

《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19辑
二〇二三

语料库语言学

CORPUS LINGUISTICS

10年
第19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 CLEC collocation local grammar word embeddings
AntConc DEAP multifactorial analysis
big data corpus WordSmith
Brown Crown TECCL
BNC corpus-as-method MDA semantic prosody
COCA co-selection concordance frequency ToRCH
iWriteBaby
corpus-as-theory ParaConc phraseology

外
研
社

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



corpus.bfsu.edu.cn

语料库语言学

(半年刊)

Corpus Linguistics

(Biannual)

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

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

刊名题字：崔希亮
主 编：许家金
责任校对：刘 华、王 斌

Journal Name Calligraphy: Cui Xiliang
Editor: Xu Jiajin
Proofreaders: Liu Hua & 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 19 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1期

目 录

语境共选

基于LDA主题建模技术的北京冬奥会话语意义研究	张 毓	卫乃兴 (1)
英语“致使-动结”构式宾补共选机制研究	刘聪颖 李潇辰	曹笃鑫 (15)
变异语言学视角下英语并列二项式词序多元定量研究	林奕冰	孟庆楠 (27)

研究论文

计算社会科学、文化组学与语言学	邵 斌	李雨飞 (38)
中外跨国公司英文社会责任报告情感分析对比	宋天祯	黄立波 (48)
学术论文中互动元话语及作者身份建构对比研究	刘国兵	张君兰 (60)
实证类汉语学术期刊论文中的自我指称与作者身份构建	王亚敏 宫 雪	安卓玛 (72)
语料库方法在汉语第二语言习得研究中的应用研究	彭家法	孙梦馨 (87)

研制开发

CQP 语法赋能语言研究及语言学习	吴良平 (98)
汉英人文社会科学文献平行语料库建设.....	邓劲雷 (115)
LawDEAP 法学学术英语语料库的创建.....	王艳伟 干 诚 李俊飞 高利杰 黄张成 (127)
ShipDEAP 船舶与海洋工程学术英语语料库的创建	田 苗 滕如玉 (136)

书刊评介

《学习者语料库研究遇见二语习得》述评	李维静 (146)
《情态与历时句式语法》述评	刘 娜 李福印 (155)
主要文章英文摘要	(160)

Corpus Linguistics

Volume 10, Number 1, 2023

Table of Contents

Featured column: Contextual co-selection approach to language

- Discourse meanings of Beijing Winter Olympic Games based on the LDA model
..... ZHANG Yu & WEI Naixing (1)
- A study on the co-selection mechanism of the object and complement in English caused-
resultative construction..... LIU Congying, LI Xiaochen & CAO Duxin (15)
- A multivariate quantitative study on word order preference of English binomials from a
variationist linguistics perspective LIN Yibing & MENG Qingnan (27)

Research articles

- Computational social sciences, culturomics and linguistics
..... SHAO Bin & Li Yufei (38)
- A comparative sentiment analysis of CSR reports between multinational enterprises home
and abroad SONG Tianyi & HUANG Libo (48)
- A comparative study of interactional metadiscourse and authorial identity construction in
academic theses LIU Guobing & ZHANG Junlan (60)
- Self-mention and authorial identity construction in empirical Chinese academic journal
papers WANG Yamin, GONG Xue & AN Zhuoma (72)
- Application of corpus methods in Chinese second language acquisition
research..... PENG Jiafa & SUN Mengxin (87)

New corpora, tools and methods

- Empowering language research and language learning with CQP syntax..... WU Liangping (98)
- Construction of Chinese-English parallel corpus of humanities and social science literature
..... DENG Jinlei (115)
- Construction of LawDEAP (legal English research article) Corpus
..... WANG Yanwei et al. (127)
- Construction of ShipDEAP (ship and ocean engineering English research article) Corpus
..... TIAN Miao & TENG Ruyun (136)

Book reviews

- B. Le Bruyn & M. Paquot (eds.). *Learner Corpus Research Meets Second Language
Acquisition*..... LI Weijing (146)
- M. Hilpert, B. Cappelle & I. Depraetere (eds.). *Modality and Diachronic Construction
Grammar* LIU Na & LI Fuyin (155)

- English abstracts of major papers..... (160)

基于LDA主题建模技术的北京冬奥会话语意义研究^{*}

北京航空航天大学 张 毓 卫乃兴

提要：LDA主题模型是目前较为常见的无监督主题建模方法，可用于批评话语分析中的主题分析。既往同类研究大多使用语料库语言学中的主题词方法发现主题，而较少使用主题建模方法。鉴于此，本研究选择境外英文媒体有关北京冬奥会的报道为语料，自建北京冬奥会英文报道语料库，采用LDA主题建模技术探究北京冬奥会的境外英文报道主题。此外，我们结合语料库驱动路径分析主题的高频关键词和语义韵，以揭示北京冬奥会的话语建构和媒体态度意义。研究发现，境外媒体在报道北京冬奥会时聚焦于3个主题，即冬奥会主办权、赛事准备工作以及政府举措。此外，境外媒体对北京成为首座“双奥”之城和中国政府积极推广冰雪运动的行为给予正面评价，但同时批评了北京缺乏冰雪运动文化传统和自然降雪条件。本研究表明，将LDA模型、语料库语言学方法和批评话语分析方法结合具有可行性，为今后研究提供了新的分析框架。

