共现情况基本与英语本族语者一致;such as和for example还存在与后助式列举类词共现过多的问题。第三,从句中位置来看,such as和for example均在句首使用过多。需要注意的是,such as和for example句首使用过多的问题,进一步印证了such as和for example左侧与句号共现过多。学习者语料库such as左侧句号比例高于句首比例的主要原因是学习者标点符号使用不当及句首首字母未大写,for example在学习者使用中也存在此类情况。在考试环境中,such as和for example与后助式列举类词共现的问题、句首使用过多的问题都进一步加重。

## 4 差异原因

本节主要从内外两方面分析中国学习者such as、for example使用特点的成因。虽然导致学习者出现上述差异的原因错综复杂,但是根据母语迁移规律以及对词典、教材教辅的调查结果,我们认为以下因素在一定程度上会影响学习者对such as和for example的使用。

## 4.1 内部原因

### 4.1.1 母语负迁移

二语习得中,母语迁移不可避免,在对第二语言的语义层面进行加工时,其对应的母语语义层面会被自动激活(孟迎芳等 2016)。其中,母语负迁移主要指母语知识与规则会对目的语学习产生一定干扰(高燕 2007)。研究发现,中国学习者 such as、for example 与后助式列举类词共现造成的冗余,以及“句首”过多使用,一定程度上受到了汉语母语的干扰。

汉语中“例如”“比如”等词的语义及功能与英语中 such as、for example 有一定对应关系。但不同于英语,汉语的多种列举类词形成了固定组合模式,徐敏(2010)将其分为 A+B、A+C、B+C、A+B+C 四类(A 为前连式列举类词,指用于列举子项前面表示列举关系的列举类词;B 为语气式列举类词,指用于列举子项后面表示列举关系的“呀,啊”等词;C 为后助式列举类词)。而英语中没有对应的语气式列举类词,所以对学习者影响最大的为 A+C 式组合。参见以下例句<sup>8</sup>。

(1) 西方有些著名的美学家,例如贺拉斯、布瓦罗、狄得罗、莱辛、泰纳和别林斯基等人,都同时是文艺批评家。(《美学》朱光潜)

汉语中 A+C 式的组合,一定程度上会迁移到英语习得中去,学习者使用 such as 及 for example 时也会在列举子项末加上 and so on、etc. 等后助式列举类词,如例(2)和例(3)。

(2) *For example*, you can use double-sided paper, use less snack box and disposable chopsticks, and so on. (TECCL\_C)

(3) They would bring various food to forefathers [fm1, -], *such as* roast little pigs, fruits, rice, boiled chicken, wine *and so forth*. (CLEC\_C)

通过国家语委语料库对几个典型前连式列举类词(黄紫墨 2020)检索发现,“例如”与“等”“等等”“一类”“之类”“此类”“什么的”等主要后助式列举类词共现 846 例,“比如”共现 137 例,“譬如”共现 27 例,“诸如”共现 196 例,分别占所检索语料总体的 19.21%、12.91%、6.94% 和 80.33%(见附录 1),占比较高,而英语本族语者 such as 和 for example 列举子项后基本不会出现后助式列举类词。可知,学习者将汉语使用习惯推广到 such as 和 for example 的使用上,造成了英语词汇使用的冗余。

此外,中国学习者 such as 和 for example “句首”使用过多(左侧与句号共

现过多)也与汉语表达有关,也是母语负迁移的结果。汉语中表达列举意义的“例如”“比如”“譬如”等均倾向于出现在句首。我们利用国家语委语料库在线检索发现,“例如”出现在句首的比例为74.28%，“比如”出现在句首的比例为69.18%，“譬如”出现在句首的比例为64.52%。学习者将汉语的使用习惯推广到such as和for example上,由此造成了二者句首使用过多的现象。

#### 4.1.2 学习者策略

TECCL\_C中such as和for example使用偏少,因为该数据为学习者大学阶段的学业作文,涉及的题目较为宽泛,学习者用词也比四六级作文宽泛很多。此外,学习者有时间和机会查阅资料,改正失误,也可以充分利用外部资源,引用他人原文或改写,尽量避免使用such as和for example等中学阶段就已经学过的普通短语,而去追求更加新颖和高级的表达,因此such as和for example出现偏少。而在四六级考试作文中(CLEC\_C),为避免出错,考生一般会采取回避策略,头脑中优先调取较有把握的、相对简单的such as和for example,而且for example的使用更加自由灵活,学习者会更青睐,造成for example的使用频数远远超过本族语者。

另外,考试环境中,考生迫于时间和心理压力,“言语失误明显偏高”(杨惠中等 2005: 160),一定程度上解释了本文观察到的现象,学习者在考试环境中such as和for example与后助式列举类词的共现更为频繁,句首出现比例更高,等等。“考试是一种非常态的语言运用,我们由此可以得出结论,如果考试的信度和效度得到保证的话,考生的实际语言能力要高于考试所反映的语言能力。”(杨惠中等 2005: 160)因此,由母语负迁移造成的学习者语言特征,考试环境中可能会进一步加重。这也说明考试环境对学习者的语言使用有所影响,“试卷命题作文所反映的是考生头脑中真正属于考生自己的表达性语言能力,与领会式语言能力比较,语言宽度要小得多。”(杨惠中等 2005: 160)

## 4.2 外部原因

外部原因主要指词典、教材教辅等原因。因为such as及for example为中国学习者初中阶段习得(李进才 2021),所以我们收集了35本英汉词典<sup>9</sup>以及八年级上册英语教材与16本教辅书<sup>10</sup>,考察了such as及for example的辨析情况,发现中国学习者的使用不当与词典及教材教辅收词不当、释义模糊、缺少举例和辨析有密切关系。

第一,教材未收录such as。外语教学与研究出版社的中学英语教材并未收录such as, for example在八年级上册Module 11明确收录,学习了for example之后,such as才会顺带辨析。

第二,词典释义不具体,甚至互释(见附录2)。35本词典中,同一词典for example及such as释义基本相同。《朗文当代高级英语辞典:英英·英汉双



解：第6版》的解释已较为全面，“*for example*: used before mentioning a specific thing, person, place etc in order to explain what you mean or to support an argument 例如，举例来说”（英国培生教育出版集团 2019: 865），“*such as*: used when giving an example of something 像，诸如，例如〔用于举例〕”（英国培生教育出版集团 2019: 2591），在词语辨析中又将*for example*解释为“used when giving an example”（英国培生教育出版集团 2019: 865），这种描述难以体现二者差异。《牛津高阶英汉双解词典第9版》等五本词典甚至直接用*for example*解释*such as*，容易误导学习者，认为二者用法一致。

第三，教材辨析过少、不准确。16本教辅中，五分之一未对二者进行辨析；辨析中最多的为位置信息，大多教辅只写*for example*“可置于句首、句中、句末”（如杨文彬 2021: 188），没有数据比例呈现；大多教材只写*such as*“常用在被列举的人或事物之前”（如李进才 2021: 214），并未对具体位置进行标注。

第四，举例偏少，且不够典型。35本词典，只有18本词典举例，得到*such as*例句35个，*for example* 27个。大多词典仅举1例，少数大部头词典举了2—3例。统计例句发现，*for example*引导短语的情况占51.85%，明显高于英语本族语者（NESSIE\_C: 3.09%），就列举子项数量而言，例句中*for example*后引导多个列举子项占比为29.63%，也高于英语本族语者（NESSIE\_C: 9.28%）；就共现标点符号而言，例句中*for example*右侧与逗号搭配占比62.96%，低于英语本族语者（NESSIE\_C: 92.78%），无标点符号占比22.22%，高于英语本族语者（NESSIE\_C: 4.12%）。说明词典中例句的选取并不能反映英语本族语者的表达习惯。而且，《尖子生学案：外研社版·八年级英语》指出*such as*表示列举时可与*and so on*搭配，《学生英汉词典：精编大字本》及《新英汉双解大词典（插图本）》对*such as*进行讲解时，列举了以下例句。*such as*的列举子项后出现了*etc.*，无疑对学习产生了误导。

（4）He knows six languages, such as Chinese, Russian, etc.

## 5 教学启示

由上可知，中国学习者在书面表达中*such as*及*for example*的使用与英语本族语者相比有诸多差异。中国学习者主要通过教师、教材、词典等获取词项信息，因此，改进教学过程，修订词典、教材教辅，对于学习者学习至关重要。本研究给我们以下启示。

一是外语教学过程中教师可以利用语料库，更好地进行预警。在后疫情时代，外语教学因跨时空、跨地域而形成了新的教学形式（黄立鹤 2021），教师提

高语料库素养（何安平 2022）更是时代所需。语料库的呈现方式可以充分调动学习者的积极性，强化其记忆力，在真实语境中强化词项学习，从而减少上述使用不当的形成。由上文可知，母语负迁移对学习者正确使用 such as 及 for example 造成了很大障碍。因此，教师在讲解 such as 及 for example 的用法时，可以利用语料库进行预警，以减少学习者母语负迁移的影响。主要操作如下：一方面，利用英语母语者语料库进行教学设计及课堂讲授。老师在教学设计时登录英语母语语料库（NESSIE Corpus 2nd release, NESSIEv2, <http://114.251.154.212/cqp/nessie2/>），检索含有 such as 及 for example 的典型例句制作语境共现，从中选取 20 个适合所教学生实际水平（中学或大学）的例句，选取依据为英语本族语者使用习惯（如 for example 在句首、句中、句末位置的比例为 6 : 3.5 : 0.5）。制作关键词居中的 PPT 等将例句加以呈现，设计任务指令引导学生观察归纳不同颜色凸显的两侧标点符号、列举项形式与数量及句中位置等情况，思考两者是否可以替换使用，学生小组讨论归纳用法及两者异同；同时，引导学生观察归纳列举子项末的搭配，列举子项之后很少出现后助式列举类词，老师对此进行预警。另一方面，利用学习者语料库进行病句修改等练习及测试。重复练习可以帮助学生巩固用法，因此教师有必要利用语料库加强学生在此方面的练习。在中国英语学习者语料库中选取 5—10 个 such as 及 for example 与标点符号共现不当、句中位置使用不当及与后助式列举类词共现的病句，学生指出偏误并加以修改。此外，为了强化学生的应用能力，学生可以通过情景造句进行产出练习，触发学生多模态互动（黄立鹤 2021），小组内互相指出错误并加以修改，增加学习趣味性，减少畏难心理。

二是结合语料库修订词典及教材教辅。首先，such as 及 for example 作为前连式列举类词使用频率较高，是《大学英语教学大纲通用词汇表（1—4 级）》的基础词组，需要对其熟练掌握，因此教材有必要将二者全部收录。其次，增加 such as 及 for example 的辨析。词典和教辅对两者辨析并不明确，学习者容易混用。Rodríguez-Abruñeiras（2015）指出 such as 及 for example 作为举例标记在语义方面并不完全等同，for example 对列举子项强调意味较 such as 弱，更为中性，所以既要避免二者互释，突出两者差异，也需补充辨析内容，必要时提供具体数据。例如，从列举子项数量来看，such as 及 for example 均可引导一项或几项，such as 引导多项情况居多，占 60% 左右；for example 引导一项情况居多，占 90% 左右。such as 左侧无标点和逗号各占一半；for example 左侧一般为句号，占 60% 左右。数据信息能帮助学习者更好地理解二者异同。最后，增加并合理选取例句。举例示范有助于学习者对正确搭配有直观感受，应根据 such as 及 for example 的不同用法做到举例多样化，增加其趣味性。同时，例句选择要符合英语本族语者使用习惯。例如，避免选取 such as 和 for example 与后助式列举类词共现的例句。

## 6 结语

列举类短语 *such as* 和 *for example* 使用频繁，但国内对其关注较少，尤其缺乏针对中国英语学习者的定量研究，因此基于语料库的学习者使用考察对英语教学有重要的指导意义。本文基于大型语料库（NESSIE\_C、TECCL\_C及CLEC\_C），对比考察中国学习者与英语本族语者 *such as* 及 *for example* 的使用差异，将以往某些主观感觉和表面印象转化为具体的语言事实和数据，并提出合理化建议。研究发现，中国学习者对 *such as* 和 *for example* 的使用与本族语者存在差异。第一，*such as* 及 *for example* 使用偏少，但考试环境下 *for example* 的使用明显高于母语者，存在过度使用的现象。第二，*such as* 及 *for example* 左侧标点符号、列举子项数量、和后助式列举类词共现情况与英语本族语者均存在较大差异。第三，*such as* 及 *for example* 句首使用过多。具体分析，这些差异与母语负迁移、词典及教材教辅辨析不到位等有一定关系。据此，我们得到的教学启示是：教师应结合语料库进行预警，提高语料库素养，引导学生自主归纳 *such as* 及 *for example* 的用法；词典及教辅材料编写应符合英语本族语者使用习惯，确保辨析的全面性及数据化，做到举例的规范化、多样化。

本文从共选理论视角切入，以 *such as* 和 *for example* 为例在词汇和语法层面对共选理论进行了实践，将词汇与出现环境综合考察，可进一步加深对共选理论的理解，不仅有助于明晰 *such as* 和 *for example* 的用法，更可为其他类似固定短语句的性质与用法研究提供有益参考。本文将定量与定性分析相结合，弥补了前人对英语列举类词习得中缺乏定量研究的不足，对教师、词典及教材教辅提出可行性建议，有利于减少传统教学材料与英语母语者实际语用不匹配的情况，为英语教学及词典、教材教辅编纂提供了实证和数据支持。此外，本文基于CIA模型着重分析学习者使用特征，重新评估教学中 *such as* 及 *for example* 的教学内容及方法，明晰了 *such as* 及 *for example* 是否与后助式列举类词共现等争议，在语言教育领域有助于提高师生数字素养，具有应用价值。后续可在以下三方面继续开展工作：首先对英语中列举类词进行总体分析与研究，对英语列举词的常用构式进行探究，进一步扩展研究广度；其次，对于现代汉语与英语中列举类词进行深入的跨语言及历时研究，丰富研究角度；最后，开展中国英语学习者与其他母语背景学习者列举类词使用情况的对比分析，讨论母语对语言习得的影响，揭示人类习得机制的共性和个性，不仅是外语教学，更是认知领域的重要课题。

### 注释

- 1 相比卡方检验，基于似然比的检验克服了正态分布假设对分析罕见事件能力的限制，相对较小的样本也能产生良好的结果，对数似然率计算器由北京外国语大学中国外语与教育研究中心许家金教授于2009年研制，能够通过对数

似然率及P值计算来考察两个词汇的出现频数是否具有显著性差异。

- 2 标准频次=原始频次 ÷ 总形符数 \*1,000,000 (结果保留整数)
- 3 Log-likelihood (对数似然比) 大于3.83、6.64和10.83, 则表明该值在0.05、0.01和0.001的显著性水平上有意义。
- 4 Sig. (P值) 大于0.05则视为无显著性差异, 若其在0.05至0.01之间则存在显著性差异 (\*), 若其值小于0.01则存在极显著差异 (0.01与0.001之间标为 \*\*, 小于0.001为 \*\*\* )。“+”表明过多使用,“-”为使用不足。
- 5 其中, TECCL\_C有一例语料为“For example. . .”, CLEC\_C有一例语料为“Such as. . .”, 均无列举子项。
- 6 TECCL\_C中 for example左侧标点因四舍五入原因各项占比相加约等于1。
- 7 CLEC\_C有一例语料为“Such as. . .”, TECCL\_C有一例语料为“For example. . .”, 句中位置不定, 计算时剔除。
- 8 国家语委语料库是全面反映汉语母语使用状况的大规模平衡语料库, 于1991年12月由国家语言文字工作委员会提出立项, 1998年底建成, 被列为国家语委“九五”“十五”科研重大项目, 得到国家科技部“863”“973”计划多个项目支持, 语料选材类别广泛, 时间跨度大, 全库约为1亿字符, 例(1)和例(2)语料均从其现代汉语语料库中在线检索获得。
- 9 35本英汉辞典为上海外语教育出版社:《新牛津英汉双解大词典第2版》《朗文初阶英汉双解词典第3版》《外教社·柯林斯初级英语用法词典》; 广东人民出版社:《小学生多功能英语词典:彩图版》; 浙江教育出版社:《小学生图解英汉词典:多功能大字版》; 华语教育出版社:《学生英汉词典:精编大字本》《新英汉双解大词典(插图本)》《彩图版小学生英汉汉英词典》; 上海辞书出版社:《辞海版英汉双解词典》; 商务印书馆:《8000词英汉词典:双色大字本》《牛津高阶英汉双解词典第九版·缩印本》《牛津中阶英汉双解词典:第5版》《精选英汉汉英词典第5版》《牛津初阶英汉双解词典第四版》; 中国青年出版社:《学生实用现代英汉双解大词典》《学生实用英汉大词典第6版》; 四川辞书出版社:《50000词英汉双解词典第三版》《现代英汉词典》; 外语教学与研究出版社:《外研社·柯林斯英汉汉英词典:第三版》《牛津袖珍英汉双解词典:第11版》《外研社实用多功能英汉词典》《柯林斯初阶英汉双解学习词典:第3版》《外研社英汉多功能词典》《英汉小词典第二版》《外研社英汉小词典大字本第2版》《外研社·柯林斯学生实用英汉汉英词典:第三版》《朗文当代高级英语辞典:英英·英汉双解:第6版》; 商务印书馆国际有限公司:《80000词英汉词典第2版》《小学生多功能英汉词典:双色插图本》《学生实用全新英汉双解大词典》; 湖南教育出版社:《学生实用英汉双解大词典》《唐文新英汉词典:新版》《新概念英汉双解词典》; 吉林教育出版社:《学生实用英汉双解词典》; 吉林出版集团有限责任公司:《新编多功能英汉大词典》。

- 10 16本教辅为现代教育出版社:《新高中英语词汇》《教材解读·英语·八年级·上册》;湖南师范大学出版社:《高中英语词汇全解》《图解高中英语词汇》;首都师范大学出版社:《高中英语必考词》;辽宁教育出版社:《新高考英语必备》《学霸同步笔记·英语八年级:WY版》;陕西人民教育出版社:《中学教材全解·八年级英语·上:外语教研版》;北京教育出版社:《1+1轻巧夺冠课堂直播·英语八年级·上》;河北少年儿童出版社:《星推荐·涂教材·初中英语八年级·上册:WY》;南京师范大学出版社:《教材帮·初中英语八年级上册:WY》;河北教育出版社:《七彩课堂·英语:外研版·八年级·上册》;吉林人民出版社:《尖子生学案:外研社版·八年级英语》《新教材完全解读:外研社版·八年级英语》;开明出版社:《教材划重点·英语八年级·上:WY》;浙江教育出版社《全易通·初中英语八年级·上》。
- 11 薛熙哲, 2015, 中国学生万篇英语作文语料库(V1.1)(Ten-thousand English Compositions of Chinese Learners, Version 1.1, 简称The TECCL corpus)。

### 参考文献

- GRANGER S. The computerized learner corpus: a versatile new source of data for SLA research [C]//GRANGER S. Learner English on computer. London: Longman, 1998a: 3-18.
- HYLAND K. Applying a gloss: exemplifying and reformulating in academic discourse [J]. Applied Linguistics, 2007, 28(2): 266-285.
- PAQUOT M. Exemplification in learner writing: a cross-linguistic perspective [C]// MEUNIER F, GRANGER S. Phraseology in foreign language learning and teaching. Amsterdam: John Benjamins, 2008: 101-119.
- RODRÍGUEZ-ABRUÑEIRAS P. Exemplifying, markers in English: synchronic and diachronic considerations [D]. Santiago de Compostela: University of Santiago de Compostela, 2015.
- RODRÍGUEZ-ABRUÑEIRAS P. The history of for example and for instance as markers of exemplification, selection and argumentation (1600-1999) [J]. Atlantis, 2021, 43(1): 133-153.
- SINCLAIR J. Corpus concordance collocation [M]. Oxford: Oxford University Press, 1991.
- SINCLAIR J. The search for units of meaning [J]. Textus, 1996, 9(1): 75-106.
- SINCLAIR J. Trust the text [M]. London: Routledge, 2004.
- TRIKI N. Exemplification in research articles: Structural, semantic and metadiscursive properties across disciplines [J]. Journal of English for Academic Purposes, 2021, 54: 1-13.



- 陈仁祯. such as 与 for example 的用法比较[J]. 第二课堂(A), 2014(9): 23-24.
- 戴卫红. for example 还是 such as [J]. 英语辅导(高中年级), 2001(8): 11.
- 高燕. 对外汉语词汇教学[M]. 上海: 华东师范大学出版社, 2007.
- 桂诗春, 杨惠中. 中国学习者英语语料库[M]. 上海: 上海外语教育出版社, 2002.
- 郭书彩, 李娜, 徐瑞华. 基于语料库的学习者例证词“for example”对比研究[J]. 河北大学学报(哲学社会科学版), 2015(1): 103-108.
- 何安平. 中小学英语教师语料库素养提升路径研究[J]. 中小学数字化教学, 2022(6): 5-9.
- 黄立鹤. 多模态范式与后疫情时代的外语教学[J]. 当代外语研究, 2021(1): 75-85.
- 黄龙旺, 龚汉忠. 英文论文中“such as, for example, e.g., i.e., etc., et al.”的用法分析[J]. 编辑学报, 2008(2): 124.
- 黄紫墨. 现代汉语列举范畴研究[D]. 南京: 南京师范大学, 2020.
- 金子铭. 举例说明 for example [J]. 考试与评价(英语八年级专刊), 2018(8): 16.
- 李进才. 教材解读(英语八年级上册)[M]. 北京: 现代教育出版社, 2021.
- 梁石. 九年级 Units 1-2 易混词语辨析[J]. 中学英语之友(新教材初三版), 2012(7): 20-22.
- 马懿. for example 与 such as 用法辨析及教学建议[J]. 课程教材教学研究(中教研究), 2021(Z1): 95-96.
- 孟迎芳, 林无忌, 林静远, 等. 双语即时切换下非目标语言语音和语义的激活状态[J]. 心理学报, 2016(2): 121-129.
- 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, 74.
- 卫乃兴, 陆军. 基于语料库的二语学习研究述评: 范式变化与挑战[J]. 外语教学, 2018(5): 47-53.
- 邢红兵, 辛鑫. 第二语言词汇习得的中介语对比分析方法[J]. 华文教学与研究, 2013(2): 64-72.
- 徐敏. 现代汉语列举类词语考察[D]. 上海: 上海师范大学, 2010.
- 薛熙哲. 中国学生万篇英语作文语料库(V1.1)(Ten-thousand English Compositions of Chinese Learners, Version 1.1, 简称 The TECCL corpus), 2015.
- 杨惠中, 桂诗春, 杨达复. 基于 CLEC 语料库的中国学习者英语分析[M]. 上海: 上海外语教育出版社, 2005.
- 杨文彬. 教材划重点(英语八年级上 WY)[M]. 北京: 开明出版社, 2021.
- 英国培生教育出版集团. 朗文当代高级英语辞典(英英·英汉双解)[Z]. 6版. 北京: 外语教学与研究出版社, 2019.



附 录1

汉语前连式列举类词与后助式列举类词共现

		等（等）	之类	一类	什么的	此类	总计
例如	频数	834	12	0	0	0	846
	占比	18.93%	0.27%	0	0	0	19.21%
比如	频数	134	1	0	2	0	137
	占比	12.63%	0.09%	0	0.19%	0	12.91%
譬如	频数	24	2	0	1	0	27
	占比	6.17%	0.51%	0	0.26%	0	6.94%
诸如	频数	142	11	4	1	38	196
	占比	58.20%	4.51%	1.64%	0.41%	15.57%	80.33%

附 录2

such as 及 for example 词典例句统计

			频数	占比
such as	列举子项形式	短语	28	80.00%
		句子	7	20.00%
	列举子项数量	一个	17	48.57%
		多个	18	51.43%
	句中位置	句首	0	0.00%
		句中	35	100.00%
		句末	0	0.00%
		逗号	0	0.00%
	标点符号	无	35	100.00%
		句号	0	0.00%
	列举子项形式	短语	14	51.85%
		句子	13	48.15%
for example	列举子项数量	一个	19	70.37%
		多个	8	29.63%
	句中位置	句首	5	18.52%
		句中	18	66.67%
		句末	4	14.81%
		逗号	17	62.96%
	标点符号	无	6	22.22%
		句号	4	14.81%

通信地址：264209 山东省威海市 山东大学（威海）文化传播学院

# 意大利语语料库及其应用研究

北京外国语大学 谭钰薇 余丹妮

**提要：**意大利是语料库建设及应用的先驱地之一，其语料库语言学自成体系且蓬勃发展。以高校为中心的各大研究团体相互合作，创建出类别多样的语料库。本文梳理意大利语语料库语言学的发展脉络以及主要的研究机构与团队，介绍主要的意大利语开源语料库和基于它们的应用研究，为国内意大利语语言学研究及语料库建设与应用研究提供参考。

**关键词：**意大利、语料库语言学、基于语料库研究、意大利语

## 1 引言

语料库语言学的发展得益于语料库电子化，意大利是应用该技术的先驱地之一。作为语文学传统的根植之地，意大利自20世纪50年代起便率先将新兴的信息技术应用于语文学研究，实现了语料库电子化，推动了20世纪60年代语料库语言学形成系统学科的进程。意大利语语料库建设蓬勃发展，应用研究成果丰富，而国内文献对此却仍然鲜有涉及。

国内对于外语语料库的研究以英语为主，目前已有对英语（许家金 2019）、西班牙语（赵冲、许家金 2023）、法语（田园 2014）、俄语（李勤、常翔宇 2018）等通用语种语料库建设与发展历程的综述论文，但尚未有任何介绍包括意大利语在内的非通用语种语料库语言学的文章。这一现状与国外非通用语种语料库研究的繁荣状况不甚相符。意大利语语料库呈现出分类繁多、应用广泛等特点，可以作为国内意大利语语料库研究及语料库建设研究的重要借鉴。

近年来，我国学界也开始关注意大利语语料库的创建。北京外国语大学研究团队于2020年创建了汉意意汉双向文学平行语料库（余丹妮 2020），又于2022年创建了当代意大利语语料库itGLOBE（喻儒辰等 2023）和意大利语新闻语料库ItalianWac。另外国内意大利语学界虽然已有少量基于语料库的研究（董丹 2019；余丹妮、张虢 2022），但仍处于起步阶段。相比之下，意大利学界的相关研究成果丰硕，涉及意大利语语言学的方方面面。对其进行介绍，可为国内意大利语语

1 余丹妮为本文通讯作者。

作者贡献：

谭钰薇：数据收集、数据分析、初稿撰写、字数占比（60%）、修改润色。

余丹妮：选题构思、研究方法、讨论结论、字数占比（40%）、修改润色。

言学界提供启示，有助于推动着眼国内教学需求与社会需求的研究。

本文介绍意大利语语料库的建设、发展与应用研究，主体内容分为两节。第一节梳理意大利语语料库发展的历史脉络以及当今发展状况，介绍意大利本土主要的语料库研究机构与团队，以及可以公开访问的开源语料库；第二节对开源语料库在不同语言学领域的应用进行引介，以具体案例阐述意大利语语料库如何用于解决不同的语言学研究问题。

## 2 意大利语语料库发展

### 2.1 意大利语语料库建设的起源与发展

意大利是最早应用计算机技术研发语料库的国家。早在1949年以前，意大利耶稣会布萨神父（Roberto Busa）就萌生出创建电子语料库的想法，他联系国际商业机器公司（IBM）寻求技术支持，在米兰创办文学分析自动化中心。1967年，布萨神父牵头完成《托马斯索引》（*Index Thomisticus*），其中收录了118篇中世纪神学家托马斯·阿奎纳的拉丁语作品，以及61篇相关作品，规模约1,100万词（Busa 1973）。意大利语语料库的索引和词汇搭配功能最早则可追溯到比萨国立大学电子计算中心于1963年建立的《神曲》索引搭配（Cresti & Moneglia 2016: 591）。

语料库的发展可以划分为三个主要阶段（Bonelli & Sinclair 2006: 208）。第一阶段为20世纪60到80年代，该阶段的原始材料基本是纸质材料，建造语料库时需要逐词录入电脑，耗时费力，难以完成100万词以上规模的语料库。该阶段具有开创性和奠基性的语料库包括1971年比萨国立大学电子计算中心为编写意大利语频率词典建立的首个意大利语笔语参考语料库（Corpus LIF），以及语言学家斯坦默约翰（Harro Stammerjohann）1965年起研制的首个意大利语口语语料库<sup>1</sup>（Corpus Stammerjohann）。第二阶段为20世纪80年代到21世纪初，扫描技术的应用使语料库规模逐渐扩大，可达2,000万词以上。该阶段英美语料库迅速发展，意大利在语料库界虽然并未处于中心地位，但也贡献了大量具有独特研究意义的语料库。第三阶段始于21世纪，互联网为语料库提供了无限量的电子语料，超大规模语料库应运而生，个人也能够根据具体研究目标制作中小型专业语料库。

### 2.2 意大利语语料库语言学主要研究机构与团队

意大利语语料库语言学的研究机构与团队数量众多，常以高校和研究院为中心，相互合作与影响。各团队可能建立类似或相同种类的语料库，但其研究方法或侧重点往往有所不同。以下参考《意大利语语料库导论》（Cresti & Panunzi

2013)、秕糠学会(Accademia della Crusca)语料库数据库<sup>2</sup>以及“说意大利语”门户网站<sup>3</sup>提供的相关信息展开介绍。

意大利最早开始制作语料库的研究机构是意大利国家研究委员会计算机语言学研究院<sup>4</sup>。该研究院在研制语料库方面成果丰硕,如1971年的现代意大利语频率词典语料库(Corpus LIF)、1991年的意大利语参考语料库(Italian Reference Corpus)、1993年的扎尼凯利意大利语文学语料库(Letteratura Italiana Zanichelli, 简称LIZ)、1997年的外语口语课程语料库(PARallèle Oral en Langue Etrangère, 简称PAROLE)、2007年的语言学语境语料库(Corpus Linguistics in Context, 简称CLiC)。与CNR研究院几乎同时启动的还有比萨高等师范学院的语言学实验室。该实验室由语言学家南乔尼(Giovanni Nencioni)牵头,除建立语料库外,还涉及语音学、音系学、形态语言学、神经语言学等广泛研究领域。2005年,语言实验室同CNR研究院合作推出意大利语笔语词频语料库与词典(Corpus e Lessico di Frequenza dell'Italiano Scritto, 简称CoLFIS)。

在南乔尼的推动下,佛罗伦萨大学文学与哲学院于1985年成立了意大利语语言实验室,领头人为克雷斯蒂(Emanuela Cresti)和莫内利亚(Massimo Moneglia)教授。该实验室主要研制口语语料库,成果包括2005年建成的罗曼语族-意大利语口语参考语料库(C-ORAL-ROM Italia)、2006年基于斯坦默约翰的意大利口语语料库完善而成的佛罗伦萨口语语料库(Corpus LABLITA),以及2013年联合众多高校建立的动态网络意大利语语料库(Risorse Dinamiche dell'Italiano in Rete, 简称RIDIRE)。

罗马智慧大学数字人文学科奠基人吉里奥齐(Giuseppe Gigliozi)于1993年创建了意大利最早的文学文本信息应用研究中心——文学信息文本协作研究中心。该中心收集不同类别的文本材料,基于电子档案建立数字图书馆和语料库,推出在线意大利语文本、意大利图书馆、网络意大利语语料库、意大利议会口语语料库,以及政治与议会语言可读性-词汇和句法语料库等语料库。语料库的词频分析和语料筛选功能是词典编撰的有力支持,罗马智慧大学同时也是语料库词典学的研究中心。20世纪90年代起,罗马大学德毛罗(Tullio De Mauro)学派基于语料库编写的词典对普通语言学作出了奠基性贡献。该学派借助相应语料库研制的词典有《千禧年词汇:意大利语计算机词典》(*Il vocabolario del 2000: Vocabolario Elettronico della Lingua Italiana*, VELI, 1989)(Italia IBM 1989)、《意大利语口语词频词典》(*Lessico di frequenza dell'Italiano Parlato*, LIP, 1993)(De Mauro et al. 1993)、《意大利语语用大词典》(*GRAnde Dizionario ITaliano dell'uso*, GRADIT, 1999)(De Mauro 1999)以及《二十世纪文学语言第一宝库词典》(*Primo Tesoro della Lingua Letteraria del Novecento*, 2007)(De Mauro 2007)等。

意大利另一所享有盛誉的高校都灵大学同为意大利语语料库建设的中心。该

大学的语料库语言学研究团队活跃于20世纪末到21世纪初,由马雷洛(Carla Marella)教授牵头于2003—2004年期间研制了大量笔语语料库,如古意大利语语料库(Corpus Taurinense)、意大利都灵大学学术文本语料库(Athenaeum)、皮埃蒙特大区新闻报刊文本语料库(Corpus Seguisinum)、“权利之羹”意大利语法律语料库(Jus Jurium)、意大利语学习者类型语料库(Varietà Apprendimento Lingua Italiana Corpus,简称VALICO)和意大利语母语者类型配对语料库(Varietà di Italiano di Nativi Corpus Appaiato,简称VINCA)。此后该团队逐渐将研究重心移至网络语料库,分别于2008年和2012年建立了新闻组用户网络语料库(Newsgroups UseNet Corpora,简称NUNC)和在线新闻语料库(Varietà Alte di Lingue Europee in REte,简称VALERE)。

博洛尼亚大学应用语言学跨学科中心是意大利语语料库语言学最大的研究中心之一。该研究中心在法弗雷蒂(Roma Rossini Favretti)教授的领导下研制了一系列功能强大且使用广泛的语料库,其中包括1997年起研制的博诺尼亚法律法规范英平行语料库(Bononia Legal Corpus,简称BoLC)、1998年起研制的意大利语笔语参考语料库/意大利语笔语动态语料库(Corpus di Riferimento dell'Italiano Scritto/Corpus Dinamico dell'Italiano Scritto,简称CORIS/CODIS)、2006年推出的意大利笔语历时参照语料库(DiaCORIS)。在和多方研究机构的合作下,博洛尼亚大学还建立了共和国报新闻语料库(corpus *La Repubblica*)、意大利语网络语料库(Web as Corpora-Italiano, ItWac)和派萨网络语料库(Piattaforma per l'Apprendimento dell'Italiano Su copia Annotati,简称PAISÀ)等语料库。

那不勒斯腓特烈二世大学的语料库语言学研究——信号分析与合成跨系研究中心同样自20世纪末21世纪初开始活跃。该研究中心由语言学家莱奥尼(Federico Albano Leoni)教授牵头,主要成果有1999年建成的意大利语口语变体语料库(Archivio delle Varietà di Italiano Parlato,简称AVIP)、2001年的意大利语口语正字转写语料库(Archivio di Parlato Italiano Trascrizione Ortografica,简称API),以及2003年建成的意大利口语笔语语料库(Corpora e Lessici dell'Italiano Parlato e Scritto,简称CLIPS)和意大利语口语语料库(Italiano PARlato,简称IPAR)。

萨莱诺大学的欧洲语言研究观察实验室自21世纪初成立起即活跃于语料库语言学研究等领域。在沃盖拉(Miriam Voghera)教授的领导下,实验室于2006年推出收录古今口笔意大利语的佩内洛佩语料库(corpus PENELOPE),2015年又以《意大利语口语词频词典》语料库(Corpus LIP)为基础制成LIP之声口语语料库(La Voce del LIP,简称VoLIP)。

意大利的外国人大学通常会发挥本校语言教学资源优势制作习得语料库,如锡耶纳外国人大学的外国人意大利语口语语料库(Lessico Italiano Parlato di



Stranieri, 简称LIPS)、意大利语二语习得语料库(Archivio Digitale di Italiano L2, ADIL2)。另有佩鲁贾外国人大学的意大利语二语习得学习者语料库(Corpus di Apprendenti di Italiano L2, CAIL2)和中国学生意大利语(口语和笔语)语料库[Corpus of Chinese Learners of Italian (written and spoken)]。

### 2.3 主要的意大利语开源语料库

自20世纪60年代起,意大利语语料库建设在国家研究委员会与各高校研究中心的推动下蓬勃发展。意大利语语料库现有类型多样,包括通用/专用、共时/历时、口语/笔语、本族语/学习者、单语/平行语料库。随着21世纪初网络语料库的出现,意大利语语料库呈现出规模更大、模态丰富、专用化强的特点。不过,目前可供公开访问和查询的意大利语开源语料库数量仍然有限,以下对主要的意大利语开源语料库进行介绍。

当前最具代表性的意大利语开源笔语语料库是博洛尼亚大学的意大利语笔语参考语料库/意大利语笔语动态语料库<sup>5</sup>(CORIS/CODIS),该语料库是意大利语首个一般现代笔语的参考语料库,其规模相当于BNC语料库(Cresti & Panunzi 2013)。目前,该语料库体量已达1.5亿词,每3年更新一次,下分新闻、小说、学术文章、法律行政文本、混杂文集、时效文本6个子语料库。开源口语语料库中,最常用的是《意大利口语词频词典》语料库(LIP)。该语料库最初于90年代在德毛罗学派的推动下建成,包含取材于4个不同城市的录音,共计60小时,有当面对话、电话对话、采访和辩论、独白、广播等口语类型,目前可通过意大利口语数据库(Banca Dati dell'Italiano Parlato, 简称BADIP)和LIP之声口语语料库<sup>6</sup>进行访问和检索。另外,博洛尼亚大学和都灵大学合作开发的“谁说”语料库(KIParla)是当前最新颖和最实用的开源口语语料库之一(Goria *et al.* 2019),该语料库包含100余小时的录音,主要特点是在收集语料过程中重点考量语域,根据地域、年龄、教育程度与发言场合等进行分类。

语料库是二语习得和外语教学发展的有效手段,意大利语开源笔语习得语料库有都灵大学的意大利语学习者类型语料库<sup>7</sup>(VALICO)。该语料库收录非意大利语母语的意大利语学习者的笔语文本,可查询文本达3,804篇,能够根据学习者年龄、母语类型、教育程度及教育经历筛选语料。学习者类型语料库诞生一年后,都灵大学又推出了规模仅为729篇笔语文本的意大利语母语者类型配对语料库(VINCA),其文本主题内容与学习者类型语料库一致。学习者和母语者类型语料库灵活对照使用,能为研究民族结构复杂的意大利语学习者的语言特点提供科学工具,用途广泛(Caruana 2020)。开源口语习得语料库有锡耶纳外国人大学的外国人意大利语口语语料库<sup>8</sup>(LIPS)。该语料库收录约2,198次口语考试中共计约100小时的录音,包括根据欧洲语言共同参考框架从A1到C2级别的意大利语,



其内容形式与题材丰富多样,有对话、独白、介入独白的对话、介入对话的独白以及对话独白交替5种口语类别,对于不同水平的意大利语学习与教学都能起到具有针对性的指导作用。

伴随着“网络作为语料库”(Kilgariff & Grefenstette 2003)的研究方法提出,意大利的语料库语言学家逐渐将潜藏着海量语言数据资源的互联网作为语料库构建的强力基础。目前规模最大的意大利语网络语料库是TenTen多语语料库家族中的意大利语语料库<sup>9</sup>(itTenTen),该语料库的规模随每次更新显著扩大,从2010年至2020年已实现由25亿词到124亿词的跨越。同样规模较大的还有“网络作为语料库倡议”语言学家社群(WaCKy)自2009年起推出的意大利语网络语料库<sup>10</sup>(itWaC),该语料库从互联网上自动收集文本,总词数达20亿。另有2012年博洛尼亚大学等研究机构合作完成的派萨网络语料库<sup>11</sup>(corpus PAISÀ),该语料库规模较小,总词数达2.5亿,可以作为各类语言研究活动的资源。

开源语料库在建库时一般会遵循代表性与系统性原则,依照一定的逻辑结构设定,在确定的抽样范围内收集语料,同时根据语料库具体用途选择几个重要指标作为平衡因子,兼顾平衡性。开源语料库为研究人员提供了极大便利,省去了大量语料收集时间,是语言学研究的有力工具,所以在必要时,在具体研究中,我们并不总能找到合适的开源语料库作为参考语料,应创建专门的语料库,以满足研究需求。

### 3 基于意大利语开源语料库的研究

为了解意大利语开源语料库的研究应用情况,我们以主要的开源语料库为关键词搜索文献,整理了围绕形态、句法、二语习得与外语教学、语用分析等方面的研究。以下结合具体案例介绍开源语料库在意大利语研究中的应用。

#### 3.1 形态学研究

意大利语属于词形变化丰富的屈折语,形态学是意大利语语言学中的重要分支,研究词形如何在不同语境下发生变化。意大利语语料库中的语言数据可以作为语法规则与词形变化机制的实证工具,辅助意大利语形态学研究。意大利语词汇中最为重要的构词形式是词汇派生,一般通过向基词增加词缀实现(Palermo 2020: 57),因此基于开源意大利语语料库的形态学研究通常集中于词缀研究。Calpestrati(2017)基于CORIS/CODIS笔语语料库分析意大利语super-/extra-/mega-/iper-/maxi-/ultra-6个强化前缀,发现ultra-/iper-/extra-使用较少,常与形容词相结合,而super-/mega-/maxi-使用较多,常与名词相结合。此外,该研究从跨语言对比的角度出发,考察德语强化前缀在COSMASII Korpus德语语料库中的使用情况,发现德语中同样更多地使用super-/mega-,而几乎不使用其他前缀。强

化前缀搭配不同词性的规律在德意笔语中也存在差异。以最常用的super-为例,其在CORIS/CODIS中出现名词搭配104次、形容词搭配43次、动词2次、名词短语3次,而在COSMASII中分别为90次、3次、24次和4次。Cacchiani (2011)则对比了英语和意大利语的词缀。该研究以corpus La Repubblica、CORIS/CODIS等意大利语语料库和英国国家语料库(British National Corpus,简称BNC)中的例句为分析对象,发现两种语言中形式相同的词缀(mega-/super-/ultra)或同源词缀(如iper-/hyper和arci-/arch-)的功能与用法存在差异,在实际翻译过程中词缀形式也无法保持对等。如强化前缀mega-在两种语言中都存在,但megaconcert对应megaconcerto或concertone, megamind对应supercervello或cervellone。性质前缀semi-同理,semicircolare可对应semicircular,但semiassiderato只能译为almost frozen。又比如,英语中缩小后缀-let可能并不具有明显的褒贬含义,但在译作意大利语时可能对应贬义后缀-uccio,如kinglet译作reuccio。意大利语中的放大和缩小后缀表意丰富,常在英语中无对应后缀或语义无法对等,如giallino译作dim yellow,其中gialletto同时包含着玩笑和喜爱的含义,更贴切的英文译法应为nice dim yellow。研究还发现,意大利语中不同强化词缀的使用与语境相关,前缀extra-常用于与专业领域相关的语境,如extraurbano和extraparlamentare。此外,搭配词本身带有的情感含义还会使得后缀-one/-ino等根据不同语境而具有褒贬、玩笑和同情等不同含义。这两项基于语料库的形态学研究聚焦词缀,不仅揭示了我们平时难以察觉的语言规律,还通过跨语言词缀对比,凸显出不同语言的词汇形态在语言应用中的差异。

### 3.2 句法学研究

意大利语句法研究通常聚焦词汇、短语和从句依照怎样的语法规则组成其上级成分,分析语言结构和语序变化如何影响语句表意。语料库可以反映语言的真实状况,是分析句子成分的定量信息来源。Marzo & Crocco (2015)基于LIP和CLIPS口语语料库以及CORIS/CODIS笔语语料库研究意大利语中c'è或ci sono引导名词或代词再接上che引导的伪关系从句的陈述结构(costruzione presentativa)。研究发现,这种句式结构语序固定,只能用于肯定句<sup>12</sup>,且在新标准意大利语中并不常见,且相比书面语更常出现在口语中。由于LIP和CLIPS口语语料库均从不同地区收集语料,所以该研究也将意大利地区之间显著的语用差别考虑在内,发现不同城市的口语中使用上述句式结构的频率也存在差异:佛罗伦萨、米兰、罗马和那不勒斯分别为27.4%、17.8%、32.2%和22.6%。Crocco (2010)则对新标准意大利语中的右脱位结构<sup>13</sup>(dislocazione a destra)进行研究。该研究同样考虑地区语用差别,结果发现,米兰人在对话中使用右脱位结构的次数显著少于佛罗伦萨人。

基于语料库的句法研究还能够结合创新方法达到研究目的。如 Tamburini *et al.* (2002) 将半自动标注分布分析法 (Brill & Marcus 1992) 应用于 CORIS/CODIS 笔语语料库, 以定量方法比较目标词分布和搭配的相似度, 得到共词聚类树状图。该研究将 *posto che/nel frattempo/per esempio/appunto/infatti/tra l'altro* 等本身没有连接含义的弱连接成分, 以及 *oltre che/dopo che/per quanto/poiché* 等具有强连接功能的成分进行聚类, 直观可视地揭示不同连接词之间的近似程度, 验证其语法规律。Mauri & Masini (2021) 则通过话语分析、跨语言共时和语言历时分析结合组成的 3D 模型法, 扩大了句法研究的外延。该研究从 VoLIP、CORIS 和 KIParla 等多个语料库中提取例证, 以构式语法 (Construction Grammar, CxG) 为理论框架, 分析了转折连词 (*connettivo disgiuntivo*)、伪并列结构<sup>14</sup> (*pesudo-coordinazione*) 和重复短语<sup>15</sup> 在语言内部或跨语言表达中的表现。结果发现, 意大利语中的转折连词 *o* 有 80% 都用于疑问或质询语境, 或是出于习惯使用, 只有 20% 用于提出另一选择。而疑问副词 *magari* 或是重复结构 *vuoi...vuoi...* 也可表示提出另一选择。伪并列结构 *mettersi lì e V* 和 *saltare su e V* 都能传达与 *prendere e V* 相同的不确定性与突发性, 但 *saltare su e* 只能与表示说话的动词搭配, *mettersi lì e* 还能与表示持续的动词搭配, 表达“承诺”和“奉献”。意大利语中的重复短语形式上类似于中文叠词, 但还有表示物品货真价实的含义。上述研究基于不同语料库探讨了意大利语中的几种句法现象, 涉及句式结构和连接词, 以及句中成分对于语句表意的影响。使用语料库这一定量分析工具时, 句法研究能够提高结果的可信度, 与创新研究方法结合时, 句法研究通常呈现出跨学科特征。

### 3.3 二语习得与外语教育研究

语料库是二语习得与外语教学的策略工具, 其最大优势在于语料真实, 能够用语境共现功能呈现出不同语境中的特定语言现象, 协助学习者掌握词汇以及词法、句法等语言规律。基于学习者语料库的研究对学习者的中介语进行分析, 能够预判不同母语背景和教育经历的学习者容易发生的各类失误, 通过比照母语语料库中的对应内容, 为教学者提供教学建议。VALICO 学习者语料库收录了大量中介语资源, 展示了不同语言背景的学生习得意大利语的特点, 可以与 VINCA 本族语语料库形成对照。基于意大利语开源语料库的二语习得与外语教育研究呈现出切入点多样化的特点。

Valentini (2018) 通过比较 VINCA 和 VALICO 语料库中英语母语者和德语母语者的语料得出结论: 与母语句法中主谓宾成分顺序较灵活 (如英语) 的学习者相比, 母语成分顺序严格 (如德语) 的学习者更加难以根据语用调节词序, 在学习和使用意大利语时也更加不愿意偏离基本词序。Caruana & Novello (2020) 聚焦马耳他和意大利博尔扎诺自治省两个教育体系分别深受英语和德语影响的多语

社区,选取中学生笔试作文为研究材料,借助 VALICO 和 VINCA 语料库比照研究两地区学生笔语中存在的特点。研究发现,两地区学生的笔语表达习惯与意大利语母语者差异较大,通常受到本地使用口语的影响,直接挪用英语或德语中的单词变为意大利语。如将英语中的 to profit 变为 *profitare*,将德语中的 *ruinieren* 变为 *ruinare* 等。Corino (2016) 则专门选取意大利语学习者难以掌握的话语标记 *cioè* 为分析对象,研究学习者在使用该话语标记时普遍出现的错误,以及不同母语学习者使用该话语标记习惯之间的差异。该研究通过 VALICO 等语料库分析得出,学习者最常用 *cioè* 表示解释说明,但通常将其与具有解释作用的冒号相混淆<sup>16</sup>。在意大利语学习者中,波兰语、法语和日语母语者使用该表达的频率最高,德语母语者通常受到母语表达 *das heisst* 的影响,将 *cioè* 和 *dire* 连用。以上两项研究基于学习者语料库研究了意大利语学习者容易产生的语言偏误,这些偏误通常与学习者的母语表达习惯存在关联。在应用层面,我们可以使用学习者语料库中的语言偏误作为测试多选题中的干扰项,帮助教师根据学生母语条件来评估学生的语言学习状况 (Marello 2009),以便对症下药。

### 3.4 语用学研究

语用学研究语言的实际应用以及语言含义如何受到语境的影响和制约,不同类型的语料库不仅能够捕捉静态的语用功能和结构,还能体现语用的动态演变,是语用学研究的有力工具。基于意大利语语料库的语用学研究呈现出主题多样和分散的特点。Lo Baido (2018) 选择研究 *non so*、*tipo* 和 *per esempio* 等例证结构的语用。该研究通过 CORIS 等笔语语料库和 LIP 口语语料库中的语用案例分析发现,这些结构原本用于进一步说明和举例,但在实际语用中也可以缓和语气,类似于英语中的 *I think for example that...* 或 *I suppose for instance that...*。Farese (2020) 则选择与意大利传统文化联系紧密的名词 *carità* 为研究对象。基于 CORIS 和 CODIS 笔语语料库,该研究发现, *carità* 通常和 *chiedere*、*accettare*、*rifiutare* 和 *fare* 等动词组成动词短语,表示“仁慈”“慈善”。然而,在特定语境下,“*carità*”也有讽刺含义,例如在形容不足挂齿的小事时。当组成短语 *per carità* 时,该名词的语用意义会变得十分灵活,除表示原有的行善意义外,还可用于表示强调、祈求、拒绝、辩解甚至消极情感。同样研究词汇不同语用功能的还有 Cruschina & Cognola (2021)。该研究以时间副词 *poi* 为研究对象,基于 corpus La Repubblica、CORIS 笔语语料库和 KIParla 口语语料库进行分析。结果发现,该副词除了充当逻辑或时间顺序连词以外,还可以作为话语标记,表示划分、总结和对比,或者作为语气助词,表示反对、关心、不确定与猜测等情绪。研究还指出, *poi* 作为时间副词时常位于句尾,作为连词时常位于句首,作为语气助词时常位于句中,作为话语标记时位置则较为灵活。由此可见,语用研究能够与句法研究形成学科交叉。



上述研究选择特定短语或词汇作为研究对象,通过语料库补充了它们在实际使用中容易被忽略的用法。语用会随着时间发生变化,选取历时语料库作为研究工具可以进行历时语用学研究。如Lorella(2020)通过DiaCORIS/CORIS、Corpus Stammerjohann、corpus LIP、C-ORAL-ROM Italia和意大利语词形历时语料库(Morfologia dell'Italiano in DIACronia,简称MIDIA)对意大利语中表示回应的话语标记non c'è problema进行了历时变化研究。研究发现,non c'è problema在交际中逐渐取代了如con piacere和volentieri等语义几乎相同的表达。语义相同的表达在不同语言文化中的语用可能不尽相同,考察不同语言的语料库则可以进行对比语用学研究。例如上述研究还对比了non c'è problema在意大利语语料库中和no problem在ARCHER和COHA历时英语语料库<sup>17</sup>中的语用。研究发现,英语中的no problem出现较早,且该表达在英语语料库中出现频率的提高与non c'è problema在意大利语语料库中出现频率的提高之间存在显著正相关,no problem可能补足了non c'è problema的语用空缺。Cappuzzo(2020)则从跨文化语用学出发,分析了意大利语中英语借词的语用。该研究基于La Repubblica语料库,对意大利语中与新冠疫情相关的英语借词如lockdown、cluster、task force和smart working进行分析。结果发现,这些英语借词表意非常丰富,仅lockdown在翻译中就能对应意大利语中的blocco、chiusura、confinamento、contenimento、coprifuoco、distanziamento sociale、isolamento and quarantena和misure restrittive/misure di restrizione/restrizione等表达。在实际应用中,英语借词通常与其翻译对应的意大利语表达混用,以避免重复。语用学还能与社会语言学发生交叉,借助语料库以实证方法分析语用能够揭示深刻的社会现象。如Nardone(2016)基于itWaC网络语料库分析意大利语中与社会职业相关单词的词频,发现其中存在严重的性别不对称现象。意大利语中相同职业名词的阴性和阳性形式词频差异较大,“建筑师”“外科医生”“工程师”的阴性形式architetta、chirurga和ingegnera的词频显著低于其阳性形式architetto、chirurgo和ingegnere,“合作者”“主任”“秘书”的阴性形式collaboratrice、direttrice和segretaria的词频则高于其阳性形式,而“检察员”“医生”“公证员”ispettore、medico和notaio则几乎没有阴性形式。此外,职业名词的阴性形式通常与阳性形式语义不对称,且会与特定语义场的单词搭配。例如segretaria常与含有从属且无尊敬意义的词语如impiegata、personale、scuola和giovane等搭配;direttrice常与文化和教学领域的词语如biblioteca、museo、didattica和rivista等搭配,而它们的阳性形式则没有这种搭配倾向,该研究从语用上反映出意大利社会部分职业领域的性别不平等现象。从以上研究可见,适当选取和组合语料库作为研究工具能够丰富语用学的研究方向,从历时、跨语言文化甚至社会学的角度来分析语言的实际应用情况。