关键词：LDA、语义韵、话语建构、北京冬奥会

1 研究背景

主题建模（topic-modeling）是文本挖掘领域中的一种无监督学习方法，它可直接对文本数据进行归纳建模，从而挖掘语料库的主题（刘文字、胡颖 2020）。其中，隐含狄利克雷分布（Latent Dirichlet Allocation，简称LDA）是常用的主题建模方法之一，该模型较为简单且应用广泛，其算法排除了人工干扰，保证了结果的客观性（Mohr & Bogdanov 2013）。Törnberg & Törnberg（2016a, 2016b）较早使用LDA主题模型结合批评话语分析方法，研究了社交媒体中有关穆斯林的话题以及Muslim和Islam的话语表征。他们认为LDA模型适合处理大规模语料库数据，并且与提前预设的关键词分析不同，LDA能够在无监督的情况下对数据进行

^{*} 本文是北京市哲学社会科学研究基地项目“境外主流英文媒体北京冬奥会新闻报道分析研究（2015年7月31日—2020年12月31日）”（19JDYYB005）的阶段性研究成果。

卫乃兴为本文通讯作者。

作者贡献：

张毓：选题构思、研究方法、数据收集、数据分析、初稿撰写、字数占比（60%）；

卫乃兴：选题构思、研究方法、讨论结论、字数占比（40%）、修改润色。

结构化归纳，是语料库驱动的研究。随后，Jo（2019）探讨了话语分析中使用主题建模的可能性，认为主题建模能够应用到话语分析中主要基于3个原因：首先，主题建模产生的主题包含了话语中的两个重要信息，即高频词汇和词与词之间的关系网络；其次，主题建模认为文本由多个话题构成，能够帮助理解多重话语的动态性（dynamics of multiple discourses）；最后，一些主题建模方法（如动态话语模型DTM和结构话语模型STM）能够追踪话语的历史变化。Jacobs & Tschötschel（2019）的研究则指出，主题建模能够在方法上为话语分析提供有力帮助，可以弥补话语分析在方法上的缺陷（如分析的主观性、缺乏系统性和操作性）；并且在元理论层面（meta-theoretical level）和认识论层面（epistemological level），两者也具有较高的适配性。Aranda *et al.*（2021）认为，将批评话语分析（Critical Discourse Analysis，简称CDA）与结构主题建模（Structural Topic Model）相结合能够拓展传统的CDA方法，并且实现互补。

上述研究表明，批评话语分析中使用主题建模方法挖掘语料库文本的主题可以实现优势互补，而且能够保证结果的客观和准确，避免人为预设和参照语料库的干扰。然而，当前基于语料库的批评话语分析较少使用主题建模，而多采用主题词方法分析文本主题，并通过词语搭配和词语索引对某一话题领域的文本进行话语建构或话语策略分析（如Baker *et al.* 2008, 2020；Engström & Paradis 2015；Song *et al.* 2021；杨敏、符小丽 2018；赵永刚 2021等）。具体而言，主题词方法需要将观察语料库与参照语料库的词表进行对比，运用卡方检验或对数似然率等手段，统计观察语料库中显著性高频使用的词汇，从而生成主题词表。不难发现，主题词方法中，使用不同的参照语料库，得到的主题词表也会有所不同。换言之，参照语料库的选择会影响最终的主题词结果。

当前有关北京冬奥会新闻报道的研究多集中于叙事框架研究和基于语料库批评话语分析的国家形象研究（刘静轩等 2022），鲜有研究运用LDA主题建模对北京冬奥会英文报道的主题进行分析。鉴于此，本研究将收集自北京冬奥会申办成功至2021年9月1日间境外媒体有关北京冬奥会的英文报道作为研究语料。研究采用LDA主题建模，结合Sinclair（2004）提出的扩展意义单位模型对相关英文报道进行分析，从宏观和微观两个层面揭示北京冬奥会英文报道中的话语意义建构。