基于意大利语开源语料库的研究具有明显的应用意义和学科交叉特征,能够

解决且不止步于解决与语言学相关的各种问题。这些研究通常围绕着意大利语本身的复杂语言特征展开,选择适当的切入点来解释相应的语言现象,同一研究领域中的主题丰富多样,值得借鉴以开拓研究思路。基于开源语料库的研究应首先验证选库的合理性,需重点关注开源语料库的规模、建库规则及语料类型,考量该语料库是否适用于解决特定的研究问题并产生可信结果。

## 4 结语

本文介绍了意大利语语料库的发展、意大利语语料库语言学的主要研究机构与团队、主要的意大利语开源语料库以及基于这些语料库进行的相关研究。意大利于20世纪50年代开启了语料库电子化的序章。随着语料储存与信息提取的工具发生革新,意大利学界在语料库发展的三个历史阶段中成果卓著,无论从语料库建设的规模还是种类而言都展现出极高的丰富度与专业性。意大利的语料库语言学研究团队数量庞大,以罗马智慧大学、都灵大学、佛罗伦萨大学等高校为中心,集中活跃于20世纪末至今。各研究团队的语料库研究方向呈现出复杂交织但各有侧重的特点,为服务于不同的研究目的而研制了多种类别的开源语料库。以这些意大利语开源语料库为基础开展的研究不计其数,涉及形态、句法、二语习得与外语教学和语用分析等方面,范围宽广且切入点多样,可为国内意大利语语言学及语料库研究提供一定的参考与启示。

### 注释

- 1 LIF 语料库收录1947—1968年戏剧、小说、电影和报纸等文本,共计50万词。斯坦默约翰当时于佛罗伦萨收集语料制作即兴口语语料库,共计42小时录音。然而该口语语料库最终并未完成转写,且样本存在诸多局限,后来他将未完成的语料库捐赠给佛罗伦萨意大利语语言实验室,用于进一步研究。
- 2 秕糠学会语料库数据库: <https://accademiadellacrusca.it/it/contenuti/banche-dati-corpora-e-archivi-testuali/6228> (2022/8/20)。
- 3 萨莱诺大学沃盖拉教授于2004年在意大利教育与研究部(Ministero dell'Istruzione, dell'Università e della Ricerca)资助下,主持整合全国语料库语言学资源,创立语料库并发布了“说意大利语”门户网站: <https://parlaritaliano.studiumdipsum.it/it/cat/11-corpora> (2023/2/15)。
- 4 研究院在1978年转型前为比萨国立大学电子计算中心(CNUCE),LIF语料库制作于1971年,故作者为电子计算中心。
- 5 corpus CORIS: <https://corpora.ficlit.unibo.it/TCORIS/>, corpus CORDIS: <https://corpora.ficlit.unibo.it/CODIS/> (2023/2/11)。
- 6 BADIP: <http://143.50.35.46/it/cerca> (2022/11/15), Corpus VoLIP: <https://www.>



- volip.it/ (2023/2/10)。
- 7 corpus VALICO: <http://www.valico.org/> (2022/11/15) 该页面也可访问 corpus VINCA。
  - 8 corpus LIPS: <https://parlaritaliano.studiumdipsum.it/it/653-corpus-lips> (2023/2/11)。
  - 9 itTenTen: <https://www.sketchengine.eu/ittenten-italian-corpus/> (2023/2/11), TenTen 语料库命名意在其目标为该语料库家族内各个语料库规模均能达到10的10次方, 即100亿词。
  - 10 itWaC: <https://www.sketchengine.eu/itwac-italian-corpus/> (2023/2/11), 语言学家社群 WaCKy 即 Web-As-Corpus Kool Yinitiative。
  - 11 corpus PAISÀ (Piattaforma per l'Apprendimento dell'Italiano Su copra Annotati corpus): <https://www.corpusitaliano.it/it/contents/partners.html> (2023/2/11)。
  - 12 例如 “C'è il gatto che ha fame.” “C'è un signore che vuole parlare con te.” 或者 “c'è la signorina che entra.”, 但不能改变语序表述为 “La signorina c'è che entra.”, 且不能用于否定句或疑问句, 如 “Non c'è la signorina che entra.” 或 “C'è la signorina che entra?”。
  - 13 右脱位结构为意大利语口语中特有的句法结构错位现象, 即改变主谓宾 (SVO) 语序, 将宾语放在右侧的句尾。例如将正常句法顺序 “Ho mangiato la pizza ieri.” 变为 “L'ho mangiata ieri, la pizza.” 或 “Hai l'ombrello?” 变为 “Ce l'hai l'ombrello?”。
  - 14 这里的 “伪并列结构” 为功能动词 (轻动词) 和助动词复合而成的谓语结构, 两动词以相同词性出现, 共同表示一项复杂事件, 而非合为一体的两项独立事件, 意大利语中的伪并列结构常以 *prendere e V* 形式出现, 其中 *prendere* 为轻动词, 如 “vagabondano, sono animali erranti. *Prendono e partono*, e non è affatto detto che ritornino.” 斜体部分表示 “突然离开”, 而非两个动作。
  - 15 重复短语在意大利语中非常常见, 如 *piccolo piccolo*、*presto presto*、*caffè caffè* (货真价实的咖啡)。
  - 16 如: *Tutto quello che portavano con loro si è trovato sul marciapiede, cioè le sue spese, la carne, le boteglia si sono rotte...*。
  - 17 ARCHER 语料库 (A Representative Corpus of Historical English Registers) 收录1650—1999年的英国和美国英语, COHA (Corpus of Historical American English) 收录1810—2009年的美国英语。

#### 参考文献

- BONELLI T, SINCLAIR J. Corpora [C]//BROWN K. Encyclopedia of language and linguistics. Amsterdam: Elsevier, 2006: 206-220.
- BRILL E, MARCUS M. Tagging an unfamiliar text with minimal human supervision [C]// Proceedings of the fall symposium on probabilistic approaches to natural language. Cambridge, MA.: American Association for Artificial Intelligence, 1992: 121-127.

- BUSA R. L'index Thomisticus e l'informatica filosofica [J]. *Revue Internationale de Philosophie*, 1973, 27(1): 31-36.
- CACCHIANI S. Intensifying affixes across Italian and English [J]. *Poznań Studies in Contemporary Linguistics*, 2011, 47(4): 758.
- CALPESTRATI N. Intensification strategies in German and Italian written language [C]// NAPOLI M, RAVETTO M. Exploring intensification: Synchronic, diachronic and cross-linguistic perspectives. Amsterdam: John Benjamins, 2017: 305-326.
- CAPPUZZO B. Anglicisms and Italian equivalents in the era of Covid-19: A corpus-based study of lockdown [J]. *European Scientific Journal ESJ*, 2020, 16(38): 7-26.
- CARUANA S. I corpora Valico e Vinca [J]. *RiCOGNIZIONI: Rivista di Lingue, Letterature e Culture Moderne*, 2020, 13(7): 153-156.
- CARUANA S, NOVELLO A. Trattati di interlingua comuni: analisi di produzioni scritte di italiano L2 di apprendenti maltesi e della provincia di Bolzano [J]. *RILA: Rassegna Italiana di Linguistica Applicata*, 2020, (2/3): 153-182.
- CORINO E. Learners and reformulative discourse markers: a case study of the use of “cioè” by students of Italian as a foreign language [J]. *Language, Interaction and Acquisition*, 2016, 7(1): 44-66.
- CRESTI E, MONEGLIA M. La linguistica italiana dei corpora [C]//LUBELLO S. *Manuale di linguistica italiana*. Berlin: De Gruyter, 2016: 581-611.
- CRESTI E, PANUNZI A. Introduzione ai corpora dell'italiano [M]. Milano: Il Mulino, 2013.
- CROCCO C. La dislocazione a destra tra italiano comune e variazione regionale [C]// PETTORINO M, GIANNINI A, DOVETTO F. *La comunicazione parlata 3*. Napoli: Università degli studi di Napoli l'Orientale, 2010: 191-210.
- CRUSCHINA S, COGNOLA F. From connective adverb to modal particle: a generative analysis of “poi” [J]. *Studi Italiani di Linguistica Teorica ed Applicata*, 2021, 50(1): 52-68.
- DE MAURO T. *Grande dizionario italiano dell'uso* [Z]. Torino: UTET, 1999.
- DE MAURO T. *Primo tesoro della lingua letteraria del Novecento* [Z]. Torino: UTET. Fondazione Maria e Goffredo Bellonci, 2007.
- DE MAURO T, MANCINI F, VEDOVELLI M, et al. *Lessico di frequenza dell'italiano parlato* [Z]. Milano: Etas, 1993.
- FARESE G. Christian values embedded in the Italian language: a semantic analysis of carità [C]//BROMHEAD H, YE Z. *Meaning, life and culture*. Canberra: ANU Press, 2020: 173-191.
- GORIA E, CERRUTI M, BALLARÈ S, et al. KIParla corpus: a new resource for spoken Italian [C]//*Proceedings of the sixth Italian conference on computational linguistics*.

- Bari, Italy, November 13-15, 2019.
- ITALIA IBM. Il vocabolario del 2000 le prospettive della lessicografia nell'era informatica. VELI Il vocabolario elettronico della lingua italiana [M]. Milano: IBM editore, 1989.
- KILGARRIFF A, GREFFENSTETTE G. Introduction to the special issue on the web as corpus [J]. Computational Linguistics, 2003, 29(3): 333-347.
- LO BAIDO M. Mitigation via exemplification in present-day Italian: a corpus-based study [J]. ELUA. Estudios de Lingüística, 2018, Anexo (IV): 11-32.
- LORELLA V. From linguistic innovation to language change [J]. Revue Romane, 2020, 55(1): 95-116.
- MARELLO C. Distrattori tratti da corpora di apprendenti di italiano LS/I2 [C]//CORINO E. Valico: Studi di linguistica e didattica. Perugia: Guerra, 2009: 179-193.
- MARZO S, CROCCO C. Tipicità delle costruzioni presentative per l'italiano neostandard [J]. Revue Romane. Langue et Littérature. International Journal of Romance Languages and Literatures, 2015, 50(1): 30-50.
- MAURI C, MASINI F. Diversity, discourse, diachrony: a converging evidence methodology for grammar emergence [C]//VOGHERA M. From speaking to grammar. Bern: Peter Lang, 2021: 55-89.
- NARDONE C. Asimmetrie semantiche di genere: un'analisi sull'italiano del corpus itWaC [J]. Gender/Sexuality/Italy, 2016, (3): 1-17.
- PALERMO M. Linguistica italiana [M]. Bologna: Mulino, 2020.
- TAMBURINI F, DE SANTIS C, ZAMUNER E. Identifying phrasal connectives in Italian using quantitative methods [C]//NUCCORINI S. Phrases and phraseology-data and description. Berlin: Peter Lang, 2002: 45-62.
- VALENTINI A. Ordini marcati delle parole nell'italiano scritto: apprendenti germanofoni, apprendenti anglofoni e parlanti nativi a confronto [J]. Incontri. Rivista Europea di Studi Italiani, 2018, 33(1): 70-88.
- 董丹. 评价理论视角下意大利主流媒体对十九大报道的积极话语分析[J]. 外国语文, 2019 (4): 17-23.
- 李勤, 常翔宇. 俄罗斯语料库语言学的学科建设与发展探微[J]. 东北亚外语研究, 2018 (2): 54-60.
- 田园. 语料库语言学在法国的学科建设[J]. 法国研究, 2014 (1): 73-77.
- 许家金. 美国语料库语言学百年[J]. 外语研究, 2019 (4): 113-118.
- 余丹妮. 汉意意汉文学平行语料库的研制[J]. 语料库语言学, 2020 (2): 83-88.
- 余丹妮, 张斌. 《习近平谈治国理政》汉意平行语料库在外交翻译教学中的应用[J]. 欧洲语言文化研究, 2022 (2): 1-13.

- 喻儒辰, 董丹, 郭垚一. itGLOBE 当代意大利语书面语平衡语料库的创建[J]. 语料库语言学, 2023 (2): 141-149.
- 赵冲, 许家金. 近百年西班牙语语料库建设与研究概述[J]. 欧洲语言文化研究, 2023, (1): 119-135.

**通信地址:** 100089 北京市 北京外国语大学欧洲语言文化学院

# arGLOBE 当代阿拉伯语书面语平衡语料库的创建

北京外国语大学 毛浚语

**提要：**arGLOBE 当代阿拉伯语书面语平衡语料库是“北外全球语料库集群”项目（即“GLOBE 语料库”项目）下的子课题，致力于依照布朗语料库的采样标准收集近十年的阿拉伯语书面语文本，建设百万词级的平衡语料库。本文首先简述面向阿拉伯语的语料库建设情况，在此基础上介绍 arGLOBE 语料库的建设理念和建库过程。此外，本文对该库可以开展的语言学研究进行探讨，并展望阿拉伯语语料库的后续建设。

**关键词：**arGLOBE 语料库、当代阿拉伯语书面语、阿拉伯语教学与研究

## 1 引言

北京外国语大学于2021年12月29日启动了“北外全球语料库集群”项目，又称“GLOBE 语料库”项目，其中“GLOBE”是“Corpus of Global Languages Out of BFSU Expertise”的首字母缩略词。该语料库集群项目依照布朗语料库的采样方案，致力于为北外开设的101个语种建设当代书面语平衡语料库。该采样依据有助于使其单语语料库与现有的布朗家族语料库具有可比性，便于开展汉英、汉外或多语对比研究。“arGLOBE 当代阿拉伯语书面语平衡语料库”简称“arGLOBE 语料库”，为 GLOBE 语料库的当代阿拉伯语子库，其设计规模为100万词，所含文本大多出版或发表于2010—2022年。

## 2 面向阿拉伯语的语料库建设简述

语料库语言学思想在阿拉伯语传统语法的发展历程中应用已久。Ditters (1990: 120-130)、Brustad (2016: 148-149) 等学者指出，基于实证主义的语料库语言学思想甚至在几千年前就被阿拉伯语语法学家使用，成为阿拉伯语语法研究的根基。例如，著名阿拉伯语语法学家西伯威 (Sībawayh) 在编纂古代阿拉伯语语法研究的经典之作《西伯威书》(Kitāb Sībawayh) 时便使用“经证实的语言” (attested language) 也即真实存在的特定语料作为语法研究的基础，这些语料包括前伊斯兰时期的诗歌、部落战争的记录、正式的演讲、阿拉伯人的对话等。Ditters

(1990: 130)指出,被参考的语料包含口语和书面语且区分不同的语域,这种具有语言学意义的构建结构有助于早期的语言学家更好地归纳古代阿拉伯人话语使用的典型特征。

尽管如此,现代阿拉伯语语料库建设却起步较晚。目前阿语语料库语言学研究刚刚兴起,但也取得了一定进展(McEnery *et al.* 2018: 8)。当前,面向阿拉伯语的电子化语料库主要包括但不限于以下几种类型。(1)通用语料库。以杨百翰大学开发的arabiCorpus(Parkinson 2018)、TenTen语料库家族的arTenTen语料库(Arts *et al.* 2014)为代表。前者库容为100万词,语料以新闻为主,另含少量中世纪阿拉伯语和埃及方言。后者语料取材自主题不同的网页,2018年版本的arTenTen18库容可达46亿词。(2)方言语料库。阿拉伯世界方言盛行,标准语和方言共存的双言现象(diglossia)使阿拉伯语方言语料具有重要研究价值。这类口语语料库例如记录巴勒斯坦、黎巴嫩、伊拉克、利比亚、苏丹和也门社交平台方言语料的CurraSat(Jarrar *et al.* 2017; Haff *et al.* 2022; Jarrar *et al.* 2022),包含突尼斯社交平台、电视剧、广播等领域方言语料的Tunisian Arabic Corpus(McNeil 2018)。(3)历时语料库。知名的语料库例如,由阿卜杜勒—阿齐兹国王大学开发的KACST阿拉伯语语料库(Al-Thubaity 2015),其库容为7亿词,包含从前伊斯兰时期至今跨越1500余年的语料。(4)学习者语料库。如利兹大学开发的Arabic Learner Corpus(Alfaifi *et al.* 2014),该语料库收集了在沙特的阿拉伯语学习者产出的书面语和口语语料。(5)专用语料库。如具有词法标注、句法标注和语义解析的古兰经语料库Quranic Arabic Corpus(Dukes & Habash 2010)。(6)平行语料库。如联合国平行语料库(Ziems *et al.* 2016),该库包含1990至2014年间编写并经人工翻译的文字内容,语种覆盖包括阿拉伯语和汉语在内的联合国六大官方语言。

以上着重列举了各类别下可公开访问的代表性语料库。相比于汉语语料库和英语等西方语种的语料库,面向阿拉伯语的语料库整体数量较少,阿拉伯语书面语的平衡语料库建设仍需进一步推进。现有阿拉伯语语料库的建设并未遵循被多个语料库共同认可并传承的采样标准,例如,arabiCorpus主要以新闻语料为主,辅以中世纪的阿拉伯语语料。这使得语料库之间的可比性不足,各阿拉伯语平衡语料库之间及其与其他语种语料库之间难以促成更具规范的语言比较和对比研究。而arTenTen语料库虽在TenTen语料库家族中具有可比性,但该语料库未严格采用平衡语料库的建库模式,而是用阿拉伯语维基百科中的高频词爬取网页数据的方式进行整合。本文介绍的arGLOBE语料库主要以近十年首次出版或发表的阿拉伯语书面语为目标语料,依照布朗语料库的采样模式进行平衡语料库建设,并与“北外全球语料库集群”项目中的各语种语料库一同构成可比语料库,既为现有的阿拉伯语平衡语料库提供有益补充,又可促成语言类型学层面的多语种对比研究。



### 3 arGLOBE 语料库的创建

arGLOBE语料库基于布朗语料库的采样方案进行建设，是库容约为一百万词的当代阿拉伯语书面语平衡语料库，所含文本主要发表于2010年至2022年间。该库包括生语料、词性赋码和词形还原三个版本，其中词性赋码及词形还原所涉及的自然语言处理工具为StanfordNLP工具包中的stanza-ar包。在此基础上，该库提供阿拉伯语词频表，内含经词性赋码和词形还原的单词以及二元词组和三元词组，可供教学与研究之用。目前，arGLOBE语料库已上传至“北外CQPweb多语种语料库平台”（<http://114.251.154.212/cqp/>）。该语料库在线检索平台提供主题词分析、搭配计算、索引分析、词表生成等功能。

#### 3.1 采样方案

arGLOBE语料库按布朗语料库的方案进行采样，所收集的文本类型及其相应篇数见表1<sup>1</sup>。

表1 arGLOBE 语料库文本类型及相应篇数

体裁大类	体裁类型	子体裁代码	子体裁类型说明	文章数量（篇）
信息类 （374篇）	新闻	A	新闻报道	44
		B	社论	27
		C	报刊评论	17
		D	宗教	17
	通用	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	29
虚构类 （126篇）	小说	L	侦探小说	24
		M	科幻小说	6
		N	历险悬疑小说	29
		P	言情小说	29
		R	幽默	9
合计				500

布朗语料库，全称为布朗大学当代美国英语标准语料库（Brown University Standard Corpus of Present-Day American English），是世界上第一个根据系统性原则进行采样的百万词级英语电子语料库。该库由布朗大学的学者Henry Kučera和W. Nelson Francis所建，所含文本为发表于1961年的500篇美国英语书面语文本，共计约一百万词。Kučera和Francis后于1967年发表《当代美国英语的计算分析》（*Computational Analysis of Present-Day American English*）一书，对布朗语料库的数据进行描述分析。布朗语料库的文本选自多个来源，包含新闻、通用、学术、小说四大体裁，下分新闻报道、社论等15个子体裁。其采样标准保证了所采语料的平衡性和代表性，进而使所收样本能较好地反映语言的整体特性，有助于开展词典编纂和各领域的语言学研究。此外，布朗语料库的建设还促成了语料库建设领域的热潮，其采样标准影响了诸多同类型语料库的建设。这些语料库包括收集英国英语文本的LOB、FLOB、B-BLOB、CLOB语料库等，以及收集美国英语文本的FROWN、CROWN和CROWN2021语料库等。这类按相同采样标准建成的语料库被称为“布朗家族语料库”，各库之间具有良好的可比性，有利于开展基于可比语料库的语言学研究，例如不同英语变体间的语言结构对比、某种英语变体在不同时期的历时研究、不同语种间的语言对比等。

arGLOBE语料库及其所属的北外全球语料库集群均按布朗语料库的采样标准进行建设，因而获得的可比性有助于更好地开展汉英、汉外以及不同语种间的对比研究，为语言结构的差异与共性分析、语言应用层面的研究提供量化分析工具。基于布朗语料库的采样方案，arGLOBE语料库从多渠道选取新闻、通用、学术、小说四大体裁的文本，下分新闻报道、社论等15个子体裁。因文化差异等原因，阿拉伯国家出版物在一些体裁或版块的设置上与布朗语料库代表的美国出版物有所不同。因此，课题组在取样过程中对相应版块进行微调。例如，布朗语料库采样方案中包含“冒险小说和西部小说”（Adventure and western fiction）一项，其中西部小说具有19世纪下半叶美国旧西部的背景设定，不适用于阿拉伯文化。因此，“西部小说”文本未被收入arGLOBE语料库，该子体裁仅收集与“冒险小说”相关的文本。此外，由于不同国别报纸的版块设定各异，“给编辑的信”（Letters to the editor）这一在美国报刊中常见的版块近年来未曾出现在阿拉伯报刊。参考布朗语料库在“社论”（Editorials）子体裁分模块采样时的三分法设计，从“机构”（Institutional）、“个人”（Personal）再到“给编辑的信”，其分类趋势呈现从“官方”“一般”到“个人”的话语风格变化，故arGLOBE语料库在采样时将“给编辑的信”替换为更偏向“个人观点表达”的文本，以符合该子体裁分类趋势的话语风格变化。这些微调考虑了阿拉伯国家的文化背景、报刊版块设定等方面的因素，有利于更好地维护语料的代表性；微调程度适度且未破坏采样标准的整体结构，这也维护了所采语料的平衡性。

## 3.2 语料采集

arGLOBE 语料库以上述采样方案为基础,采集 500 个 2,000 词左右的阿拉伯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、文本录入和语料库元信息标注这 3 个环节。

### 3.2.1 文本收集与取样

arGLOBE 语料库在第一版规划中面向主要阿拉伯国家的阿拉伯语标准语出版物,暂不涉及各国的阿拉伯语方言变体,文本内容因特定的话语风格所需而涉及的方言语料除外。尽管该库原则上收录位于西亚和北非总共 22 个阿拉伯国家的语料以充分体现语料平衡性,但考虑到不同国家因文化影响力等因素造成出版物质量、数量、代表性有别,arGLOBE 语料库实际收集的文本仅涉及阿拉伯世界三大地区的主要国家,即北非地区、海湾地区和沙姆地区的主要国家,且确保三大地区的文本均有涉及。这样的语料采样来源设定在确保语料代表性的同时也最大程度维护了平衡性。该库所收文本的第一作者国籍原则上均为阿拉伯国家,且所收文本均为原创阿拉伯语,从其他语言译入阿拉伯语的文本不在该库的收集范围内。此外,由于阿拉伯语出版物公开流通的规模较英语等西方语种而言总体较小,考虑到语料收集的可操作性,arGLOBE 语料库和布朗家族语料库相比扩大了语料发布的时间范围。其中,新闻类语料发布时间主要为近三年,其他三类体裁(通用、学术、小说)所收文本的首次出版时间大多为近十年。所收语料的时间与数量分布可见图 1。

为了尽力符合与布朗家族语料库的可比性以及维护语料库的均衡性,arGLOBE 语料库除了考虑语料发布时间的因素外,在采样过程中也参考了布朗语料库对语料来源、主题的划分标准。例如,在“新闻报道”子体裁下,arGLOBE 语料库选取政治、体育、社会、热点新闻、经济、文化等主题多元化的报道,并在一定程度上兼顾日报、周报的取材来源划分;在“政府或机构公文及文宣”子体裁下,该库在采样时依照原标准选取政府文件、基金报告、工业报告等方面的语料;在“宗教”“日常技艺及消遣爱好”“通俗读物”等子体裁下,该库在采样时尽力确保书籍、期刊的来源划分,并对篇幅较长的书籍进行前、中、后三部分拼接采样,以维护文本内容在书籍内部的代表性。此外,对于同一子体裁的语料出现于不同阿拉伯国家出版物的情况,该库尽力确保在阿拉伯世界三大地区(北非地区、沙姆地区、海湾地区)各选取一定数量的语料来反映阿拉伯世界出版物的整体情况,进而体现语料的代表性和平衡性。

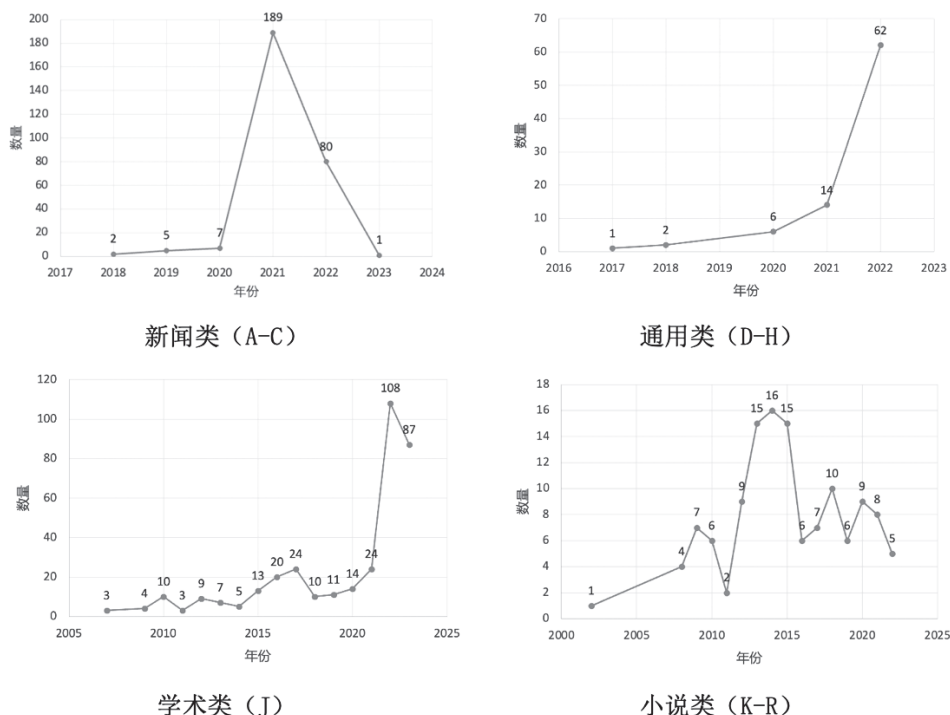


图1 四类体裁所收语料的时间和数量分布

### 3.2.2 文本录入

arGLOBE语料库所收文本长度为每篇2,000词左右，通过正则表达式  $[\text{ا-زA-Z0-9.-}]+$  进行计数。当所选文本篇幅较短时，则将多篇同类型的文本进行拼接，并在其文件名末尾添加A/B/C等字母标记以示区分。例如，阿拉伯语新闻报道一般篇幅较短，“新闻”体裁中“政治”子体裁下的第一篇语料文本便由6个总词数共为2,000词的文本文件构成，其文件名分别为ARA01A、ARA01B、ARA01C、ARA01D、ARA01E、ARA01F。所选文本均经过人工校对，并以UTF-8编码格式储存为txt文本文件。

### 3.2.3 元信息标注

arGLOBE语料库共包含500个2,000词的文本，共存储为847个文本文件。文件的命名规范为“两位字母语种代码—一位字母体裁编码—数字编号—字母编号”，进而清楚地通过文件名表明语料所属的语言、体裁、位置等信息。

除了文件名所体现的分类信息，该库将详细的元信息记录成表，提供所采集语料的文件编号 (File)、出版物标题 (Title)、作者 (Author)、词数统计 (Word Count)、出版商 (Publisher)、出版年份 (Year)、来源 (Source) 和链接 (URL)

八项基本信息。其中，出版年份和作者信息视原文本所提供信息的详细程度加以记录，若原文本所提供的信息足够明确，那么出版年份会进一步精确至具体日期。此外，该元信息表格还提供体裁类型、子体裁文本数量、所属领域等信息，以对所采集文本进行清晰明确的分类。

3.3 语料版本及应用

建成后的arGLOBE共包含生语料、词性赋码和词形还原3个版本，后两个版本为经过StanfordNLP工具包中的stanza-ar包处理所得。

表2以ARA01A中的第一段为例，展示3个版本的语料区别。生语料即未经任何加工的版本。词性赋码则在生语料的基础上，按“单词\_词性码”的形式呈现每个词及其对应的词性编码，例如名词为NOUN、动词为VERB、限定词为DET等。词形还原版本则将原文本中的所有单词逐一还原为该单词的原形，例如动词、指示代词、人称代词等均还原为阳性第三人称单数形式，所有的确指名词均去掉冠词“al-”而成为泛指名词。各研究者可选取特定的语料库版本进行分析，以满足其研究目的。

表2 三种语料版本示例

语料版本	语料标注示例	来源文件
生语料	قال رئيس الحكومة الليبية عبد الحميد الدبيبة الخميس إنه تم توحيد أكثر من 80% من مؤسسات الدولة الليبية تحت مظلة هذه الحكومة، ولم تبق إلا المؤسسة العسكرية.	ARA01A
词性赋码	قال VERB_ رئيس NOUN_ الحكومة NOUN_ الدبيبة ADJ_ عبد X_ الحميد X_ الدبيبة X_ الخميس NOUN_ إن SCONJ_ ه PRON_ تم_ VERB_ توحيد NOUN_ أكثر ADJ_ من SYM_ 80 NUM_ تحت ADJ_ المؤسسات NOUN_ الدولة NOUN_ الليبية ADJ_ و PUNCT_، NOUN_ الحكومة DET_ هذه NOUN_ مظلة NOUN_ لم PART_ تبق VERB_ إلا PART_ المؤسسة NOUN_ العسكرية ADJ_ PUNCT_.	ARA01A_POS
词形还原	قال رئيس حُكومة ليبيّ عبد الحميد الدبيبة خميس إنّ هُوَ تمّ تَوحيد أكثر من 80% من مُؤسّسة دَوْلَة ليبيّ تحت مِظَلّة هَذا حُكومة، وَلَمْ يَبقِ إلّا مُؤسّسة عَسْكَريّ.	ARA01A_LEMMA

目前，arGLOBE语料库已上传至“北外CQPweb多语种语料库平台”，平台提供简单查询（simple query）和CQP专属检索语法（CQP syntax）模式供不同需求的用户使用，用户可根据自身研究目的进行检索关键词、计算特定词语的典型搭

配、生成词频表等操作。详细的CQPweb平台使用说明可参考北外语料库语言学网站发布的使用手册（<http://corpus.bfsu.edu.cn/info/1082/1875.htm>）。CQPweb是基于网络的第四代语料库分析工具，其采用的索引技术（indexing）能预先对语料库进行数据索引，使上亿词级的复杂检索得以在短时间内完成，有利于促成大规模的语料库研究。CQPweb将语料文本按特定格式存储于服务器，用户只需联网即可进行语料分析。这一操作使得搜索结果只显示有限的上下文语境，用户无法根据检索结果重构语料库，进而在为其他研究者提供复制和验证前人研究的机会同时规避了版权问题（许家金、吴良平 2014：12）。

## 4 研究展望

### 4.1 arGLOBE语料库的应用

arGLOBE为百万词级的当代阿拉伯语书面语平衡语料库，收录了多来源、多体裁的阿拉伯国家出版物语料，具有良好的语料代表性和平衡性。利用该库可开展诸多语料库驱动的研究和基于语料库的研究，为阿拉伯语语言本体研究、语言学与其他邻近领域的跨学科研究提供机遇。

在语言本体层面，该库的词频表、短语表可作为阿拉伯语词典编纂、教学研究的参考，词性标注信息可用于阿拉伯语结构复杂度的分析，研究者还可对关键词及其上下文语境进行标注以开展丰富的语体研究或语法分析。此外，该库按照布朗语料库的采样标准创建，与相同采样标准的布朗家族语料库、北外全球语料库集群等共为可比语料库，研究者还可利用该特征开展阿拉伯语与汉语、英语等语言的对比研究，以及多语种语言类型学研究，促进国内外语料库建设与应用的交流。

语料库研究也可向邻近学科延伸并与之构成学科界面，这些领域包括但不限于政治学、社会学、翻译学、文学、传播学等。研究者可提取该库中不同子体裁的语料进行搭配词计算、索引行分析等操作，对该体裁下的语料进行话语分析与讨论，或进行语域变异研究。此外，研究者可自建小型语料库，并将arGLOBE作为参照语料库进行关键词提取，进而开展符合自身需要的研究。

### 4.2 相关语料库的建设发展

1.0版本的arGLOBE语料库在一定程度上为阿拉伯语语料库语言学研究提供了新的机遇，但与汉语语料库、英语等西方诸语言语料库相比，阿拉伯语语料库的发展仍有极大提升空间。为促成各语种研究间更广泛的学术对话并更好地挖掘阿拉伯语语料库的价值，更多类型的语料库仍有待建设。



arGLOBE 语料库收集的语料大多为 2010 年至 2022 年的当代阿拉伯语书面语文本，这些语料仅能反映一个有限时期的语言面貌，难以表现阿拉伯语长期的动态发展。因此，可参考布朗家族语料库的发布时期建设反映不同年代语言使用特点的阿拉伯语语料库，或文本发布时间跨度更大的语料库，以开展对阿拉伯语语言结构变化、文本语用变化的历时研究。此外，也可开发监控语料库（monitor corpus），持续动态地收集相关领域的阿拉伯语语料，进而更为全面地反映语言使用的全貌。

除了平衡语料库外，各领域的专用阿拉伯语语料库也同样值得关注。例如，中国阿拉伯语学习者的写作语料可用于建设学习者语料库，研究者可结合二语习得的理论和假说开展进一步语言学分析并服务于外语教学；阿拉伯国家各时期的文学作品可用于建设专门的文学语料库，分类依据可以是作品、国别、作家等，进而促成更深入的文体学研究；除书面语语料库外，口语语料也具有重要地位，阿拉伯国家突出的双言现象使得建设阿拉伯语方言口语语料库具有重要价值，有助于推动基于口语语料特征分析的社会语言学研究。

## 5 结语

阿拉伯语语料库研究仍是一个新兴领域，相关语料库的建设与发展值得进一步推进。受益于“北外全球语料库集群”项目的开展，arGLOBE 语料库得以建成。该库参考布朗语料库的采样标准进行建设，依托“北外 CQPweb 多语种语料库平台”提供数据检索功能。相关研究者可根据自身研究需要利用 arGLOBE 语料库或其他可比语料库，开展受语料库驱动或基于语料库的研究，进而推动阿拉伯语的语言本体研究和跨学科研究的发展。此外，“北外全球语料库集群”的建设基于“共建共享”理念，希望并倡导更多同行加入语料库的建设和利用，进一步推动阿拉伯语教学与研究。

### 注释

- 1 参考并改编自：<https://varieng.helsinki.fi/CoRD/corpora/BROWN/basic.html>。

### 参考文献

- ALFAIFI A, ATWELL E, HEDAYA I. Arabic learner corpus (ALC) v2: a new written and spoken corpus of Arabic learners [C]//ISHIKAWA S. Proceedings of learner corpus studies in Asia and the world (LCSAW) 2014. Kobe International Communication Center, 2014: 77-89.
- AL-THUBAITI A. A 700M+ Arabic corpus: KACST Arabic corpus design and construction [J]. Language Resources and Evaluation, 2015, 49: 721-751.

- ARTS T, BELINKOV Y, HABASH N, et al. arTenTen: Arabic corpus and word sketches [J]. Journal of King Saud University - Computer and Information Sciences, 2014, 26(4): 357-371.
- BRUSTAD K. The iconic S ī bawayh [C]//HEIDEMANN S, HAGEN G, KAPLONY A, et al. Essays in Islamic philology, history, and philosophy. Berlin/Boston: Walter de Gruyter GmbH, 2016: 141-165.
- DITTERS E. Arabic corpus linguistics in past and present [C]// CARTER M, VERSTEEGH K. Studies in the history of Arabic grammar II. Amsterdam: John Benjamins Publishing Company, 1990: 129-141.
- DUKES K, HABASH N. Morphological Annotation of quranic Arabic [C]//CALZOLARI N, CHOUKRI K, MAEGAARD B, et al. Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10). European Language Resources Association (ELRA), 2010: 2530-2536.
- HAFF K, JARRAR M, HAMMOUDA T, et al. Curras + baladi: towards a levantine corpus [J]. 2022. *arXiv: 2205.09692v1*
- JARRAR M, HABASH N, ALRIMAWI F, et al. Curras: an annotated corpus for the Palestinian Arabic dialect [J]. Language Resources and Evaluation, 2017, 51: 745-775.
- JARRAR M, ZARAKET F, HAMMOUDA T, et al. Lisan: Yemenu, Irqi, Libyan, and Sudanese arabic dialect copora with morphological annotations [J]. 2022. *arXiv: 2212.06468v2*
- MCENERY T, HARDIE A, YOUNIS N. Introducing arabic corpus linguistics [C]// MCENERY T, YOUNIS N, HARDIE A. Arabic Corpus Linguistics. Edinburgh University Press, 2018: 1-16.
- MCNEIL K. Tunisian arabic corpus: creating a written corpus of an 'unwritten' language [C]//MCENERY T, YOUNIS N, HARDIE, A. Arabic corpus linguistics. Edinburgh University Press, 2018, 30-55.
- PARKINSON D. Under the hood of arabiCorpus [C]//MCENERY T, YOUNIS N, HARDIE A. Arabic corpus linguistics. Edinburgh University Press, 2018: 17-29.
- ZIEMSKI M, JUNCZYS-DOWMUNT M, POULIQUEN B. The United Nations parallel corpus v1. 0 [C]//CALZOLARI N, CHOUKRI K, DECLERCK T, et al. Proceedings of the tenth international conference on language resources and evaluation (LREC'16). European Language Resources Association (ELRA), 2016: 3530-3534.
- 许家金, 吴良平. 基于网络的第四代语料库分析工具CQPweb及应用实例[J]. 外语电化教学, 2014 ( 5 ): 10-15.

通信地址: 100089 北京市 北京外国语大学阿拉伯学院

# faGLOBE 当代波斯语书面语平衡语料库的创建<sup>\*</sup>

北京外国语大学 李彦军 陈帅楠 胡 奇 周汀鹭

**提要：**faGLOBE 当代波斯语平衡语料库是“北外全球语料库集群”项目（又称“GLOBE 语料库”项目）下的一个子课题，旨在收集近十年伊朗所使用的波斯语书面语文本，创建百万词级的平衡语料库。本文首先简述当前国际上波斯语的语料库建设情况，在此基础上，从语料采集、文本录入和元信息标注等方面对 faGLOBE 的建库过程进行较为全面的论述，并对基于该语料库的语言研究、教学及共享作出规划。

**关键词：**faGLOBE 语料库、当代波斯语书面语平衡语料库、波斯语教学与研究

## 1 面向波斯语的语料库建设

faGLOBE 波斯语平衡语料库是“北外全球语料库集群”中首个建成的非通用语种语料库，收集的全部为伊朗使用的现代波斯语语料。

### 1.1 已有波斯语语料库的建设情况

波斯语属于印欧语系，是伊朗和塔吉克斯坦的官方语言，也是阿富汗境内两种主要语言之一。波斯语为拼音文字，共有 32 个字母，其中 28 个是阿拉伯字母，4 个是波斯字母。现有的波斯语语料库种类繁多，主要用于语言学研究、波斯语教学、人工智能开发等目的，在此根据不同类型简单举例如下。

共时语料库：波斯语语料库 Peykare (Bijankhan *et al.* 2011: 143)，包含约 1.1 亿字的现代波斯语书面和口语文本，根据事实性、格式、风格和语言材料等标准进行分类，约有 1,000 万个单词被随机选择并根据 EAGLES Guidelines 进行标注。

<sup>\*</sup> 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）及北京外国语大学 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。李彦军是本文通讯作者。

作者贡献：

李彦军：选题构思、研究方法、讨论结论、修改润色；

陈帅楠：数据收集、初稿撰写；

胡 奇：数据收集、初稿撰写；

周汀鹭：数据收集、初稿撰写。陈帅楠、胡奇、周汀鹭在本文撰写中的贡献相同。

专用语料库：波斯语通讯社的标题语料库 *پیکره تیتیر خبرگزاری های فارسی زبان* (۱۳۹۵ میرزایی، آ. و صفری، پ) 汇编了来自13个著名波斯语官方新闻机构的110,198个标题，包括标题、新闻导语、新闻主要分组、新闻子类别或次要、刊发日期等，总字数超过一百万。德黑兰大学工程学院数据库研究小组 (DBRG) 开发的Hamshahri语料库 (AleAhmad *et al.* 2009: 382)，由1996—2002年出版的Hamshahri报纸中共计345MB的新闻文本组成，用于不同研究领域的信息检索。

学习者语料库：伊斯兰科学计算机研究中心制作的波斯语话语语料库 *پیکره گفتامانی زبان فارسی* (Mirzaei & Safari 2018)，对约30,000个波斯语句子进行话语标注，以研究波斯语句子之间的关系和文本逻辑。波斯语句法语料库 *پیکره وابستگی نحوی زبان فارسی* (Rasooli *et al.* 2013: 306) 是波斯语的第一个句法语料库，采样自各种来源的当代波斯语文本，包括约30,000个已标记的句子，所有句子都标有句法关系和词性。

口语语料库：萨罕德科技大学基于五种波斯语口音（阿塞拜疆口音、马赞德兰口音、库尔德口音、德黑兰口音、伊斯法罕口音）开发的萨罕德口音语音数据库 *دادگان گفتار لهجه دار سهند* (Sedaaghi 2008)，主要用于语音处理领域，尤其是口音识别的研究。谢里夫情感语音数据库 Sharif Emotional Speech Database (Mohamad Nezami *et al.* 2019)，涵盖了87位波斯语说话者共计3,000个带有不同情感的语音片段，包括“愤怒”“恐惧”“快乐”“悲伤”“惊讶”以及中性模式，用于波斯语语音情感检测。今日波斯语会话语料库 *پیکره گفتار محاوره ای زبان فارسی امروز* (بیجن خان، م) (۱۳۹۵) 由350小时的口语数据组成，记录作为志愿者的波斯语使用者在各种交流情况下的声学信号，目的是对会话系统进行应用研究、培训和测试，语料库输出年龄、性别、口音、教育水平、语境类型、话语持续时间、音频文件和相应文本网络、书面标记的文本文件、语料库词汇文件和语料库文档等信息。

此外还有其他类型的波斯语语料库，如波斯手写离散字母数据库 *بانک اطلاعات حروف گسته دستنویس فارسی* (Khosravi *et al.* 2007: 1)，由10,023,640个图像样本组成，总计约120 GB。库中的图像以BMP格式呈现，分辨率为300 dpi，灰度为256级，用于帮助开发和训练波斯语光学字符读取 (OCR) 系统。

## 1.2 faGLOBE 语料库的建设情况

faGLOBE 语料库是北京外国语大学“全球语料库集群”（又称“GLOBE 语料库”）中的一个子项目，总容量约为100万词。语料库搜集的绝大部分语料发表于2010至2022年间，少量文本发表于本世纪最初几年。

本语料库是按照布朗语料库模式创建的当代波斯语平衡语料库。布朗语料库 (The Standard Corpus of Present-Day Edited American English) 是1961—1964年美国布朗大学Francis和Kučera教授建设的根据系统性原则采集样本的机读语料库，

是世界上第一个计算机可读的语料库，标志着语料库建设进入了电子化时代。布朗语料库对欧洲的计算机语料库建设和语料库语言学的发展起到了重要的引领和催化作用，特别是其语料取样方法为欧洲一系列语料库的建设提供了重要借鉴（Leech & Johansson 2009）。在布朗语料库诞生后的数十年中，语料库建设蓬勃发展，数个语料库根据布朗语料库的设计理念先后建成，包括英国兰卡斯特大学的LOB语料库、法国国家科学研究中心与美国芝加哥大学的法语TLF语料库、芬兰赫尔辛基大学的历史语料库、英国伦敦大学的国际英语语料库等。

faGLOBE语料库的取样标准使之与现有的Crown和CLOB等布朗家族语料库具有高度可比性，可搭配进行搜索和分析。目前faGLOBE语料库包含生语料版本，提供波斯语词频表与短语列表，可供教学与研究之用，语料库已上传至“北外多语种语料库平台”（<http://114.251.154.212/cqp/>）。

faGLOBE语料库收集的语料类型主要分为信息类和虚构类两种类别，其中信息类包含新闻和通用两种体裁，虚构类包含学术和小说两种体裁，语料库共搜集2,000词语料500篇。具体语料类型及相应篇数详见表1。

表1 语料类型及相应篇数<sup>1</sup>

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	
		D	宗教	17	
		E	日常技艺及消遣爱好	36	
	通用	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
	小说	P	言情小说	30	
		R	幽默	9	
合计				500	

与原始布朗语料库相比，faGLOBE语料库调整了收集语料的类别，使之更符合对象国的国情和文化。例如，faGLOBE中删去了不符合对象国国情的“冒险或西部小说”，转而将其替换为“历险悬疑小说”；将“教科书”替换为覆盖范围更广的“学术”；将“混杂的文本”替换为“政府或机构公文及文宣”等。

与已有的波斯语语料库相比，faGLOBE语料库选取的底层文本时效性更强、内容更广。faGLOBE语料库录入的底层文本种类多样，不仅包含已有语料库收集的新闻类文本，还包含大量虚构性文本；这些文本生成时间多集中在近十年，可以更好地反映当代波斯语的特点。

faGLOBE语料库按照上述采集方案，收集了500篇约2,000词的波斯语文本，并对其进行初步处理。波斯语语料采集工作主要由3个环节构成：文本收集与取样、文本录入和语料库元信息标注。

## 2 faGLOBE语料库的创建

### 2.1 文本收集与取样

在faGLOBE语料库1.0版本的规划中，主要面向伊朗本土的波斯语文本，本语料库所收录文本的第一作者原则上均为伊朗人，且所收集的文本均为原创的波斯语文本，由其他语言翻译到波斯语的译本不包含在本语料库的收集范围之内。此外，流通中的波斯文本总量（特别是波斯语文本的网络资源）普遍少于英文文本，考虑到语料收集的可操作性，faGLOBE语料库相比布朗家族语料库扩大了语料收集的时间范围：考虑到新闻类语料的时效性较强，因此所选取语料多为语料收集当年（即2022年）的最新报道，偶有过去四年的内容。其他3类体裁（通用、学术、小说）所收大部分文本的首次出版时间为近十年左右，部分难以收集文本的范围扩大到了近二十年，但此类文本占比较少。所收集语料的时间与数量分布可见图1。

为平衡语料库规模，尽可能扩大所收集语料的范围，除考虑语料的时间数量分布外，faGLOBE在建库过程中还兼顾了语料的来源及其主题的多样性。针对新闻类语料，兼顾多家伊朗本土媒体，如IRNA（伊朗伊斯兰共和国通讯社）、ISNA（伊朗大学生通讯社）、Tasnim News（塔斯尼姆通讯社）、Fars News（法尔斯通讯社）、Mehr News（梅赫尔通讯社）等，同时涵盖政治、经济、卫生、医疗、文化、社会、旅游、运动等多种话题，在进行语料收集时，还着重收录了伊核协议、美伊关系、中伊关系、“一带一路”等与伊朗密切相关的热点话题语料。对于学术类语料，综合考虑核心期刊文章、普通期刊文章、学术专著或文集、学术会议中的文章，涉及自然科学、医学、数学、机械、社会学、人类学等不同学科门类及研



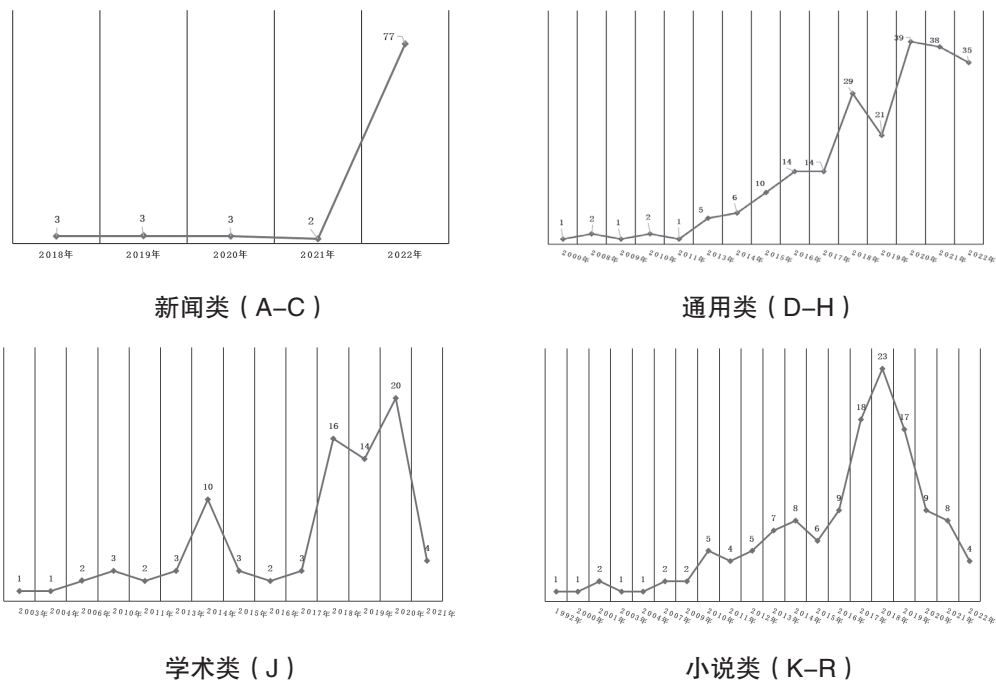


图1 四类体裁所收集语料时间数量分布<sup>2</sup>

究领域，语料多源自科学信息数据库（SID）、伊朗文化研究杂志（JICR）、综合人文门户网站（ENSANI）、MagIrans等伊朗数据库。至于通用类、小说类中的书籍，则源自多家出版社，涵盖不同的题材，语料多源自Fidibo、绿皮书（Ketabesabz）、DL图书（Dlketab）、伊朗图书（Iranketab）等多家伊朗本土电子书网站。通用类下的子体裁H“政府或机构公文及文宣”则摘自伊朗多个政府部门、多家企业及机构。

## 2.2 文本录入

faGLOBE语料库所收文本长度为2,000词左右，当涉及篇幅较短的语料体裁类型时，如部分新闻报道、社论、政府或机构文书及文宣等，则将多篇同类型的语料拼合起来，并将用于拼合的语料单独保存，在其文件名末尾添加A/B/C/D等大写英文字母以示区分。如faGLOBE库所收的第二号文本FAA02即包含四篇新闻，因此存储为四个独立的txt文件：FAA02A、FAA02B、FAA02C、FAA02D，这四篇语料共同构成表1子体裁A下的第二个语料文本。所有语料均经过人工校对，并以UTF-8编码格式储存为txt文本文件。

## 2.3 元信息标注

总体来看,当前 faGLOBE 语料库包含 500 个 2,000 词长度的文本,共存储为 798 个语料子文件,其命名遵循“两位字母语种代码——位字母体裁编码——数字编号——字母编号”的格式。通过这种方式,仅从文件名就能知晓语料所属体裁信息。除文件名本身承载的分类信息外,faGLOBE 语料库还以 Excel 表格的形式记录元信息标注,以便为后续研究提供更多的语料信息。

针对所有语料文件,faGLOBE 提供文件编号 (File Name)、文本类型 (Category)、文本数 (Text Count)、词数 (Word Count)、作者 (Author)、出版年份 (Year)、出版社或报刊名称 (Publisher)、语料收集人 (Collector)、文章或书籍标题 (Book Name or Article Title) 和网址出处 (URL) 10 项基本信息。

## 3 研究展望

### 3.1 基于波斯语语料库的教学研究应用

faGLOBE 语料库 (1.0 版) 采集的文本涵盖新闻报道、社论、宗教、传记、科技、公文等 15 个类别,选取的文本绝大部分发表于 2010—2022 年,语言资料真实客观,对于波斯语学习者而言是宝贵的学习资源,可以作为课后资料补充,扩展语言知识,也可以作为实用的检索工具,帮助学生自行对语言实例进行归纳,从而增强其自主学习的能力。教师可以利用语料库词频功能列出波斯语中最常见的词汇,借助语料库方法进行词汇教学、大纲词汇的选择和分级辅助教学 (许家金 2009: 45)。

本语料库还可以充当波斯语课堂的教学工具,教师利用该语料库可以将大量的语料和实例迅速、直接地展示给学生,从而使学生在真实鲜活的语境中掌握相关词汇的语义,并总结使用规律,同时强化学生对波斯语中的语言现象以及重点难点的理解。例如波斯语中常见的“一词多义”“一词多音”,可以通过语料库丰富的例证深化学生的记忆,并激发课堂上的师生互动。

语料库对语言学研究大有裨益,如词汇语法研究,基于此,在本语料库的帮助下,波斯语教学者可以将词汇教学的范畴扩大,利用本语料库进行词语搭配、语义韵、词语辨析教学 (王家义 2012)。此外,考虑到本语料库容纳文本量较大,语料涉及场景多样且难度不一,教学者亦可依据这些文本资源为不同学习年龄段学生的相关课程 (如阅读、写作课等) 制定教学大纲,设计教学活动,以及编写教材等。

### 3.2 基于波斯语语料库的语言学应用

近年来,语料库结合分析软件的使用可以帮助研究者更清晰地从语言学角度辨别语言形式(钱毓芳 2010: 199)。研究者可利用faGLOBE语料库提供的词频和搭配功能辨别波斯语词汇的总体分布情况。当与其他波斯语语料库进行比较时,研究者还可通过对比词频获知语料库之间的差异性。对语体进行历时性分析时,可以获知波斯语语体的总体变化趋势。

除此以外,研究者可以考察某一波斯语词汇左右搭配的惯用情况,即词丛。词语搭配之父Firth曾说过,词的意义在与它结伴出现的词中体现(Firth 1957: 182)。词的结伴规律、结伴词项间的相互期待与吸引、搭配成分类联接关系等都是词语搭配研究的重要内容(卫乃兴 2002: 101)。在此基础上,研究者可对文本进行索引分析,并分析波斯语的话语韵和语义偏向等特征,了解某一波斯语词汇搭配的常见形式和规律,并掌握词汇与语言使用者态度之间的关联。

### 3.3 后续语料库的建设

faGLOBE语料库目前只包含生语料版本,暂未完成对语料的词形还原、词性赋码、句法标注和语义标注等深层处理。在语料库建设中,生语料指未进行语言学标注的版本,词性赋码指标注生语料中单词的词性,词形还原指在自然语言处理中,去掉单词的词缀,提取单词的主干部分,通常提取后的单词会是字典中的单词。经过词形还原的语料库可以减少单词的变化形式,统一单词的表示,方便后续的分析 and 处理。例如,动词的过去式可以还原为原形。未来,faGLOBE语料库可以结合许家金、梁茂成教授开发的Tree Tagger等自然语言处理软件对语料库中的生语料进行标注,不同版本的语料将为使用者提供更丰富的使用场景,也将有助于语料库更好地识别单词在语境中的含义。

## 结语

faGLOBE语料库作为我国第一个波斯语语料库,以真实波斯语语言数据为研究对象,将从宏观角度对波斯语语言事实、语言交际和语言学习进行多层面研究。语言是社会及人类生活的一面镜子,本语料库的上线将加深我国波斯语学习者对伊朗政治、社会、历史、文化等多方面的认识。我们应进一步加强faGLOBE语料库的资源共建和共享,运用技术创新不断丰富语料库的内容和使用方式,让faGLOBE语料库成为助力波斯语教学和伊朗研究的有力工具。

### 注释

- 1 改编自 <https://varieng.helsinki.fi/CoRD/corpora/BROWN/basic.html>。

- 2 由于存在单篇语料词数少于2,000词的情况,需由多篇语料拼接成一个文本,此处所统计为实际收录的语料篇数,因此折线图中语料数量与表1中的数量有所出入。

### 参考文献

- ALEAHMAD A, AMIRI H, DARRUDI E, et al. HAMSHAHRI: a standard Persian text collection [J]. Knowledge-Based Systems, 2009, 22(5): 382-387.
- BIJANKHAN M, SHEYKHZADEGAN J, BAHRANI M, et al. Lessons from building a Persian written corpus: Peykare [J]. Language Resources and Evaluation, 2011, 45(2): 143-164.
- FIRTH J. Paper in linguistics 1934-1951 [M]. London: Oxford University Press, 1957.
- KHOSRAVI S, RAZZAZI F, REZAEI H, et al. A comprehensive handwritten image corpus of isolated Persian/Arabic characters for OCR development and evaluation [C]//2007 9th International Symposium on Signal Processing and Its Applications. Sharjah, United Arab Emirates: IEEE, 2007: 1-4.
- LEECH G, JOHANSSON S. The coming of ICAME [J]. ICAME Journal, 2009, 33: 5-20.
- MIRZAEI A, SAFARI P. Persian discourse treebank and coreference corpus [C]// CALZOLARI N, CHOUKRI K, CIERI C, et al. Proceedings of the Eleventh International Conference on Language Resources and Evaluation. Miyazaki, Japan: European Language Resources Association (ELRA), 2018.
- MOHAMAD NEZAMI O, JAMSHID LOU P, KARAMI M. ShEMO: a large-scale validated database for Persian speech emotion detection [J]. Language Resources & Evaluation, 2019, 53(1): 1-16.
- RASOOLI M, KOUHESTANI M, MOLOODI A. Development of a Persian syntactic dependency treebank [C]//VANDERWENDE L, DAUMÉ III H, KIRCHHOFF K. Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Atlanta, Georgia: Association for Computational Linguistics, 2013: 306-314.
- SEDAAGHI M. Documentation of the Sahand emotional speech database (SES) [R]. Technical Report, Department of Electrical Engineering, Sahand University of Technology, Iran, 2008.
- بی‌جن‌خان، م. پیکره گفتار محاوره‌ای زبان فارسی امروز [C] // مجموعه مقالات دومین همایش ملی زبان‌شناسی پیکره‌ای. تهران، ایران: نشر نویسه پارس، ۱۳۹۵.
- میرزایی، آ. و صفری، پ. پیکره تیتتر خبرگزاری‌های فارسی‌زبان [C] // مجموعه مقالات دومین همایش ملی زبان‌شناسی پیکره‌ای. تهران، ایران: نشر نویسه پارس، ۱۳۹۵.
- 钱毓芳. 语料库与批判话语分析[J]. 外语教学与研究, 2010 ( 3 ): 198-202.

- 王家义. 基于语料库的英语词汇教学: 理据与应用[J]. 外语学刊, 2012 (4): 127-130.
- 卫乃兴. 基于语料库和语料库驱动的词语搭配研究[J]. 当代语言学, 2002 (2): 101-114.
- 许家金. 词汇中心教学法的交际观——理论溯源与反思[J]. 中国外语教育, 2009 (4): 38-45.

**通信地址:** 100089 北京市 北京外国语大学亚洲学院

# itGLOBE 当代意大利语书面语平衡语料库的创建<sup>\*</sup>

北京外国语大学 喻儒辰 董 丹 郭焱一

**提要：**itGLOBE 当代意大利语书面语平衡语料库是“北外全球语料库集群”项目（又称“GLOBE 语料库”项目）下的一个子课题，也是该项目首批建成的百万词级平衡语料库之一。本文简述了意大利语语料库建设历程，并在此基础上讨论 itGLOBE 的语料采集方案及编码过程。此外，本文对基于 itGLOBE 的研究及后续语料库建设作出了展望。

**关键词：**itGLOBE、意大利语书面语平衡语料库、意大利语教学与研究

## 1 引言

“itGLOBE 当代意大利语书面语平衡语料库”（以下简称“itGLOBE 语料库”）是“北外全球语料库集群”项目首批建成的百万词级语料库之一。该语料库集群旨在囊括北外开设的101个语种，借鉴布朗语料库的采样方案建成一系列以“GLOBE（Global Languages Out of BFSU Expertise）”命名的单语平衡库。GLOBE 家族语料库与现有的布朗家族语料库具有可比性，可促成双语或多语对比研究。本文将着重介绍 itGLOBE 语料库的建库背景及过程，并讨论其应用价值。

## 2 意大利语语料库建设历程

意大利是语料库语言学研究的前驱地之一。20世纪40年代末，意大利耶稣会士罗贝尔托·布萨（Roberto Busa）使用计算机编写圣托马斯·阿奎那（St Thomas Aquinas）的拉丁语著作索引，并于1951年出版了初步成果（Busa 1951）。学界普遍认为，这标志着世界上首个机器可读语料库的诞生（McEnery & Hardie

<sup>\*</sup> 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）及北京外国语大学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。本文由中央高校基本科研业务费专项资金资助（2023JX041）。董丹是本文的通讯作者。

作者贡献：

喻儒辰：选题构思、数据收集、初稿撰写（文字占比50%）；

董 丹：数据收集、初稿撰写（文字占比25%）、修改润色；

郭焱一：数据收集、初稿撰写（文字占比25%）。



2012: 37)。此后,意大利语语料库建设大致经历了起步、注重词典编纂和全面发展三个阶段。

20世纪50年代初到70年代初是以著作索引编写为目标的起步阶段。这一阶段,大批学者开始搜集和整理个人语料,秕糠学会等权威研究机构也将目光投向语料库语言学,几所大学陆续开设相关实验室以开展语料库研究工作(董丹 2022)。例如,比萨大学电子计算中心和帕多瓦大学均受布萨研究成果的影响,在秕糠学会和IBM意大利分公司的帮助下整理编写意大利语经典著作索引(Zampolli 2004)。此外,维托里奥·桑托利(Vittorio Santoli)等学者组织了意大利民歌目录的编纂项目,并开展了相关的词频及韵律研究(Duro 1968)。

20世纪70年代初至90年代末的主要任务是基于语料库的词典编纂。该阶段的首个重要成果是比萨大学电子计算中心基于五万词电子语料库制成的《当代意大利语词频库》(Lessico di frequenza della lingua italiana contemporanea,简称LIF)(Bortolini *et al.* 1971)。LIF的语料库已经初具平衡思想,其文本由5种不同体裁的书面意大利语构成。受LIF影响,图里奥·德·毛罗(Tullio De Mauro)先后基于千万词的新闻语料库和57小时的口语语料库编写了《词汇使用指南》(Guida all'uso delle parole)(De Mauro 1980)和《口头意大利语词频库》(Lessico di frequenza dell'italiano parlato,简称LIP)(De Mauro *et al.* 1993),并在两部著作的基础上编纂了影响深远的《意大利语实用大词典》(Grande dizionario italiano dell'uso)(De Mauro 1999)。

20世纪90年代末起,意大利公开的大规模语料库建设日臻成熟,逐渐形成了以权威机构为依托、分类明确、功能多样的语料库格局。所涉及的权威研究机构和项目主要有:比萨计算语言学协会、秕糠学会意大利词语工程项目、德·毛罗学术项目、比萨高等师范学校语言实验室、都灵大学Unito项目、博洛尼亚大学系列语料库和意大利语言实验室。已建成的大型语料库包括但不限于:(1)书面语语料库,如意大利首个大型参照语料库——笔语参照语料库及动态笔语语料库CORIS/CODIS<sup>1</sup>、意大利语书面语料资源语料库NUNC(Newsgroups UseNet Corpora)、学术论文语料库Athenaeum、意大利文学语料电子库CRILet、《共和国报》语料库以及历时意大利笔语参照语料库DiaCORIS等。(2)口语语料库,如外国人口语语料库LIPS、C-ORAL-COM语料库、口语种类档案库AVIP、意大利语口语档案API、意大利口笔语语料库及词库CLIPS<sup>2</sup>等。(3)学习者语料库,如变异意大利语习得库VALICO和意大利语等级考试语料库CELI等。(4)特殊用途语料库,如儿童语言库CHILDES-Italia(该语料库可反映不同健康状况儿童的语言使用情况)、精神分裂症患者口语语料库、意大利语手语及病理性语言语料库、法律术语语料库、医学术语语料库等。

意大利语语料库对于意大利语语言资源的分类、储存与保护起到了重要作用,

可用于语言学习、意大利语研究、语言资源记录与保护等（董丹 2022）。意大利语语料库在语料库建设和使用过程中体现了较强的自然语言处理技术，各项语料库研究间也具有较强的传承性，著作索引、LIF、LIP等成果充分体现了“语料库作为方法（corpus-as-method）”的应用价值。此外，意大利语语料库的建设理念也相对统一，从LIF到CORIS/CODIS，从LIP到CLIPS，平衡性、代表性等原则贯穿始终。不论从研究成果还是建库模式上看，面向意大利语的语料库建设都已相对成熟，但相关语料库大多使用独立的取样方法，难以与现有的英语（如布朗语料库）及汉语语料库进行直接对比。本文介绍的itGLOBE语料库可以有效填补这一空缺。此外，itGLOBE也为国内意大利语研究提供了重要的平衡语料资源。

3 itGLOBE 语料库的创建

itGLOBE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2013年之后出版和发表的意大利语书面语文本。目前，itGLOBE语料库已上传至“北外CQPweb多语种语料库平台”（<http://114.251.154.212/cqp/>），该平台在线提供索引分析、搭配计算、词表生成和主题词分析等功能。此外，itGLOBE还随库提供意大利语词频表与短语列表，供教学与研究之用。

3.1 采样方案

itGLOBE语料库借鉴布朗语料库的采样方案，所收文本类型及相应篇数参见表1。

表1 itGLOBE 语料库文本类型及相应篇数<sup>3</sup>

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
信息类 (374篇)	通用	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

（待续）

(续表)

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）
虚构类 ( 126 篇 )	小说	K	一般小说	50
		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

itGLOBE 语料库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, 并在 4 种体裁类型下进一步细分出 15 个子体裁类型。对于个别子体裁类别的语料采样, 课题组根据意大利语言文化状况进行调整。例如, 在子体裁类型 L “侦探小说” 类别下, itGLOBE 语料库主要收集本土侦探/犯罪小说, 符合意大利小说类别划分标准, 与相应语言文化背景契合, 具有明显的意大利语语料特征。此外, 在语料文本选用与采样方面, 部分收入与时事热点及社会文化环境变化相关的语料, 主要集中于子体裁类别 A、B、J 下, 以进一步保证及维护 itGLOBE 语料库的平衡性与代表性。

### 3.2 语料采集

itGLOBE 语料库的语料采集工作基于以上适用于意大利语的采样方案, 依据该方案采集 500 篇词数为 2,000 词 (+/-50 词) 的意大利语文本, 在其基础上进行初步加工。意大利语语料采集主要包括文本选择与取样、文本录入以及语料库元信息标注 3 个主要环节。

#### 3.2.1 文本收集与取样

itGLOBE 语料库选用的文本均为意大利语本土原文语料, 不涉及由其他语言译入意大利语的翻译文本。此外, 意大利语文本整体流通量低于英语文本, 考虑到语料收集的可操作性, 相较于布朗家族语料库而言, itGLOBE 语料库所收录语料文本的年限范围更大: 新闻体裁类型文本发布时间均集中于 2019—2021 年; 通用、学术、小说三类体裁类型下, 主要选用发布时间为 2013—2022 年的文本, 仅存在个别发布时间早于 2013 年的文本, 所有文本均发布于 2000 年后。四类体裁选用语料文本时间与数量分布<sup>4</sup>如图 1 所示。

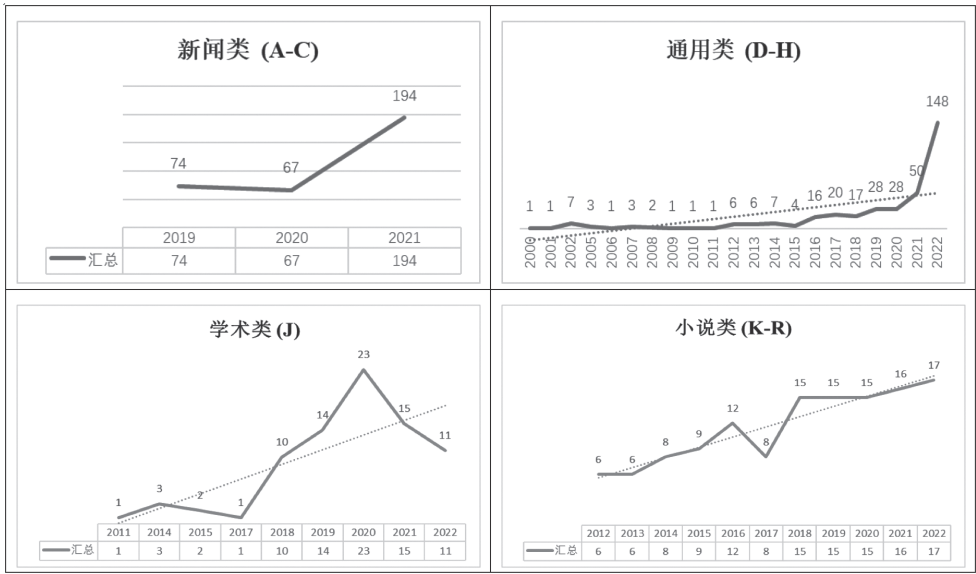


图1 四类体裁选用语料文本时间与数量分布

在itGLOBE建库过程中，课题组兼顾语料来源、主题、内容及时间、空间分布等特征的多样性。新闻类语料文本的选用覆盖意大利各大规模时报，新闻文本均来自大、中型报社媒体，兼顾意大利全国性与地方性（意大利南、北方及各大区专刊）新闻报道；同时，主题内容涵盖时事政治、国际热点、经贸、科技、文化、体育等多个方面，出处包含综合型报刊与专业型报刊（如体育、经济等）。表2为新闻类语料文本来源及地区分布。

表2 新闻类语料文本来源

来源类型	综合型报刊	地方报刊	专门型报刊
数量	212	97	26

学术类文本主要来自意大利各大高校论文库、论文检索门户网站及各学科期刊。文本内容主题广泛，涉及时事热点、国际政治、文学、历史、社会学、数学、物理、天文学等多门学科及研究领域；针对通用类文本，课题组均从意大利线上媒体网站收录语料。其中，子体裁H“政府或机构公文及文宣”语料文本的主要来源为意大利各政府部门、大区政府及多个机构；在学术类、小说类文本语料的收集过程中，针对篇幅较长的语料篇目，选择截取前、中、后三部分拼接成词数为2,000（+/-50）的语料文本，以增强语料的代表性，维护语料库的平衡性。

3.2.2 文本录入及元信息标注

itGLOBE 语料库收录文本词数均为 2,000 (+/-50) 词，通过正则表达式 [A-Za-zàéèìòùÈ0-9-']<sup>+</sup> 进行计数（定冠词与其他单词以 “'” 相连时，整体视为一个单词）。对于部分文本篇幅较短的体裁类型，如新闻类、通用类，将多个文本拼接合并为一个满足收录词数要求的文本，每一个用于拼接语料篇目的子文本单独保存为一个文件，并在其末尾进行 A/B/C 等字符标注，以示区分。例如：itGLOBE 语料库所收新闻类体裁的第一篇文本 ITA01 由五篇各自独立的新闻文本拼接而成，编号为 ITA01A、ITA01B、ITA01C、ITA01D、ITA01E。所有语料均经过人工校对，以 UTF-8 编码格式储存为 txt 文本文件。

针对所有收录的文本，itGLOBE 均注明文件编码（File Name）、出版方（Publisher）、书名或文章标题（Book Name or Article Title）、资源链接（URL）、发布年份（Year）、作者（Author）以及检索来源（Source）7 项语料基本信息，存储于独立的 Excel 表格中。

3.3 语料版本及应用

itGLOBE 为意大利语研究者及学习者提供了生语料、词性赋码和词形还原三个不同的版本。其中，生语料指未添加任何语言学标注的文本，可以帮助使用者迅速掌握单词或短语所在的上下文语境。词性赋码、词形还原则是文本经过自然语言处理工具包 spaCy 处理后的版本，能够为语言学研究提供便利。各版本语料示例见表 3。

表 3 三种语料版本示例

语料版本	示例	来源文件
生语料	La notizia anticipata dal Daily Mail ha già registrato i primi commenti negativi	ITA10A
词性赋码	La_DET notizia_NOUN anticipata_VERB dal_ADP Daily_PROPN Mail_PROPN ha_AUX già_ADV registrato_VERB i_DET primi_ADJ commenti_NOUN negativi_ADJ	ITA10A_POS
词形还原	il notizia anticipare dal Daily Mail avere già registrare il primo commento negativo	ITA10A_LEM

表 3 以 ITA10A 文件中的一个小句为例，展示了 3 个不同版本的语料。其中，词性赋码版语料的基本格式为“单词\_词性码”。例如，定冠词 la 的赋码为 DET、过去分词 anticipata 和 registrato 的赋码为 VERB、助动词 ha 赋码为 AUX、名词为

NOUN、形容词为ADJ、副词为ADV等（完整的词性赋码集见表4）。该版本的语料可以帮助研究者快速检索特定句式。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，如将阴性定冠词la替换为阳性定冠词il，将过去分词anticipata替换为不定式anticipare，将阳性复数形容词negativi替换成阳性单数原形negativo等。在统计词频和检索搭配时，该版本的语料可以减轻屈折变化的干扰。

表4 spaCy 意大利语词性赋码表

词性赋码	词性	词性赋码	词性
ADJ	形容词	PART	小品词
ADP	介词	PRON	代词
ADV	副词	PROPN	专有名词
AUX	助动词	PUNCT	标点
CCONJ	并列连词	SCONJ	从属连词
DET	限定词	SYM	符号
INTJ	感叹词	VERB	动词
NOUN	名词	X	其他
NUM	数词	SPACE	空格

3个版本的itGLOBE均可在“北外CQPweb多语种语料库平台”上在线使用。平台提供两种检索模式：简单查询（simple query）和CQP语法（CQP syntax），既能满足普通语料库用户检索特定单词、短语的需求，又能实现高级的检索和分析。

4 研究展望

4.1 基于语料库的意大利语研究

itGLOBE语料库的取样具有平衡性和代表性。通过检索itGLOBE全库，可以深入探究2000年后（尤其是近十年）意大利语的词汇、短语、句法等特征。此外，由itGLOBE生成的词频表及n-gram表，可应用于中高级意大利语的课程设计和基于语言实例的词典编纂。itGLOBE语料库中的新闻、通用、学术、小说四大类文本也可单独构成子语料库，助力篇章体裁研究、语域变异研究。

itGLOBE语料库具有可比性，可与布朗家族语料库和其他GLOBE语料库对比使用。例如，itGLOBE可与近期建成的CROWN2021组合使用，开展意英对比



研究；可与ToRCH系列汉语语料库（2009、2014、2019）组合使用，开展意汉对比研究；可与上文提及的英汉语料库及其他GLOBE语料库（如deGLOBE德语语料库）组合使用，开展类型学研究。此外，itGLOBE的取样方案与现有的意大利语参照库近似，其新闻、小说等子语料库可以与CORIS/CODIS及其历时语料库DiaCORIS中的同类别文本组合使用，推动意大利语历时变化研究。

另外，itGLOBE收录的语料均为意大利本土的原创文本，其中的小说子库可与意汉双向平行语料库CHCPC（余丹妮 2020）中的意语翻译文本进行对比，开展语料库翻译学研究。

## 4.2 后续语料库建设

现已上线的itGLOBE语料库仍有待进一步完善。首先，参考deGLOBE的方案（周顾盈等 2022），后续版本可进一步更新语料元信息。例如，收录新闻文本的作者生平及所属栏目、学术文本的DOI编码、小说文本的章节信息等。此外，后续版本应进一步降低2013年前语料的比例，使语料库更精确地反映近十年的意大利笔语使用情况。

此外，还可对itGLOBE进行扩容，促成不同类型的对比研究。当前itGLOBE仅包含笔语，后续可在条件允许的情况下加入口语语料补充库。此外，还可在布朗语料库取样方案的基础上加入近十年的新兴体裁，如在通用类别中加入社交媒体文章、企业可持续发展报告等。

## 5 结语

当前，意大利语语料库的建库理念先进、研究成果丰富，其发展已经进入了成熟时期。然而，直接适用于多语对比研究的平衡语料库仍具有较高的创建价值。在“北外全球语料库集群”项目的支持下，itGLOBE基本依照布朗语料库的取样方法建成，可以直接与布朗家族语料库及一系列GLOBE语料库组合使用，成为现有国内外意大利语语料库的有益补充。itGLOBE语料库可在“北外CQPweb多语种语料库平台”上在线使用，平台提供的检索功能可以满足语言学习者和科研工作者的不同需求。itGLOBE语料库遵循“共建共享”理念，希望并倡导更多同行加入后续建设，为意大利语教学与研究做出新的贡献。

### 注释

- 1 该语料库参考了LIF等语料库的采样方法，每3年更新一次，由专门的配套标记程序对语篇进行注释及词条化处理。截至2021年，累计收录了1,650万词的文本。
- 2 该语料库收录了广播电视节目（16小时30分）、对话（约48小时15分）、阅

读（16小时20分）、正音（3小时40分）、电话（16小时40分）等多种形式的录音。每类录音都在国内15个不同的地点采集，同时保证了性别的平衡，为意大利语的语言变异研究提供了极大的便利（见Albano Leoni 2003）。

- 3 参照、改编自 <https://varieng.helsinki.fi/CoRD/corpora/BROWN/basic.html>。
- 4 此表根据实际收集语料篇数统计——itGLOBE语料库所收四类体裁语料中，新闻类、通用类及文学类均存在单篇语料词数低于2,000词的情况，因此由多篇语料拼接成为一个完整文本。

### 参考文献

- ALBANO LEONI F. Tre progetti per l'italiano parlato [C]//POGGI SALANI T, MARASCHIO N. Italia linguistica, anno Mille, anno Duemila. Roma: Bulzoni, 2003: 1000-1009.
- BORTOLINI U, TAGLIAVINI C, ZAMPOLLI A. Lessico di frequenza della lingua italiana contemporanea [M]. Milano: IBM Italia, 1971.
- BUSA R. Rapida e meccanica composizione e pubblicazione di indici e concordanze di parole mediante macchine elettrocontabili [J]. Aevum, 1951, 25(6): 479-493.
- DE MAURO T. Guida all'uso delle parole [M]. Roma: Editori Riuniti, 1980.
- DE MAURO T. Grande dizionario italiano dell'uso (GRADIT) [Z]. Torino: UTET, 1999.
- DE MAURO T, MANCINI F, VEDOVELLI M, et al. Lessico di frequenza dell'italiano parlato [M]. Milano: Etas, 1993.
- DURO A. Humanities computing activities in Italy [J]. Computers and the Humanities, 1968, 3(1): 49-52.
- MCENERY T, HARDIE A. Corpus linguistics: method, theory and practice [M]. Cambridge: Cambridge University Press, 2012.
- ZAMPOLLI A. Filologia e informatica: le origini della filologia computazionale [J/OL]. Euphrosyne, 2004, 32: 11-24.
- 董丹. 意大利国家语言能力研究[M]. 北京: 外语教学与研究出版社, 2022.
- 余丹妮. 汉意意汉文学平行语料库的研制[J]. 语料库语言学, 2020 (2): 83-88.
- 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(喻儒辰)  
100089 北京市 北京外国语大学欧洲语言文化学院(董丹、郭壹一)

# MineDEAP 矿业工程学术英语语料库的创建

中国矿业大学（北京） 张汝莹

**提要：**MineDEAP 矿业工程学术英语语料库是北京外国语大学中国外语与教育研究中心联合国内多所高校创建的“DEAP 学术英语语料库”的子语料库。本文主要介绍矿业工程学术英语语料库的建库过程、语料库构成及应用前景。该语料库为矿业工程学科的专业英语教学及研究提供了大规模真实语料，在专业英语词典编纂、跨学科学术英语语体研究等领域具有借鉴意义。

**关键词：**MineDEAP、矿业工程、学术英语、语料库

## 1 引言

学术英语（English for academic purposes, 简称EAP）作为专门用途英语的分支（文秋芳 2013），其内容主要涉及某学科跨国界的普遍知识（文秋芳 2014），具有区别于通用英语的语体特征（姜峰 2020）。语料库不仅为EAP量化实证研究提供了大规模真实语料（王立非 2019），还可直接融入教学实践中（徐秀玲、许家金 2017），在EAP教学中为教师提供教学材料。现有的EAP语料库涵盖不同语体，如密歇根大学学术口语语料库（MICASE）、密歇根高阶学生论文语料库（MICUSP），可分为本族语者语料库，如牛津学术英语语料库（OCAE），以及学习者语料库，如英国学术英语写作语料库（BAWE）。既有囊括多个学科的通用语料库，如上海交通大学科技英语语料库（JDEST），也有针对某一具体学科的专用语料库，如华中农业大学的农业科学学术英语论文语料库（AEC）。多数为单语语料库，也有少量平行语料库，如中国法律法规汉英平行语料库（PCCLD）。

在矿业工程领域，已建成的专用语料库主要是研究者根据自身研究目的需要，自建的小型语料库（如孟莹莹 2016；李雅玲、李绚丽 2016）。该类语料库库容较小，且语料遴选标准及子学科划分标准不一，语料来源时间跨度较小。因此，本文作为中国外语与教育研究中心学术英语语料库建设项目（Database of English for Academic Purposes, 简称DEAP）的子课题，旨在建设矿业工程专业学术英语语料库（MineDEAP），为新工科建设背景下矿业工程专业的EAP研究及教学实践提供借鉴。以下主要介绍该语料库的建库目标、建库过程及应用前景。

## 2 建库目标

本文在DEAP学术英语语料库的总体建库方案指导下,结合矿业工程专业的具体学科特点,旨在创建覆盖该专业所有二级学科、囊括主要语类、涵盖该专业内广泛认可的高水平SCI英语期刊、库容为500万词的学术英语全文语料库,从而助力矿业工程专业的学术英语教学、教材及专业英语词典编纂、语言本体研究、跨学科对比、学术话语体系建构及国际学术交流。

## 3 建库过程

### 3.1 学科领域及来源期刊

根据教育部发布的2018年版《学位授予和人才培养学科目录》,矿业工程一级学科(0819)下设3个二级学科:采矿工程(081901)、矿物加工工程(081902)、安全技术及工程(081903)。本文依据Web of Science 2020年版SCI期刊引文报告(Journal Citation Report,简称JCR分区)及中国科学院文献情报中心期刊分区表(2020年版),分别选取上述3个二级学科中期刊类别位于Q1区,且影响因子(influence factor)、h因子(h-index)综合排名<sup>1</sup>前三的期刊作为备选期刊。同时,通过咨询这3个二级学科的多位学者教师,具体了解备选期刊的刊文偏好、业内知名度及同行认可度等相关指标,经相关专业学者推荐,选取7本期刊作为语料来源。收集每本期刊在2016—2022年发表的文献,包含研究性论文、综述性论文、通讯文章3种主要语类。同时,为确保所选语料具有学科代表性,能够反映各二级学科的核心研究成果,且语言质量较高,符合专业英语术语传统及写作规范,本文按照文献的被引次数由高到低进行语料采集,将可能存在明显学科交叉的文献交由相关领域学者进行复核,剔除研究对象及研究方法明显偏向其他学科的文献。最终建成的语料库共收集829个文本,库容为5,169,118形符(token),70,527类符(type),类符/形符比(type/token ratio)为1.44%,语料平均篇幅为6,235形符。语料库具体构成见表1。

### 3.2 语料收集及整理

本文在语料收集集中,主要下载文献的PDF格式或HTML格式文档,前者先通过Adobe Acrobat Pro将格式转换为Word文档,而后经过语料清理,保存为TXT文本格式。语料中包含论文的标题、作者姓名、摘要、正文共四部分内容,删除了作者机构信息、基金资助情况、脚注、尾注、参考文献及附录。文档以“一级学科\_二级学科\_期刊代码\_文献类型\_文本编号”方式命名。为了便于区分,矿业工程(Mineral Engineering)一级学科以其英文名称的第一个单词首字母M

表 1 MineDEAP 学术英语语料库构成

序号	二级学科	来源期刊 (出版商)	期刊代码	影响因子 <sup>2</sup>	文献类型	发表年份	文本数	库容 (字符)
1	采矿工程 (Mining Engineering)	International Journal of Rock Mechanics and Mining Sciences (Pergamon-Elsevier)	RMM	7.4	RA	2017-2022	185	1,179,174
		Ore Geology Reviews (Elsevier)	OGR	3.8	RV	2017-2022	50	457,566
		Minerals Engineering (Pergamon-Elsevier)	ME	5	RA	2019-2020	106	636,731
					RV		6	
					C		4	
2	矿物加工工程 (Mineral Processing Engineering)	Mineral Processing and Extractive Metallurgy Review (Taylor & Francis)	EMR	4.8	RV	2018-2020	71	431,449
		International Journal of Coal Preparation and Utilization (Taylor & Francis)	CPU	2.1	RA	2016	135	522,031
					RV	-2020	1	
3	安全技术及工程 (Safety Technology and Engineering)	Applied Energy (Elsevier)	AE	11	RA	2016	103	1,046,252
					RV	-2021	24	
		Energy (Pergamon-Elsevier)	E	8.2	RA	2019	130	895,915
					RV	-2022	14	

作为代码，二级学科则以其英文名称中实词的首字母命名：采矿工程（Mining Engineering）记为“ME”、矿物加工工程（Mineral Processing Engineering）记为“MPE”、安全技术及工程（Safety Technology and Engineering）记为“STE”；文献类型中研究性论文（Research article）记为“RA”、综述性论文（Review）记为“RV”、通讯文章（Communication）记为“C”；期刊代码采用期刊英文名称的代表性单词首字母，如Ore Geology Reviews的期刊代码记为“OGR”；文本编号依据文献下载的顺序进行统一编号。比如，语料文件“M\_ME\_RMM\_RA\_01”表示

采矿工程专业中来源于International Journal of Rock Mechanics and Mining Sciences期刊的第一篇研究性论文。

语料的文本整理遵循最大程度保留文献原始结构的原则,对照下载的论文原件对每篇语料进行人工文本清理,通过批量清理与手动修正相结合的方式,主要清理断头句、乱码、全角/半角、多余连字符及空格等格式问题。同时删除正文中的表格和图片,仅保留其标题;将正文中的数学公式以#E进行替换;手动删除文内注(in-text citation),仅保留作为正文句子成分的文内注,如“The same case study from Ref. [40] was chosen in Ref. [41]...”(M\_STE\_E\_RA\_14)。因文本清理工作量较大,为避免文本清理中可能存在的遗漏、错误或清理标准不一致的情况,项目组在前期培训的基础上,首先对10篇文献进行了试处理,对发现的问题进行及时反馈纠正,而后以50篇语料为单位,在语料清理完成后进行组内互查,查漏补缺,确保语料的整理质量。整理后的语料保存为UTF-8编码格式,方便后期使用语料库软件进行分析。

### 3.3 元信息标注

本文在每篇语料的收集过程中,将语料的元信息汇总并保存为Excel格式,涵盖语料的所属二级学科、出版商、期刊名称、论文标题、体裁、DOI号、发表时间及卷号/期号。需要说明的是,因论文普遍具有线上发表时间及出版时间两个发表时间,本文以线上发表时间为准进行记录。

## 4 应用前景展望

本研究建设的MineDEAP矿业工程学术英语语料库将作为中国外语与教育研究中心学术英语语料库(DEAP)的子语料库,已在北京外国语大学BFSU CQPweb平台上发布,实现语料库资源共享。该语料库在投入使用后,凭借其在“用、量、聚、器”(许家金 2017: 52-53)上的独特优势,可在相关语言学研究、教学研究及实践中发挥积极作用。

在语言本体研究方面,MineDEAP语料库可为矿业工程一级学科以及三个二级学科在词、句、篇、语域、语体等层面的学术英语语言特征研究提供量化支撑。同时,结合DEAP语料库的其他子库,依托其1亿词的超大库容及科学的学科划分标准等优势,进一步探究学术英语的共性特征及跨学科差异,助力语言变体的相关研究。此外,该语料库可以作为学习者学术英语语料库的参照库,为二语学习者的学术英语中介语对比研究提供借鉴。

在教学方面,该语料库可为矿业工程专业及其二级学科的学术英语教学提供大规模真实语料。教师可结合词频、搭配强度、主题词表等语料库软件统计结果,



择选具有代表性的索引行或论文段落作为教学材料，使教学内容重点更突出，实践性更强。同时，通过诸如数据驱动学习等教学法，将语料库技术融入教学设计中，促进启发式教学以及学生的自适应学习。另外，该语料库可广泛应用于专业英语词表的创建、专业英语教材及词典的编纂、语言测试等领域，为数字化环境下的语言智能教学提供参考。

### 注释

- 1 因评价期刊质量的不同因子计算方法不同，所以 Web of Science 的期刊 JCR 分区与中科院期刊分区对同一期刊的分区不尽相同。且在不同因子排序中，同一期刊的排名也会有所不同。为体现期刊在专业内的综合评价结果，本文将 JCR 与中科院分区结果，以及多种因子排名进行综合考量，作为期刊选择依据。
- 2 Web of Science 中期刊的 5 年平均影响因子（检索日期为 2023 年 12 月 28 日）。

### 参考文献

- 姜峰. 基于多维分析的学术语篇语体特征的历时考察[J]. 外语教学与研究, 2020 (5): 663-673.
- 李雅玲, 李绚丽. 基于语料库的矿业科学期刊论文中的报道动词使用特征分析[J]. 技术与创新管理, 2016 (3): 337-339.
- 孟莹莹. 国际矿业期刊中外作者英文学术论文的转述语对比研究[D]. 徐州 中国矿业大学, 2016.
- 王立非. 王立非谈语料库与 ESP 研究[J]. 语料库语言学, 2019 (2): 1-10.
- 文秋芳. 输出驱动假设在大学英语教学中的应用: 思考与建议[J]. 外语界, 2013 (6): 14-22.
- 文秋芳. 大学英语教学中通用英语与专用英语之争: 问题与对策[J]. 外语与外语教学, 2014 (1): 1-8.
- 徐秀玲, 许家金. 我国外语教学中的语料库应用 40 年[J]. 中国外语教育, 2017 (4): 62-68.
- 许家金. 语料库研究学术源流考[J]. 外语教学与研究, 2017 (1): 51-63.

**通信地址:** 100083 北京市 中国矿业大学（北京）文法学院外语系

# SET多版本高中英语教材语篇语料库的创建<sup>\*</sup>

广州协和学校 陈运良

**提要：**SET多版本高中英语教材语篇语料库选取人教版、外研版、北师大版三种新版教材的语篇为语料，由使用该版本的国内五地教师分工联合建成。该库依托2017年版《普通高中英语课程标准》的设计以及教学实际需要进行了语篇来源、主题、类型及易读度4个方面的文本头部元信息标记。据此标记，本文介绍了该库的总体信息、语篇主题、语篇类型以及语篇易读度的分布情况。作为以教学使用为目的的基础教育阶段语料库，该库在提升教学效果、丰富教师理念、促进学生发展方面具有应用前景和价值。

**关键词：**高中英语、新课标、多版本、教材语篇、语料库

## 1 引言

Meunier & Gouverneur (2009: 179-180) 认为，有必要将学习者所使用的教材纳入语料库大家庭之内，将新型教材语料库的理念引进语言教学当中，以迎接未来语言教学的挑战。罗庆铭 (2017: 59-62) 指出，教材语料库能够变隐性信息为显性信息、变分散信息为整体信息，同时其语料加工和标注更符合教学需要，这两方面都将有益于教学。因此，建立教材语料库是必要的、有益的。

观察旧版高中英语教材语料库的使用 (罗颖 1999; 谢家成 2010; 陈妍玲、陈杰鑫 2014; 陈丽勤 2016; 何安平等 2017; 梁红梅 2018)，笔者发现仍存在两个问题：一是只能在大学局域网线上检索，中学一线教师难以实时使用；二是建库碎片化，所建微型语料库只能满足一时或一课的需求。

《普通高中英语课程标准》(教育部 2017，以下简称“新课标”) 的颁布促成了新教材的编写发行。新版高中英语教材从2019年或2020年启用至今，时间不长但问题已然显现。针对教材使用中忽视语篇或是囿于特定语篇的问题，梅德明 (2021) 指出，教师不应受限于单一教材，要以“用教材教”替代“教教材”，从而培养核心素养。刘道义 (2021) 指出，教师要提升有效整合教学资源的能力，

<sup>\*</sup> 本文是2021年度中国外语教材研究专项课题“基于语料库的多版本高中英语新教材使用互鉴研究”(ZGWYJCYJ2021ZZ06) 的阶段性成果，得到北京外国语大学中国外语教材研究中心资助。

而如果有的语篇偏难或偏易，可灵活整合，从而满足学生的需求。可见教师需要根据学生发展的需要灵活使用教材，而跨版本使用教材汲取众长也可以是“用教材教”的一个做法。陶伟、古海波（2020）基于外语类CSSCI期刊载文，对2010—2019年我国外语教材研究进行了文献综述，未见跨版本教材使用研究。

由此可见，建立新版多版本高中教材语篇语料库既可充实现有语料库，发挥建库尤其是数据收集对语言学的研究价值（McEney & Hardie 2012: 6），也将方便跨版本借鉴，从而有利于解决当前新教材使用中出现的問題，体现建库填补空白为时所用的价值（Reppen 2022: 13）。

## 2 建库目标与方案

### 2.1 建库目标

考虑到新教材的编写在新课标框架下进行，SET（Senior High English Texts的简称，同时意指多版本组成“一套”）也将遵循新课标框架进行创建。SET的建设以教学使用为基本诉求，具体包括如下方面。

其一，建成体现新课标理念的人教版、外研版、北师大版三个主流版本的高中英语教材语篇语料库。新课标强调主题与语篇，这也是采集教材语料的基本依据。以此路径编制的语料库便于教师在不同版本语篇中切换，能更好地照应梅德明、刘道义所提出的灵活使用教材的问题，加深教师对新课标的理解。

其二，丰富和提升语料库教学理念。语料库不仅是素材库和手段，同时也是具有自身理念的一门学科。语料库注重“观察”（杨惠中 2017），注重观察的教学活动将引发自主学习、合作学习、探究学习等学习方式，有利于践行新课标所倡导的英语学习活动观。同时语料库核心理念中的语篇类型、语境、数据驱动、频数、关键词、搭配、类联接、同现、构式、型式和语义韵等（Sinclair 1991; Biber *et al.* 1998; Thornbury 2004; Ellis *et al.* 2014; Hunston & Su 2017; Crosthwaite 2020; O’Keeffe & McCarthy 2022）都将丰富高中教师的教学理念，并影响教学活动。

其三，促成国内多校联动开展跨地区跨版本融合。目前不同地区使用不同教材，相对独立，缺乏沟通和融合。针对这个现象，SET选取最常用的版本进行跨版本建库，而建库成员来自国内不同地区、分别使用这些教材。其优势在于成员在本地教材使用中能精准领会该教材的特点，接触使用中的问题，提出针对性建库建议及教学建议，继而汇总来自3个版本的不同声音，达到教学实施的有效整合，达成新课标理念的有效贯彻。

## 2.2 建库方案

### 2.2.1 语料来源

从出版社或省级发行部购买已出版使用的普通高中英语教科书人教版(2019)、外研版(2020)及北师大版(2019)共21册,各套分别包括必修部分第一至三册、选择性必修部分第一至四册,同时购买配套的教师用书辅助教材解读。

获取以上教材及教师用书的pdf电子书,利用WPS转换成Word文档。从Word文档中选取语篇,参照所购纸质教材及教师用书进行文字、版式校对。

### 2.2.2 人员分工及程序

桂诗春、梁茂成提倡语料库资源的共建与共享(桂诗春等2010)。徐秀玲、许家金(2017)认为“统一标准,多方合作,共建共享,建研并重”是解决专门用途语料库所面临问题的有效途径。基于以上指引,笔者依托主持的2021年度“中国外语教材研究专项课题”,营造共建共享生态。课题组8名成员分别来自广州、杭州、北京、成都、西安,为建库核心成员,并形成各自团队,最终参建人员达25人。

首先五地教师在统一学习建库体例后开展分工联建,于2021年12月建成初版SET;其次课题组分版本负责人合作对初版进行全面检校和补充;最后总负责人对修订版进行第三遍检查,特别做好元信息标记的标准统一,于2022年5月建成定版SET。

## 3 文本的采集、清理与命名

Reppen(2010:32)认为建库前需认真考虑如下4个问题:一是文本由什么构成,二是如何命名文件,三是各个文件包含什么信息,四是以什么格式保存文本。其基本观点是这些问题了解得越清楚,文本的采集就越精准,文本的命名就越便于解读。

### 3.1 文本采集与清理

建库团队采集教材所有书面语篇。除常规语篇外,还包括名言、谚语、调查问卷、微技能指引,还有图表、流程图等非连续性文本。如果语篇出自填空、改错、改写、排序等练习,参照教师用书答案相应调整。外研版选择性必修第一至四册附录部分含补充读物,采集时纳入对应单元;北师大版全套都设置了文学角,但不对应具体单元,仍一并采集,与各单元并列。与其他两个版本不同,北师大版教材书末专设录音脚本栏,也一并采集,归入对应单元末尾。

统一用Microsoft Word进行编辑,使用半角英文输入法,设置语法错误显

示,但不设置自动更正,以免将英式拼写自动改为美式拼写。原则上删除文本中用以简化阅读难度的中文注释,个别保留了用以说明中华文化的汉字例字或注释,以保证文意完整。将文本中仅有的3处“<>”符号改为“()”,如将“Wang Ying <wangying@\*\*\*.com>”(出自北师大版必修一第一单元)改为“Wang Ying (wangying@\*\*\*.com)”,这客观上有损电邮的格式规范,但可以避免在隐藏头部元信息时被一并隐去,造成信息缺漏。此外,文本中的破折号统一用“--”标示,以尽可能被各检索软件读取。

完成word层面的校对工作后,用AntFileConverter(Anthony 2015)将文本从word文档转换为txt文档,达到较好的降噪效果。例如转换后的非英文字母符号“ə”(出自人教版选择性必修二第二单元)、“π”(出自北师大版必修三第九单元)、“鱼”(出自北师大版必修一第三单元)就能够被AntConc(Anthony 2019)等检索软件读取。

### 3.2 文本命名

设置三个层级进行文件夹及文件命名。第一层级为总库,名为SET。第二层级为分库,名字是总库名加三个版本的拼音缩写,分别为SET-RJB(人教版)、SET-WYB(外研版)、SET-BSD(北师大版)。第三层级为分单元文件,如SET-WYBBIU1表示外研版必修一第一单元语篇,又如SET-BSDB7A表示北师大版选择性必修四附录文学角的语篇。

## 4 文本标记

### 4.1 标记与标注的取舍

李文中(2012)准确厘清了标记与标注的区别。本库进行文本标记,但不作文本标注,具体有如下考虑。

Burnard(2005: 35)强调,元信息标记对语料库语言学极其重要。Kirk & Andersen(2016)也认为,标记对分析者了解语言学之外的信息尤为关键。元信息标记多寡不定,以语篇来源、主题、类型为多见(Burnard 2005; Reppen 2010; 崔刚、盛永梅 2000; 陈小荷 2021)。事实上,头部元信息标记是本库体现新课标理念的最佳落脚点。本库将新课标的“主题语境内容要求表”(教育部 2017: 14-15)以及“语篇类型内容要求表”(教育部 2017: 17-18)作为元信息标记的核心依据。为便于跨教材互鉴时区分语篇难度,本库设置易读度标记,采用在线工具计算易读值。因此,本库的元信息标记包括语篇来源、语篇主题、语篇类型、语篇易读度4个方面。



文本标注可为语料库增加挖掘价值。兰卡斯特大学各代CLAWS软件作为文本标注的重要工具广为使用 (Leech 1993; 李文中 2012), 然而随着第四代语料库检索工具的陆续出现, 建库时专门的词性、用法赋码就略显多余, 如兰卡斯特大学#LancsBox 6.0 (Brezina *et al.* 2021) 对所导入的语料会自动添加词性 (POS)、词元 (lemma) 等信息, 让检索一步到位。基于此项考虑, 本库不进行文本标注。

## 4.2 头部元信息

如前所述, 本库头部元信息包括四个方面, 其中语篇来源、主题、类型三项在文本采集时同时标记, 语篇易读度在校对完毕后标记。4个方面信息写入角括号“<>”内, 参考先例为中国学习者英语语料库CLEC (桂诗春、杨惠中 2003)。标记样式为“<RJBB1U5T5> <I-5> <E> <R=79.5>”, 意为该语篇选自人教版必修一第五单元, 是第五个语篇, 主题为第一大主题语境“人与自我”之“语言学习的规律、方法等”子主题, 类型为说明文, 易读值为79.5 (值越大越容易读)。

新课标的“主题语境内容要求表”(教育部 2017: 14-15) 第一层级为主题语境, 包括“人与自我”“人与社会”“人与自然”三大主题语境, 本库分别标示为“I”“II”“III”; 第二层级为主题群, 如“人与自我”包括“生活与学习”“做人与做事”两个主题群, “人与社会”包括“社会服务与人际沟通”等4个主题群, “人与自然”包括“自然生态”等4个主题群, 本库不作标示; 第三层级为主题语境内容要求, 或称子主题, 如“人与自我”包括“个人、家庭、社区及学校生活”等9个子主题, 其中第1至5个对应“生活与学习”主题群、第6至9个对应“做人与做事”主题群, 本库分别用阿拉伯数字“1”至“9”进行标示, “人与社会”“人与自然”分别包括16个和7个子主题, 标示方法相同。事实上, 对第三层级的标示足以显示第二层级信息, 所以“第一层级”加“第三层级”的标记样式 (例如“I-5”) 可以保证语篇主题信息被完整标记, 具有自足性。

依据新课标的“语篇类型内容要求表”(教育部 2017: 17-18), 制定8种语篇类型, 具体如表1所示。

需要说明的是, 本库将对话、访谈、讲座、报告、演讲、讨论、辩论等语篇归入“口头类”语篇类型, 对原文内容作了整合。另外, 新课标将“新媒体语篇”单立为一种语篇类型, 包括一般网络信息、电子邮件、手机短信、博客、知识类科普类网页等。在具体标记过程中, 有时就出现语篇类型相交叠的情况, 如教材显示为博客, 同时又是一篇说明文。对此, 本库的处理办法是优先标示为“新媒体语篇”, 以体现对新课标视角的照应, 也便于该项语篇的教学、研究提取。

语篇易读度用“R”表示, 全拼为readability。本库用在线计算器Flesch Kincaid Calculator (<https://goodcalculators.com/flesch-kincaid-calculator/>) 计算易读值。将文本粘贴至该网页对应窗口, 即刻生成结果。下面仍以人教版必修一第五



单元第五个语篇为例，其结果如图 1 所示。

表 1 SET 语料库文本头部语篇类型信息标记对照

语篇类型	标记代码	全拼
记叙文	NA	narration
说明文	E	exposition
议论文	A	argumentation
应用文	P	practical discourse
口头类	O	oral discourse
新闻报道	NR	news report
新媒体语篇	NM	new media discourse
其他语篇类型	M	miscellaneous

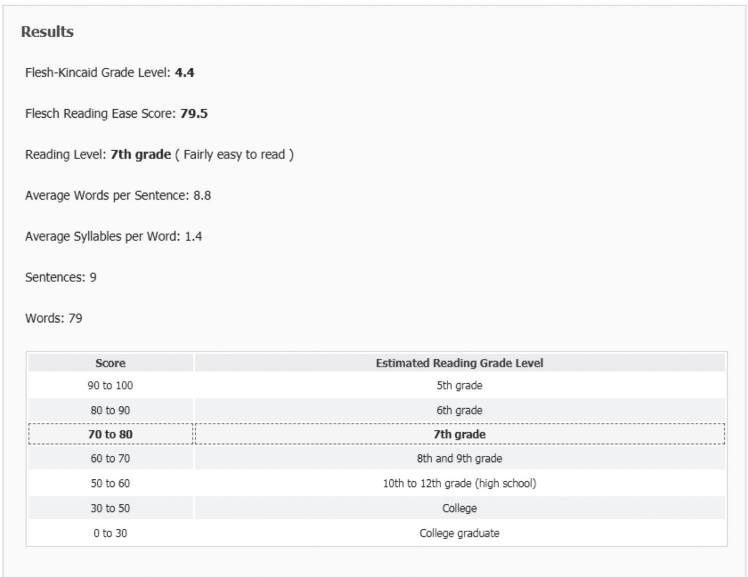


图 1 人教版必修一第五单元第五个语篇在线生成易读值结果

图 1 第二行显示，该语篇易读值是 79.5，适合母语七年级读者阅读，级别为“比较容易阅读”。图 1 下部为分值与母语读者阅读水平预计年级的对应表，共分 7 个级别，该语篇处于第 3 级别。易读度作为元信息的第四项内容，本库将之标示为“R=79.5”。

5 SET基本数据

5.1 SET总体信息

本库共收入目前已发行的三个版本新教材共21册，具体统计如表2所示。

表2 SET语料库总体信息

	册	单元	语篇	类符	形符	形符/语篇	类符/形符
人教版	7	36	693	8,699	116,001	167.39	0.075
外研版	7	42	547	8,762	97,320	177.92	0.090
北师大版	7	28	609	8,055	122,778	201.61	0.066
总计	21	106	1,849	25,516	336,099	181.77	0.076

表2显示，三个版本单元分配各有特点，人教版为每册5个单元另加入学预备单元共36个单元，外研版为每册6个单元共42个单元，北师大版为每册3个单元另加文学角共28个单元（本库视角）。各版本语篇数分别为人教版693篇、外研版547篇、北师大版609篇。各版本纯文本类符数为8,055至8,762不等。各版本纯文本形符数为97,320至122,778不等，合计336,099，这也是SET目前的库容。形符语篇比为167.39至201.61不等，新教材收入大量各类型短小语篇是造成语篇平均长度不高的主要原因。类符形符比（TTR）可作为语篇词汇变化度的部分参考，从这个角度看外研版的语篇词汇变化较大（比值为0.090），北师大版的语篇词汇变化较小（比值为0.066），本库收入北师大版的录音脚本，增大了其口语类语篇的占比，可能导致语篇词汇变化变小。

依据头部元信息标记，进一步获取如下语篇主题、类型、难度分布情况。

5.2 SET语篇主题分布

本库对新课标三大主题语境的反映情况如图2所示。

图2显示，人教版、外研版、北师大版各版本中，归入“人与社会”的语篇最多，依次为347篇（占50.07%）、232篇（占42.41%）、297篇（占48.77%）；归入“人与自我”的语篇次之，依次为221篇（占31.89%）、192篇（占35.10%）、226篇（占37.11%）；归入“人与自然”的语篇最少，依次为125篇（占18.04%）、123篇（占22.49%）、86篇（占14.12%）。事实上，新课标主题语境内容要求项目由多到少也以这个顺序分布，分别为“人与社会”16个、“人与自我”9个、“人与自然”7个。由此可见，SET各版本教材语篇在三大主题语境中的分布吻合新课

标设计。

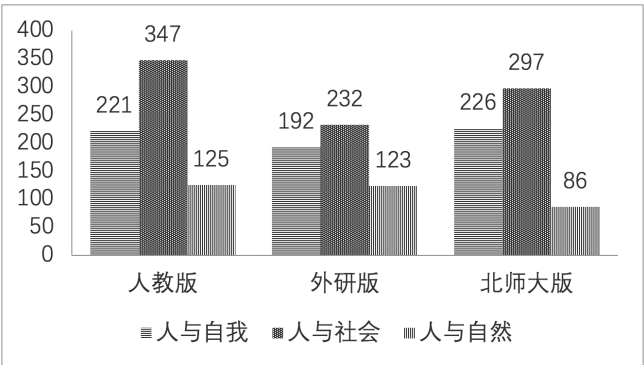


图2 SET 归入新课标三大主题语境的篇数统计

5.3 SET语篇类型分布

本库对新课标8个语篇类型的反映情况如图3所示。

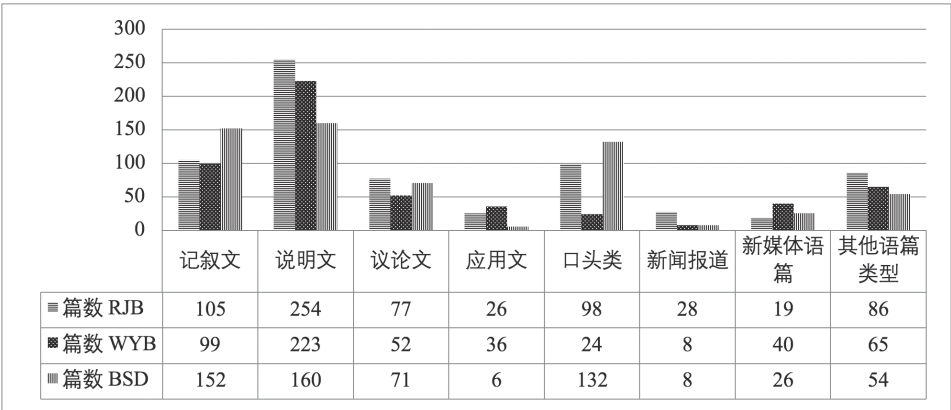


图3 SET 归入新课标八种语篇类型的篇数统计

图3显示三个版本有两个共同之处，一是所有8种语篇类型均有分布，二是语篇数最多的均为说明文和记叙文，其中人教版合计359篇（占51.80%），外研版合计322篇（占58.87%），北师大版合计312篇（占51.23%），占据所有语篇的一半以上。而各个版本横比又各有特别之处，人教版以说明文、新闻报道、其他语篇类型比较突出，外研版以新媒体语篇比较突出，北师大版则以记叙文、口头类比较突出（口头类的突出表现与该教材设立录音脚本栏有关）。

## 5.4 SET语篇易读度分布

以10分为一个区间获取三个版本语篇的易读度分布情况，结果如图4所示。

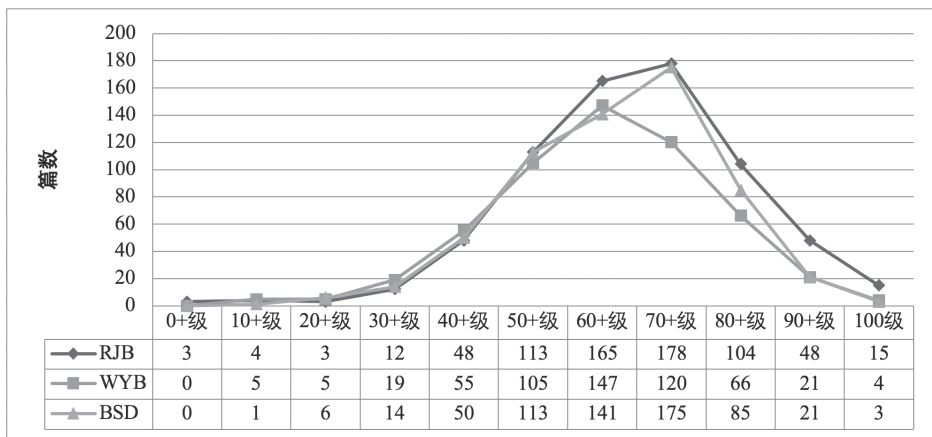


图4 SET易读度分级篇数统计

从图4可以看出，就易读度而言，人教版、北师大版的峰值在“70+级”，分别有178个语篇和175个语篇，对应图1所示母语七年级水平；外研版的峰值在“60+级”，有147个语篇，对应图1所示母语八、九年级水平。各版本的特点是，人教版难度分布最为全面，存在“0+级”语篇，同时最易读的“100级”语篇最多，共15篇；外研版曲线横比整体偏左，显示其语篇的整体难度较大；北师大版进入峰值后出现陡降，其“80+级”至“100级”语篇数量相对较少。

## 6 应用前景

笔者曾于2012年主持建成旧版多版本高中英语教材语篇语料库，用于辅助高中英语有效教学，其实际应用见于陈运良（2014），何安平（2017），徐秀玲、许家金（2017），何安平（2020）。新库SET参照新课标丰富了头部元信息标记，为语料库增值赋能，能更好地服务教学活动。SET具体有如下应用前景。

首先，SET将为“教学加工”提供支撑。本库的语料与日常教学密切相关，本库能提供具有互文性的文本网络，适合针对性地选取内容辅助教学。具体包括：

（1）通过跨版本借鉴，丰富课内教学素材，优化教学活动，做到“用教材教”。涉及的教学领域包括语音、词汇、语法、阅读、写作等方面。

（2）综合各版本语篇，形成第二课堂或课后教学素材，开展分级阅读、群文阅读、主题式拓展阅读以及读后续写等教学。

(3) 结合各版本的图片、音频、视频等素材, 建构多模态教学模式。

(4) 强化跨地区跨版本教学沟通与互鉴, 提升所在地区教材版本的使用效果。

其次, SET将促进教师发展。教师在调用语料的同时会将语料库技术与理念融入教学设计和课堂实施中。何安平指出, 未来教师将从“语料库产品(如教材)”的“消费者”转变为“合作研发者”“实践反思者”和“教师自我发展的行动者”(桂诗春等 2010)。

最后, SET将促进学生成长。针对如何加强语料库语言学在外语教学中应用的问题, 杨惠中指出, 语料库提供“数据驱动学习”, 教师可以开发针对性微库, 让学生直接接触复杂的语言现象, 利用检索软件单独或合作探索英语的用法, 展开观察、分析、对比、综合、归纳活动, 从而掌握地道的英语、发展英语语感(桂诗春等 2010)。SET将在以“数据驱动学习”为模式的科技化智能化校园中成为促进学生成长的有效力量。

## 7 结语

在基于新课标的新教材推出之际建立其语料库可谓适逢其时。以新课标为依托, 参照其主题语境、语篇类型框架对教材文本进行元信息标记, 这是SET创建的一个特色之处。“跨版本”实现“一标多本”语境下的扩容与互补, “跨地区”提高建库效率、拓宽分享借鉴平台, “跨版本”与“跨地区”相结合的做法是SET创建的另一个特色之处。“跨”的目的在于“融”, 将别地经验融会贯通、将别版本的语篇为我所用、将语料库理念融入教学活动, 将能收到更好的教学效果。“融”的发展方向是“创”, 交叉和融合激发灵感带来创新, 这既将体现在教的方面, 也将体现在学的方面, 将能更好地发展学生的学科核心素养。

当然, SET的库容相对各类大型语料库仍然很小, 仍然有扩大空间。不同版本语料内容仍有不完全匹配的地方, 比如北师大版收入了录音脚本, 其他两个版本因为教材上没有录音脚本就缺失这一部分语料。此外, 文本标记的细项归类也可能存在主观判断的偏误。这些方面有待进一步的完善和提升。

### 参考文献

- ANTHONY L. AntFileConverter 120 [CP/OL]. 2015. <http://www.laurenceanthony.net/software.html>.
- ANTHONY L. AntConc 3.5.8 [CP/OL]. 2019. <http://www.laurenceanthony.net/software.html>.
- BIBER D, CONRAD S, REPPEN R. Corpus linguistics: investigating language structure and use [M]. New York: Cambridge University Press, 1998.

- BREZINA V, WEILL-TESSIER P, MCENERY A. #LancsBox v. 6.x. [CP/OL]. 2021.  
http://corpora.lancs.ac.uk/lancsbox.
- BURNARD L. Metadata for corpus work [C]//WYNNE M. Developing linguistic corpora: a guide to good practice. Oxford: Oxbow Books for the Arts and Humanities Data Service, 2005: 30-46.
- CROSTHWAITE P. Data-driven learning for the next generation [M]. New York: Routledge, 2020.
- ELLIS N, O'DONNELL M, ROMER U. The Processing of verb-argument constructions is sensitive to form, function, frequency, contingency and prototypicality [J]. Cognitive Linguistics, 2014, 25(1): 55-98.
- HUNSTON S, SU H. Patterns, constructions, and local grammar: a case study of “evaluation” [J]. Applied Linguistics, 2017, 40(4): 567-593.
- KIRK J, ANDERSEN G. Compilation, transcription, markup and annotation of spoken corpora [J]. International Journal of Corpus Linguistics, 2016, 21(3): 291-298.
- LEECH G. Corpus annotation schemes [J]. Literary and Linguistic Computing, 1993, 8(4): 275-281.
- MCENERY T, HARDIE A. Corpus linguistics: method, theory and practice [M]. Cambridge: Cambridge University Press, 2012.
- MEUNIER F, GOUVERNEUR C. New types of corpora for new educational challenges: collecting, annotating and exploiting a corpus of textbook material [C]//AIJMER K. Corpora and language teaching. Amsterdam: John Benjamins, 2009: 179-201.
- O'KEEFFE A, MCCARTHY M. The Routledge handbook of corpus linguistics [C]. Oxon: Routledge, 2022.
- REPPEN R. Building a corpus: what are key considerations? [C]//O'KEEFFE A, MCCARTHY M. The Routledge handbook of corpus linguistics. Oxon: Routledge, 2010: 31-37.
- REPPEN R. Building a corpus: what are key considerations? [C]//O'KEEFFE A, MCCARTHY M. The Routledge handbook of corpus linguistics. Oxon: Routledge, 2022: 13-20.
- SINCLAIR J. Corpus, concordance, collocation [M]. Oxford: Oxford University Press, 1991.
- THORNBURY S. Natural grammar [M]. Oxford: Oxford University Press, 2004.
- 陈丽勤. 广西基础教育英语教学语料库建设研究[J]. 基础外语教育, 2016 ( 1 ): 18-23.
- 陈小荷. 留学生汉语语料库杂谈[J]. 语料库语言学, 2021 ( 1 ): 1-4.
- 陈妍玲, 陈杰鑫. 基于语料库的积木式英语课堂教学效果研究[J]. 教学研究, 2014 ( 5 ): 48-54.



- 陈运良. 广东高考英语语篇中的高阶词及其在教材语篇中的反映[J]. 中小学外语教学, 2014 (2): 5-9.
- 崔刚, 盛永梅. 语料库中语料的标注[J]. 清华大学学报(哲学社会科学版), 2000 (1): 89-94.
- 桂诗春, 冯志伟, 杨惠中, 等. 语料库语言学与中国外语教学[J]. 现代外语, 2010 (4): 419-426.
- 桂诗春, 杨惠中. 中国学习者英语语料库[M]. 上海: 上海外语教育出版社, 2003.
- 何安平, 梁红梅, 唐洁仪. 语料库辅助英语教学入门[M]. 北京: 外语教学与研究出版社, 2017.
- 何安平, 许家金, 张春青. 语料库辅助中学英语教学案例选编[M]. 北京: 外语教学与研究出版社, 2020.
- 教育部. 普通高中英语课程标准[M]. 北京: 人民教育出版社, 2017.
- 李文中. 语料库标记与标注: 以中国英语语料库为例[J]. 外语教学与研究, 2012 (3): 336-345.
- 梁红梅. 共选理论视角下中学英语教材中短语教学的设计特征[J]. 基础外语教育, 2018 (6): 27-36.
- 刘道义. 高中英语新课标实施中的问题和挑战[J]. 英语学习, 2021 (1): 15-22.
- 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.
- 罗颖. 利用语料库分析中学英语课堂提问技巧[J]. 国外外语教学, 1999 (4): 26-31.
- 梅德明. 高中英语新课标实施过程中的问题: 缕析与建议[J]. 英语学习, 2021 (1): 4-14.
- 陶伟, 古海波. 我国外语教材研究综述(2010—2019)[J]. 浙江外国语学院学报, 2020 (3): 77-82.
- 谢家成. 中学英语教材词汇语料库分析[J]. 外语教学理论与实践, 2010 (1): 55-61.
- 徐秀玲, 许家金. 我国外语教学中的语料库应用40年[J]. 中国外语教育, 2017 (4): 62-68.
- 杨惠中. 叩鸣录: 杨惠中先生答客问[J]. 语料库语言学, 2017 (2): 1-22.

通信地址: 510160 广东省广州市 广州协和学校

# 《如何利用语料库进行语言教学》述评

浙江外国语学院 张春青

Averil Coxhead. 2022. *Connecting Corpora and Language Teaching*. Beijing: Foreign Language Teaching and Research Press. 264pp.

## 1 引言

随着语料库语言学和语料库检索工具的发展，特别是线上线下便捷检索工具的增多，语料库辅助语言教学受到了越来越多的关注。语料库为英语学习提供真实、丰富的素材，同时，学生能以研究者身份（Johns 1991/1994）自主探究、发现和掌握语言规律。Boulton & Cobb（2017）的元分析发现，语料库辅助教学能显著提升学生学习成效。语料库辅助语言教学已经成为主流语言教学技能（Ma *et al.* 2021）。但是，语料库在课堂教学中的直接应用仍然非常有限。其困难在于，在二/外语教学中，缺乏适合学习者语言水平的语料库，直接使用通用语料库素材会产生不可理解输入，挫伤学习动机。另外，语料库辅助英语教学课程在师范生中开设较少，教师普遍缺乏职前和在职期间的语料库辅助教学培训。以上两点都涉及了教师的语料库素养问题（Callies 2019）。语料库素养既包括语料库知识，例如知晓语料库方法的优势和劣势，会作索引分析等，也包括语料库语言学知识和技能、语言知识和技能以及教学法三者的结合，体现在教师通过加工语料、设计相应教学活动来促进学习者的语言发展。以上语料库素养所包括的知识和技能被Meunier（2020）纳入“整合技术的学科教学知识能力框架”（technological pedagogical content knowledge，简称TPACK）中。

Coxhead教授的新著《如何利用语料库进行语言教学》是英语教师专业素养（PCK for English Language Teachers）丛书之一。作者写作本书的目的是介绍已有的语料库分析成果，提供资源和工具，帮助教师运用语料库辅助课堂教学（p.22）。Coxhead教授的词汇研究成果丰富，她研发的学术英语词汇表在应用语言学领域有较大影响和广泛应用。本书从英语词汇的视角，透视了语料库辅助英语教学的原理、方法和资源。

## 2 内容简介

本书共八章，第一章是概论，第二章到第五章分别从词汇、多词单位、学术英语和专门用途英语角度介绍了语料库辅助语言教学，第六章介绍语料库与课堂教学结合的途径，第七章介绍了语料库辅助语言教学的建库方法和研究方法，最后一章提供了一系列的资源。

第一章介绍了语料库，并简要说明了语料库与教学的关系。语料库是可被计算机分析的文本集合，作者将语料库分为笔语、口语、多模态和口笔语4种类别。首先，作者从文本来源的角度举例说明了笔语语料库，这些来源包括报纸和期刊等出版物、简写读物和学生作文等。口语语料库的文本来源包括教师话语和学生话语，另外，也有口笔语两种形式构成的语料库。多模态语料库的语料包括文本、图片、图表和视频等多种模态来源。然后，作者推荐了教师可以使用的BNC和COCA等网络语料库，并简要介绍了使用方法。作者提醒，语料库的选择要与搜索目的相匹配，使用者要了解语料库大小、语料种类、来源和语料清洁方法。本章最后，作者总结了语料库有利于教学之处，她认为，通过语料库，师生可以获得大量的语例，学习者可以从不同语域和其他角度更多地了解词汇，语料库可以帮助验证学习者词汇用法和搭配的直觉。但是，使用语料库也有很多挑战，例如，很难找到完全适合学习者水平的语料库，需要花费时间来学习和使用语料库。

第二章将语料库与单词联系起来。本章的两个基本概念是词频和词族：词频对词汇教学具有指导意义，了解词频可以帮助确定教学目标，更重要的是可以帮助了解口头和书面语篇理解所需的词族数量；词族由词头（headword）和与其相关的派生词、屈折变化所产生的词形等构成。作者首先介绍了高频词，一般指英语中前2,000—3,000词族，作者通过多个例子展示了高频词的重要性，例如，英语中前2,000词族覆盖了95%的以教学为目的的流行歌曲文本词汇和ESL教师课堂话语词汇。然后，作者介绍了中频词，并简要提及了低频词。中频词包括英语前4,000—9,000的词族，在小说和报纸的阅读中发挥重要作用。最后，作者分别介绍了测量书面语文本和口语文本词频的方法。作者在这两部分通过图表展示了报纸、小说、教材和学术书籍的词汇负载量（vocabulary load）。作者指出，如果学习者掌握覆盖书面语文本95%的词族，即4,000—5,000词，即可达到对文本的基本理解；口语文本方面，要掌握3,000词族，才能达到对科学教师课堂话语文本的理解。但是，要想达到更好的理解，学习者的词汇量应覆盖文本98%的词汇，但所需词族数量会显著增加，例如，要想较好地理解科学学科教师话语，即学习者词汇量覆盖科学学科教师话语的98%，需要学习者掌握7,000词族。

第三章聚焦多词单位。作者认为，多词单位数量众多，掌握多词单位能提升学习者阅读的流利性，还能帮助提高学习者的语言输出质量，应该受到重视。多词单位是词语的不同组合，涵盖了搭配、短语、词块等多种概念，多词单位并非

无序组合,而是有其型式(pattern)。作者先从搭配角度介绍了多词单位的划分方法,这些角度包括:搭配语法结构,例如名词+名词,动词+名词等;目标词左右两侧的型式;连续性和非连续性搭配,例如,highly likely是连续性搭配,两词之间不能加入词汇,而the consequence of则是非连续搭配,consequence之前可以加入economic等词。然后,作者以英语口语常用搭配表(Shin & Nation 2008)为例,探讨了通用英语中多词单位表的研制方法和原则。作者还为对英语变体有需求的教师提供了多词单位地域变体的参考文献。在本章最后,作者从短语角度介绍了如何使用语料库来辨别文本中的多词单位,其主要方法是使用网络工具和已经研制出的词表,她推荐了Compleat Lexical Tutor网站和两个短语表:短语动词表(The Phrasal Verb Pedagogical List)和短语表(The Phrasal Expression List)。

第四章关注学术英语语料库及其在教学中的应用。作者首先将学术英语定义为口笔学术英语语篇中使用的语言,包括通用学术英语和专门用途英语。然后,作者介绍了三个通用学术英语词表:学术词表(Academic Word List,以下简称AWL),学术词汇表(Academic Vocabulary List,以下简称AVL),学术口语词表(The Academic Spoken Word List,以下简称ASWL)。AWL尽管开发较早,且所使用的语料库较小,但是其开发过程严谨,语料来源具有代表性,词表经过了效度验证,并提供了网络资源(包括词表和标注工具)。作者从教学角度提醒,如果词表按首字母顺序排列,而非按出现频数排列,学习者一是不能体会到某些词汇的重要程度,二是那些开头相同的词汇会对学习造成干扰(例如,多个以con-开头的词会让学生感到迷惑)。AVL基于来源于更多学科的较大语料库研发,其选取原则包括词性、词频、比率分布等,并提供了实用的网络资源。ASWL辨识了通用口语词汇,其选取原则包括范围、词频和分布三方面,包含的1741个词族覆盖面更广,ASWL研发者还做了严谨的效度验证。在介绍完通用学术英语词表后,作者又从语料库、研制原则和用法等角度介绍了大学科学专门词表、两个中学不同学科词表和一个科学口语词表。除单词外,研究人员还研发了多词单位表,包括学术搭配表、学术程式语表、学术多词单位表等,作者提供了这些多词单位表的网络下载方式和线上识别工具。最后,作者简要介绍了学术文本中的成语。

第五章聚焦语料库在专门用途英语中的应用。作者首先描述了专门用途英语词汇的特征,专业词汇一般出现在某一特定学科,必须是学习者在听说读写等运用过程中出现的。接着,作者提出了使用语料库辨识专业词汇的词频原则和技术性原则,词频通过语料库来确定,词汇的技术性通过核对语料、咨询专家、查阅专业词典和线上材料来作出判断。作者提醒读者,不要忽视那些高频的具有专业意义的日常词汇。然后,作者用较大篇幅讨论了如何使用语料库来辅助专业词汇教学,首先是专业词汇的辨识,学习者可以借助词汇专业性的质性分析量表来辨识专业词汇,还可以通过语料库对比辨识专业词汇,尤其是通过关键词技术来辨

识专业词汇。另外,还可以使用ESP词表来辅助某一特定领域的教学,例如,教师可以使用工程词汇表来辅助工程学科教学。有的词表开发者还提供了专业词汇的在线辨识工具,识别文本中的专业词汇,帮助精准确定教学目标。最后,作者介绍了如何使用ESP词表估计专业词汇占比,作者认为,了解ESP词汇占比有助于确定具体专业词汇数量和所需教学投入,为精准确定教学重点目标提供了指导。

第六章从词汇角度介绍语料库辅助课堂教学。分为词汇教学的计划、词汇学习策略的训练、词汇测试和词汇教学四部分(Nation & Webb 2011)。首先,作者介绍了在聚焦意义的输入和输出、流利性和语言学习四个维度中(Nation 2007),如何使用语料库来辅助教学,例如,在输入和语言学习环节,可以运用索引行;在输入和流利性训练中,可以使用基于语料库数据加工过的阅读材料。在具体工具方面,教师可以用词汇表帮助计划词汇教学,教师从词表中挑选目标词汇后,在语料库中搜索目标词的信息,用来设计教学材料和练习,也可以指导学生自主查询。除了词汇的一般教学,教师还可以和学生一起对比词汇或多词单位在不同语料库和不同语域中用法的异同。然后,作者介绍了如何培训学生使用词汇学习策略,在课堂教学中,教师可以先让学生在文章中划出多词单位,再运用词典和语料库验证自己的假设,在此部分,作者详细介绍了使用COCA查询搭配等多词单位的方法。在测试方面,作者介绍了两类词汇测试:一是基于词频的测试,用来诊断学生的词汇掌握情况;二是词汇量测试,可以发挥分班测试或进展测试的作用。作者还介绍了词汇测试在教学中的运用,在写作教学中,语料库可以帮助学习者诊断自己写作中的词汇问题,例如高频词的使用量、重复使用的词语等。在教学部分,作者最后较为详细地介绍了自己的一个基于语料库的词汇教学活动。

第七章聚焦语料库辅助语言教学的建库方法和研究方法。首先,作者介绍了教学语料库建设要遵循的原则和要做出的决策,包括语料代表性、借鉴已有语料库的优缺点、语料库大小、语料库结构、语料库数量、词数计算方法、语料库清理和伦理等。作者提及,她开发的学术英语词汇表以多个大学科目的语料为基础,保证了语料的代表性。作者特别强调了词表计算方法的重要性,以不同的单位来计算,例如,词型、词符、词族或词元等,会产生较大差异。另外,她还提醒研究者注意语料清洁和语料使用的伦理,确保每个词都有意义,确保不侵犯语料的知识产权。然后,作者介绍了Nation的词表评估框架,该框架包含八个维度:词表目的、词表计数单位、来源语料库、主词汇表、附属词汇表、词表研发方法、自我批评和可及性。接着,作者展示了如何使用以上框架来评估基于语料库的研究。最后,在研究方法方面,作者建议,读者可以从复制研究入手,使用已有的研究方法,做基于语料库的研究,作者还推荐了能够发表复制研究的期刊和免费语料库软件。

第八章为教师推荐了一系列的语料库资源。在网站方面,作者推荐了本书中



多次提及和使用的 Laurence Anthony 的网站、Compleat Lexical Tutor 网站和 SKELL 网站。作者还推荐了在线语料库,包括通用英语语料库(如 BNC)、学习者英语语料库(如 ICLE)和学术英语语料库(如 BAWE 等)。另外,作者推荐了两个词汇量测试和一个词汇水平测试及其网络链接。在文献阅读建议部分,作者推荐的文献既包括经典之作(如 Sinclair 1991),也有面向下一代数据驱动教学研究的文集(Crosthwaite 2020)。最后,作者推荐了五种与语料库辅助教学关系密切的期刊及其网站。

### 3 简要评价

Coxhead 从语料库概念和分类出发,介绍了基于语料库的单词、多词单位、学术英语词汇和专门用途英语词汇领域的最新研究成果及其在教学中的应用,然后系统介绍了如何利用这些成果进行语言教学和语料库辅助教学的研究方法,最后推荐了语料库资源。下面从理论贡献、实践价值和资源价值角度进行简单评价。

首先是其理论价值。除了作者旁征博引的各类前沿研究成果之外,与本书话题最相关的且最具理论价值的,是作者在第六章建构的一个语料库辅助英语教学的 Coxhead 框架。该框架的第一层次是教学的 4 个重点:计划、策略、测试和教学(Nation & Webb 2011),这四个重点与我国现在提倡的教学评一体化大体重合。Coxhead 框架的第二层次是较为具体的活动,计划部分的第二层次首先用 Nation (2007) 的四维教学理论指引方向,避免教学在输入、输出、流利性和语言知识各个方面失衡,然后推荐了辅助计划的其他工具,包括词表、索引行和专门用途语料库。策略部分的第二层次包括培训学习者进行各类查询的方法。测试部分推荐了多种用途的测试工具。教学部分包括语料库辅助教学的间接应用和直接应用。Coxhead 框架的特点是兼顾了语料库语言学知识和技能、教学法、英语学科知识和学生自主学习 4 个方面,从课程的角度审视语料库辅助语言教学,较为宏观。相比之下,Nesi (2013) 提出的语料库辅助学习的个人模式、何安平等(2020)采用的教学设计框架与笔者曾提出的教学模式(张春青 2021),都较为微观。由于本书以词汇教学为主,Coxhead 框架也是以词汇教学为基础展开论述的,所以笔者认为,这个框架具有开创性和开放性,可供学界进一步丰富和补充。

第二是其实践价值。本书的实践价值体现在多个方面。一是每个章节中穿插的任务,全书共有 49 个任务,大致可分为三类:一类是给出文本,请读者辨识和比较各类词语单位和意义,帮助读者理解和廓清概念,例如,第三章请读者划出多词单位,第四章请读者辨识学术英语词汇等;第二类是在讲解各类标准和理论之后,请读者联系实际,思考和应用,例如,第五章请学习者使用词汇技术性的判断标准,第六章请学习者运用四维理论;第三类是语料库和相关工具的使用,



这些工具包括 Compleat Lexical Tutor、COCA 和词汇测试等。实践价值体现之二是每个章节末尾为学习者布置的作业，这些作业任务分为两类，一是概念和理论的回顾、理解和运用，二是对语料库资源和工具的实际运用。实践价值体现之三是每个研究成果中体现出的研究方法，本书介绍了多个词汇表及其研发方法和过程，虽然简略，但是具有窥斑见豹的作用，读者可以在完成前述各类任务后，按照研究者的思路自己探索。

第三是资源价值。在语料库语言学中，资源包括语料库、网站和测试等，掌握的资源越多，教师在辅助语言教学方面越得心应手。本书提供了大量资源，除了单独辟出一章讨论资源之外，还通过参考文献提供了大量的论文、论著和教材资源，且这些文献反映了本领域的最新进展，其中最新的是 Rogers *et al.* (2021) 的学术英语多词单位词汇表。另外，每章的节选、图表和本书的附录，也是资源的一部分。最后，作者提供了大量的网络资源，包括网络语料库、词汇测试和可以下载的工具等。事实上，无论是教师还是学生，初学语料库，应该先熟练使用一两个工具或资源，例如，初学者可以学习使用 AntConc 软件分析常用的文本，配以网络资源 Compleat Lexical Tutor 或 SKELL 就可以完成绝大多数分析任务了。如有必要，再扩展到 COCA、BNC 和 WordSmith 等工具的使用。

尽管有上述优点，但是本书也有一些可改进之处。首先是索引行的处理。一般来说，在索引行中，被检索词应该纵向居中排列，这样利于学习者观察和发现语言规律，这也是语料库索引行能够帮助学习者自主探究和发现规律的优势条件之一。但是本书却把索引行变成长短不一的句子，与词典中的排版形式类似，其结果与查词典无异。再者，与学科教学知识有关的书籍往往重视理论的落地，书中有很多有价值的语料库辅助语言教学的理念和做法，如果能够做出教学设计并提供短视频，则学习者能够获得更直观的感受。何安平等（2020）编选的案例提供了微本和视频讲解，反映了国内语料库辅助英语教学的发展，如果 Coxhead 教授能够提供国外的案例和视频则更能拓宽读者的视野。综合来看，本书为读者提供了基于语料库的研究的最新成果，并提供了将这些成果应用于教学的途径，在理论和实践两个层面上为语料库辅助语言教学领域作出了贡献。

### 参考文献

- BOULTON A, COBB T. Corpus use in language learning: a meta-analysis [J]. *Language Learning*, 2017(2): 348-393.
- CALLIES M. Integrating corpus literacy into language teacher education [C]//GÖTZ S, MUKHERJEE J. *Learner corpora and language teaching*. Amsterdam: John Benjamins Publishing Company, 2019: 245-263.
- CROSTHWAITE P. Data-driven learning for the next generation: corpora and DDL for pre-tertiary learners [C]. London: Routledge, 2020.

- JOHNS T. From printout to handout: Grammar and vocabulary teaching in the context of data-driven learning [C]//ODLIN T. Perspectives on pedagogical grammar. Cambridge: Cambridge University Press, 1991/1994: 27-45.
- MA Q, TANG J, LIN S. The development of corpus-based language pedagogy for TESOL teachers: a two-step training approach facilitated by online collaboration [J]. Computer Assisted Language Learning, 2021, 35(9): 2731-2760.
- MEUNIER F. A case for constructive alignment in DDL: rethinking outcomes, practices, and assessment in (data-driven) language learning [C]//CROSTHWAITE P. Data-driven learning for the next generation: corpora and DDL for pre-tertiary learners. London: Routledge, 2020: 13-30.
- NATION I. The four strands [J]. Innovation in Language Learning and Teaching, 2007, 1(1): 1-12.
- NATION I, WEBB S. Researching and analyzing vocabulary [M]. Boston: Heinle, 2011.
- NESI H. ESP and corpus studies [C]//PALTRIDGE B, STARFIELD S. The handbook of English for specific purpose. Malden: Wiley Blackwell, 2013: 239-246.
- ROGERS J, MULLER A, DAULTON F, et al. The creation and application of a large-scale corpus-based academic multi-word unit list [J]. English for Specific Purposes, 2021, 62(4): 142-157.
- SHIN D, NATION P. Beyond single words: the most frequent collocations in spoken English [J]. ELT Journal, 2008, 62(4): 339-348.
- SINCLAIR J. Corpus, concordance, collocation [M]. Oxford: Oxford University Press, 1991.
- 何安平, 许家金, 张春青. 语料库辅助中学英语教学案例选编 [M]. 北京: 外语教学与研究出版社, 2020.
- 张春青. 语料库辅助英语教学: 原理、模式与实践 [J]. 教学月刊, 2021 (4): 3-7.

通信地址: 310023 浙江省杭州市 浙江外国语学院应用外语学院

# 《学术界的跨学科实践：写作、教学与评估》述评

北京外国语大学 陈雅刚

Louisa Buckingham, Jihua Dong & Feng (Kevin) Jiang (eds.). 2023. *Interdisciplinary Practices in Academia: Writing, Teaching and Assessment*. Oxon: Routledge. xi+196pp.

## 1 引言

跨学科研究的主要目标是解决科学界和社会中复杂的现实问题，在当今学术界扮演着重要角色（Barry & Born 2013）。在全球高等教育机构的积极推进下，相关学术期刊和著作不断涌现。跨学科研究的蓬勃发展对应用语言学研究提出了新的挑战。近年来，不同学科的边界逐渐模糊，融合了传统学科特征的交叉学科不断产生，使学术英语和专门用途英语的教学与研究更为复杂。EAP/ESP领域致力于描述不同学科的交际行为，有非常强的学科导向，在描写单一学科语言特征、学科内和学科间语言特征差异等方面成果丰硕（Hyland 2004；Gray 2015；Ziaeeian & Golparvar 2022等），但与新兴学科、交叉学科和跨学科相关的研究尚显不足。跨学科研究的语言特征是什么，跨学科语境下如何进行EAP/ESP教学，高等教育机构如何培养跨学科的学术素养和技能等等，都是EAP/ESP研究亟待解决的问题。该书聚焦跨学科语境下的EAP/ESP研究，从文本体裁及跨学科话语特征、EAP/ESP课堂教学与评估、跨学科专业设置和课程研发3个方面对上述问题进行了深入探讨。

## 2 内容简介

全书由导言和正文两部分组成。导言由Louisa Buckingham和Jihua Dong共同撰写。该节首先阐述了跨学科的概念、形式及其对应用语言学，尤其是在语言教学和高等教育方面的影响。随后，作者对各章内容作了概述，并肯定了本书的学术价值。正文分为两个部分，共十章。第一部分包括一至四章，是关于跨学科话语特征的研究，主要关注跨学科书面语文本中的语言特征和话语资源，包括体裁（genre）、引用（citation）、文本声音（textual voice）、评价语（evaluation）、立场标记语（stance marker）等。第二部分包括五至十章，从更宏观的角度讨论了跨学

科的相关问题,包括学术英语课程设计,学术素养课程的开发和教学,以及语言能力评估等。

第一章中,Kalan对3位英语二语者在日常英语使用中的体裁灵活性(genre agility)作了理论分析。体裁灵活性即灵活使用不同体裁的能力。跨学科(interdisciplinarity)的“跨(inter)”意味着跨学科知识产生于既有学科“之间”。这种中间性(in-betweenness)的认知状态使我们在汇报跨学科的研究内容时需要灵活使用写作体裁。因此,体裁灵活性是跨学科交流的基础。作者在本章中提出了基于体裁灵活性的新理论模型,该模型是一个包括跨学科、话语灵活性(discursive agility)、诠释灵活性(hermeneutic agility)、符号灵活性(semiotic agility)和体裁灵活性的嵌套模型。Kalan分析了一项对3位作家的民族志研究案例,表明跨学科语境能够催生话语、诠释和符号灵活性,从而使3位作家习得了体裁灵活性。Kalan认为这种混合式跨学科体裁形式在修辞和认识论方面都是多维度的,能够丰富我们对跨学科现象的理解,促进EAP研究的发展。

第二章中,Schembri探讨了引用文本的特定语言特征在构建跨学科文本声音中的作用。她以某跨学科教育专业的本科毕业论文为语料,就引文的不同形式进行了讨论。较弱的引文框架表现为非融合性、非转述性和文本转换性。强势的引文框架则表现为融合性、转述性和文本复制性。她发现在较弱的引文框架中,跨学科文本声音较强,反之亦然。其中,转述动词的使用被认为是弱势文本声音的重要标志,而显著较低的引文密度也不利于文本声音的构建。Schembri也提出了一些对于跨学科作者在引用文本方面的建议。

第三章中,Muguiro对比研究了教育神经科学、经济史和科技研究这3个交叉学科及其发源学科的期刊论文中表示“重要”含义的评价类形容词,并以形容词important为例进行了定性与定量研究相结合的个案研究。作者建立了一个重要性的三维评价模型,包括评价对象(主题导向、研究导向、学科导向)、评价者(原作者、其他研究者、参与者)和评价语境(研究语境、学科语境、真实世界语境、其他)。研究发现,形容词important在跨学科期刊论文中的使用频数显著高于其发源学科。在跨学科的文章中,形容词important的评价对象、评价者和评价语境都与发源学科文章不同。这一研究为捕捉反映学科差异和互动的语言特征提供了思路,有助于推进跨学科写作研究。

第四章中,Dong和Buckingham关注了生物信息学及其发源学科(生物学和计算机科学)的书面学术文本,探讨三者立场标记语使用方面的异同。总的来说,在生物信息学文章中,立场标记语的使用显著低于两个发源学科。认知类立场标记语是3个学科中最高频的立场标记语类型,而模糊限制语是最低频的立场标记语类型。在各类型立场标记语的使用比例上,生物信息学与计算机科学更为相似。作者将3个学科在立场标记语使用上的相似特征归因于它们相似的研究对

象和研究方法，而差异则来源于生物信息学这一新兴学科独特的学科范式。该研究有助于学术英语教学，加深对于跨学科合作中语言和话语风格的理解，促进学术交流及跨学科实践。

第五章中，Davies介绍了中国昆山杜克大学学术英语教学框架的设计和实施情况。该教学框架在为学生打下各自所属学科坚实基础的同时，非常注重跨学科的学术交流与合作。在这一框架下，学术英语课程分为学术交流技能课程和学术写作课程。第一阶段的学术交流技能课程让不同学科的学生在同一学术英语课堂中进行学习交流，为跨学科的融合、理解和交流提供了平台和机会。在掌握通用学术英语技能的同时，学生能够体验不同学科的思维方式，提高对所属学科和学科边界的认识。第二阶段的学术写作课程则进行了分学科教学，为学生打下了各自所属学科的坚实基础。基于该课程的设计、实施，以及来自师生双方的教学反馈，作者为发展跨学科的研究生学术英语课程提供了一系列建议。

第六章中，Mahawattha等学者指出，培养学术素养的任务在本质上是跨学科的。他们阐述了学术英语教师和理工科类讲师的跨学科关系，通过跨学科协作的方式，借鉴Carr *et al.* (2018)提出的理论框架，设计了以培养科研能力为目标的物理系本科生学术素养课程和教学方法。作者详述了这一过程的4个层次（背景、过程、中期成果和教育产出）在个体层面的影响。该课程研发项目使不同学科背景的研究人员形成了共享的话语空间，并在此基础上进行了个体学习和跨学科实践。这一跨学科合作案例对于斯里兰卡全英文授课（EMI）的学位课程发展具有积极作用。

第七章中，Bedeker和Gaye首先描述了阿联酋某大学健康科学专业本科生的学术困境，探讨了阈值概念，并指出了现行学术英语课程的问题。研究者结合专门用途英语的体裁知识和系统功能语言学的教学循环框架提出了一种跨学科教学法，旨在帮助学生学习健康科学研究中的元语言特征。这一课程框架主要包括四个阶段：第一、二阶段是对元语言、修辞资源和学科知识的学习和模仿，通过多样的任务和活动，学生逐渐熟悉学术写作的规范和惯例，能够阅读和理解学术文章；第三、四阶段包括独立写作和协同写作，使学生能够系统地掌握文献综述的写作技巧。该案例说明相较于传统学术英语课程，基于跨学科和体裁的语言教学法有诸多优势，能够提高学生的参与度和积极性。

第八章中，Badenhorst介绍了一门博士生学术研究写作课程。该课程既探讨了学术研究写作中的阈值概念，也能够灵活地跨越学科和研究范式。在该课程中学生们进行了跨学科交流，并确定了情商、自我意识、写作过程知识、话语共同体知识、修辞知识、体裁知识和学科知识这七个重要阈值概念及其优先级，这些概念对于学生学术写作实践的元认知发展有重要作用。

第九章中，Altiparmak展示了一项结合心理语言学和专门用途英语的跨学科



方法在航空英语口语测试中的应用。该研究从心理语言学的角度关注言语流利性问题,旨在分析飞行员候选人在航空英语考试中执行两个不同语音任务时产生的言语不流利现象,并根据航空英语考试的评分标准对其进行评分。该研究能够加深我们对于考生不流利表达的类型以及考官们评定方式的理解,对于航空英语语境下的教学内容和形式具有启示作用。

最后,在第十章中,Kinuthia调查了肯尼亚高等教育机构中具有跨学科导向的活动类型。研究发现,肯尼亚私立大学已将跨学科嵌入其核心活动:专业设置、社区参与和学术研究中。在专业设置方面,私立高校提供了兼具灵活性与综合性的跨学科、多学科和超科学学位课程。在社会参与方面,私立高校与多个机构、企业和学校建立了合作伙伴关系,为学生提供了实践机会,使其获得了跨学科的技能,优化了学习体验。在学术研究方面,肯尼亚私立大学鼓励其师生开展跨学科研究。

### 3 简要评述

本书内容丰富,涉及话题广泛,既包括语言内部的语言本体研究,也包括语言外部的语言教学和更为宏观的高等教育研究。纵览全书,本书具有以下特色。

第一,本书具有学术前沿性。在学术英语领域,跨学科研究是近年来逐渐兴起的学术焦点,也是学科未来发展的趋势。对于跨学科领域的语言特征研究目前仍处在探索阶段。本书的探索呈现了跨学科研究和学术英语研究的前沿动态,体现了本书的创新性。跨学科的定义和概念在学界至今众说纷纭,缺乏明晰的界定,本书收录的研究对跨学科也有不同的理解:有的研究聚焦由传统学科结合演变而来的新兴学科(例如,教育神经科学、经济历史学、生物信息学)。有的研究致力于促成传统学科之间的相互合作(例如,来自不同传统学科的学生共同完成一项研究)。有的研究关注学界与社会各界(例如,企业、政府部门、中小学等)的超学科合作等。这些不同角度的研究能够丰富和发展我们对于跨学科概念的理解,有助于推动学科/跨学科领域内的知识边界,激发更多相关研究和探索。

第二,本书兼具理论和应用价值。理论方面,从语言研究角度,Kalan(第一章)指出,学术英语对于新兴跨学科研究的关注不应该是创造新体裁,而是挑战狭义的和过分简化的体裁认识论框架。Kalan提出了跨学科语境下的新体裁理论——体裁灵活性。该理论是对现有体裁理论在跨学科语境下的深化和发展。Muguiro(第三章)提出了由评价对象、评价者和评价语境构成的重要性三维评价模型。该模型有助于对跨学科文本的重要性评价进行细颗粒度分析。从语言教学角度,Mahawattha(第六章)、Bedeker和Gaye(第七章)均继承和发展了前人的



理论框架和教学框架。应用方面，首先，本书为教学实践提供了新思路。本书涵盖了跨学科背景下不同非英语本族语国家的学术英语、学术素养、学术写作课程等多种课程的教学及反思，能够给EAP/ESP领域的教师带来启示。此外，本书研究方法多样，包括个案研究、访谈研究、基于语料库的研究等等，量化、质性和混合的研究方法和数据类型均有涉及，对从事跨学科语言研究的科研人员具有借鉴意义。

第三，本书具有文化多元性。本书收录的研究在多元语言和不同文化语境下开展，研究对象背景多样、身份多元，包括移民加拿大的多语言作家、阿联酋健康科学专业本科生、中国中外合办类大学师生、斯里兰卡的学术英语教师和物理专业课教师、肯尼亚私立大学师生等。其中多个研究的研究对象为非英语本族语者，为跨学科实践提供了多样的文化视角。

## 4 未来研究展望

整体上看，本书是跨学科语境下不同类型的学术英语和专门用途英语研究的有效整合。但同时需要注意的是，本书包含的跨学科学术话语研究角度较为分散，相互之间没有建立起有效的联系。这在一定程度上也与当前跨学科话语研究领域的发展状况有关：前期文献较少，学界尚未形成跨学科语言特征的整体面貌。本文认为未来这一领域的研究可从以下3个方面进行系统性的拓展：第一，可从词汇、句法、语义、语用等不同层级进行更为广泛充分的实证研究；第二，在文本的选择上，未来可以扩展到其他学术文本类型，如演讲、学术会议口头报告、跨学科教材、课堂录音等等；第三，在研究方法的使用上，可以利用更加前沿的技术深入挖掘实证数据，如采用多维分析、聚类分析、主成分分析等统计方法或词向量技术，尝试发现跨学科学术话语的系统性特征。

其次，本书中部分研究的样本容量较少。例如，第二章中Schembri的引文研究只包含了六篇本科生学位论文的文献综述部分，使得该研究结果的代表性受限。未来研究时，可在一定程度上扩大样本容量，与本书中的研究结果进行对比。

另外，本书中多项研究都将跨学科话语的特殊性归因于认识论上的差异。例如，Kalan认为跨学科流动的写作体裁是由其在认识论上的中间性决定的；Muguiro的研究佐证了多元认识论是跨学科的固有特性，跨学科知识是在不同认识论的互动之间产生的；Mahawattha等人认为，尊重不同学科本体论和认识论的差异是学习和沟通的重要内容。在未来的跨学科研究中，可以考虑进一步对比分析不同学科/跨学科之间认识论的差异及其对跨学科的话语特征和合作难度的影响。

总体而言，该书为未来的跨学科语言研究者提供了新思路，为学术英语/专门用途英语的教学和评估人员提供了新框架、新方法、新路径，为教育行政管理人

员的跨学科实践提供切实可行的指导。本书颇具学术价值和教学意义，对于跨学科话语和学术实践感兴趣的读者可以参考。

### 参考文献

- BARRY A, BORN G. Interdisciplinarity: reconfigurations of the social and natural sciences [C]//BARRY A, BORN G. Interdisciplinarity: reconfigurations of the social and natural sciences. London: Routledge, 2013: 1-56.
- CARR G, LOUCKS D, BLÖSCHL, G. Gaining insight into interdisciplinary research and education programmes: a framework for evaluation [J]. Research Policy, 2018, 47(1): 35-48.
- GRAY, B. Linguistic variation in research articles: when discipline tells only part of the story [M]. Philadelphia: John Benjamins, 2015.
- HYLAND, K. Disciplinary discourses: social interactions in academic writing [M]. Ann Arbor: University of Michigan Press, 2004.
- ZIAEIAN E, GOLPARVAR S. Fine-grained measures of syntactic complexity in the discussion section of research articles: the effect of discipline and language background [J]. Journal of English for Academic Purposes, 2022, 57: 1-12.

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

# English abstracts of major articles

---

## Lexical semantics-based multidimensional analysis of register

.....*QIAN Yubin & SUN Ya* (1)

This study proposes a comprehensive methodology for multidimensional analysis grounded in lexical semantics, incorporating 104 semantic and 67 grammatical features. The method was applied to evaluate international news related to the Beijing Olympics, revealing eight functional dimensions. These dimensions encompassed emotion-driven interactions, the creation of informational content, depiction of landscapes, crafting of abstract narratives, exploration of power dynamics, juxtaposition of objective and subjective assessments, as well as considerations pertaining to humanities, arts, health, safety, community activities, facilities, and technical contributions. The findings of this study validate the effectiveness of the proposed method in distinguishing the unique communicative functions embedded within discourse. Moreover, it highlights the method's potential to augment traditional corpus approaches, which predominantly focus on lexical grammar while often overlooking the critical aspect of semantic value.

## The polysemy of the perception verb “看 (kan)” : a corpus-based behavioral profile analysis

.....*HUANG Jingwen, LI Jinmei & HU Zhiyong* (15)

This study explores the polysemy of the perception verb “看 (kan)” using a corpus-based behavioral profile analysis. The investigation highlights two primary categories of “kan”: one linked to social activity via cognitive metaphor, and another merging perceptual and cognitive senses. The evolution of “kan” from a perceptual to a cognitive sense is ongoing, with some senses showing closer relations than others. This research offers a comprehensive semantic network for the polyseme “kan”, advancing our understanding of polysemy in the Chinese language through behavioral profile analysis.

## Usage and L1 transfer effects of delexicalized verb collocations among Chinese EFL learners across proficiency levels

..... YAN Shengde & GAO Xia (27)

Drawing from the EFCAMDAT corpus, this research examines the progression of delexicalized verb collocations with “make” and “take” in the English writing of Chinese EFL learners across proficiency levels, contrasting them with Spanish English learners to discern L1 transfer effects. Initial findings reveal varying degrees of usage based on proficiency and instances of generalization among learners. This study provides insights into the influence of L1 typological differences on L1 transfer effects, offering guidance for second language instruction and textbook creation.

## Impact of Chinese anti-COVID-19 tweets on overseas audience engagement

..... JIANG Jinlin & WANG Jiahui (43)

This study analyzed 2,500 tweets regarding China’s response to COVID-19, examining how relational acts, engagement quality, and linguistic features influence overseas audience behavior. Results highlighted varying levels of engagement based on tweet content, interactivity, linguistic choices, and more. Collectively, the three tiers of tweet features accounted for 25.2% of the variance in audience engagement. The study’s findings inform overseas communication strategies, content creation, and writing styles in narrating Chinese experiences.

## News value analysis of “novel coronavirus pneumonia pandemic” reports in *China Daily*

..... HAN Cunxin & ZHAO Yufei (58)

Utilizing a custom-built multimodal corpus, this study evaluates news values in reports about the Novel Coronavirus Pneumonia Pandemic by *China Daily* through Corpus-assisted Multimodal Discourse Analysis. Findings indicate a focus on proximity, eliteness, timeliness, and negativity in texts, while images emphasize eliteness, superlativeness, and proximity. This underscores China’s commitment to authoritative, objective, authentic, and scientific communication during major public crises, fostering a positive, efficient, and responsible global image while bolstering national solidarity against the pandemic.

## **Comparative corpus-based critical discourse analysis of Chinese and American news coverage on TikTok**

.....HUANG Xin & LUO Weihua (75)

Integrating Critical Discourse Analysis (CDA) with corpus methodologies, this research scrutinizes the textual corpus of news reports concerning TikTok from *China Daily* and *The New York Times*, utilizing Fairclough's three-dimensional framework. This approach encompasses description, interpretation, and explanation. The findings reveal that while both media outlets underscore the widespread appeal of TikTok, Chinese coverage places a greater emphasis on its safety features. In contrast, American reporting tends to highlight potential security risks associated with the platform. The study traces these divergent perspectives back to the distinct political, economic, social, and cultural contexts inherent to each country, revealing underlying ideological differences between the two media outlets.

## **Acquisition analysis of enumerative conjunctions in the CIA model**

.....LI Yanjiao, LI Qi & ZHUANG Huibin (89)

Leveraging both a Chinese English learner corpus and a native English speaker corpus, this study offers both quantitative and qualitative insights into the usage patterns of “such as” and “for example”. The aim is to discern usage differences between Chinese English learners and native speakers. Key findings include disparities in frequency, co-occurrence patterns, and placement within sentences. Root causes are linked to native language interference, suboptimal example sentence selection, and incomplete dictionary or textbook analysis. Recommendations include enhancing corpus literacy among educators, early error detection, and refining dictionary and educational materials to align with native English usage patterns.

## 语料库语言学

CORPUS LINGUISTICS

## 要目

- |                                 |             |
|---------------------------------|-------------|
| 语域实证研究：基于词汇语义的多维分析法             | 钱玉彬 孙 亚     |
| 视觉动词“看”的多义性：基于语料库的行为特征分析        | 黄静雯 李金妹 胡志勇 |
| 不同水平中国英语学习者虚化动词搭配使用及迁移效应研究      | 闫盛德 高 霞     |
| 中国抗疫推文特征对海外受众行为参与度的影响研究         | 江进林 王佳慧     |
| 多模态话语分析视阈下的新闻价值研究               | 韩存新 赵宇飞     |
| 基于语料库的中美媒体关于 TikTok 新闻报道的批评话语分析 | 黄 馨 罗卫华     |
| 基于 CIA 模型的列举类词习得研究              | 李艳娇 李 齐 庄会彬 |

外研社·期刊出版分社  
电话：010-88819267  
E-mail: qkzx@fltp.com  
网址: www.bfsujournals.com



记载人类文明  
沟通世界文化  
www.fltrp.com



北外学术期刊



iResearch 微信公众号

责任编辑：赵 雪  
责任校对：白小羽  
封面设计：锋尚设计



定价：35.00元