2 隐含狄利克雷分布（LDA）主题模型

本节将介绍LDA主题建模的发展脉络、原理思想，以及实践中操作LDA模型的主题数目设置。

2.1 LDA主题建模的发展脉络及基本原理

LDA主题建模由Blei *et al.*（2003）提出，是一个可用于文本语料库的生成概

率模型。LDA模型是在概率潜在语义索引（probabilistic Latent Semantic Index，简称pLSI）模型的基础上发展而来。pLSI模型由Hofmann（1999）提出，用于计算文档中主题的概率。此模型认为，文档的主题符合多项分布，每个文档以一定的概率生成某个主题，一篇文档由多个占据不同比例的主题组成；而文档中的每个词按照一定的概率由某个主题产生，也符合多项分布。例如，在一个包含 N 篇文档的语料库 D 中，文档中的词通过以下过程生成：（1）以概率 $P(d)$ 选定一篇文档 d ；（2）以概率 $P(z|d)$ 选择一个不可观测的主题 z ；（3）以概率 $P(w|z)$ 生成一个单词 w （Hofmann 1999：51）。

pLSI模型的任务是根据可观测变量，即文档（ d ）和词（ w ），用概率统计的方法求解隐含变量——主题（ z ）的概率。pLSI模型中，文档的主题概率是确定的，并没有在文档层面生成概率模型，因此会导致模型中的参数随语料库规模扩大而线性增长，从而出现过度拟合的问题；同时，也无法将主题的概率分布应用到除训练集外的其他文档中（Blei *et al.* 2003：994）。换言之，随着语料库中文本量的增长， $P(z|d)$ 的参数也会随之增加，从而导致模型过度拟合。另外，对于语料库 D 之外的新文档 d_m ，我们无法获取其对应的 $P(d_m)$ 。

为此，Blei *et al.*（2003）引入了贝叶斯统计，将pLSI模型发展为LDA主题模型。LDA是一个三层贝叶斯模型，由“文档—主题—词”构成，通过主题和单词的狄利克雷先验分布，结合观测到的数据（即单词）来求解主题的后验分布。LDA模型的基本理念是语料库中的文档可表示为若干随机的隐含主题，每个主题是若干单词的概率分布（Blei *et al.* 2003：996）。

LDA模型中，一个文档中的某个单词分两个阶段生成（Blei 2012：78）：

首先，随机产生一个主题的概率分布。

其次，对于文档中的每个单词：（1）在上一阶段生成的主题中，随机选择一个主题；（2）在所选主题对应的单词中随机选择一个单词。

在以上过程中，第一步中随机产生的主题概率分布即是一个狄利克雷分布，是文档中主题概率的先验分布。换言之，LDA模型为pLSI模型中的 $P(z|d)$ 加了一个先验分布——狄利克雷分布。因此，LDA模型中，主题的概率分布是随机变量；而pLSI模型中，文档主题的概率 $P(z|d)$ 则是确定值。

现用一个例子说明LDA模型中文档的生成过程。假设一个有关学术论文写作研究的英文学术文本语料库，包含“语篇结构”“语言特征”“学科差异”“作者群体差异”4个主题。根据LDA模型，如果要生成语料库中的一篇文章，首先需要分别给前述4个主题随机分配一个概率，我们可以假设概率分布为{“语篇结构”0.4，“语言特征”0.3，“学科差异”0.2，“作者群体差异”0.1}（所有主题的概率之和为1）。然后根据概率选取其中一个主题，比如“语篇结构”。在

“语篇结构”这一主题下会有若干与之有关的单词，也具有不同的概率分布，如{“move” 0.03, “CARS” 0.01, “step” 0.02……}。根据词语的概率，可以选择一个单词，比如step，这样便在文档中生成了一个单词。之后需要不断重复“选择主题——选择单词”这一过程，直至生成文档中的所有单词。值得一提的是，语料库中的其他文档可能只包含“语言特征”和“学科差异”两个主题，概率分布可能是{“语言特征” 0.7, “学科差异” 0.3}。换言之，“语篇结构”“语言特征”“学科差异”“作者群体差异”是语料库中所有文档共享的主题，但是在不同的文档中，这四个主题所占的比例或概率是不同的。

在LDA模型中，只有文档中的单词是可观测变量，而整个语料库的主题、每个文档中的主题分布以及每个主题中的单词分布均是隐藏结构，这也是LDA名称的由来（Blei 2012: 79）。LDA主题模型的任务是根据可观测的单词去推断隐含的主题结构，也就是文档生成过程的逆过程。在上述例子中我们可以看到，在生成文档中的单词时，并没有关注其顺序。因此，LDA的算法基于词袋（bag-of-words）模型，将文档视为高维空间内的词频向量，而忽略单词在文中出现的顺序（Blei & Lafferty 2007）。

如前所述，Blei *et al.*（2003）在LDA模型中只给主题分布加了狄利克雷分布作为先验分布。随后Griffiths & Steyvers（2004）又在单词分布上增加了狄利克雷分布作为先验分布，即为pLSI中的 $P(w|z)$ 加了先验分布，最终形成了如今普遍使用的LDA主题模型。

2.2 LDA模型的主题数目设置

实践中，LDA主题建模一般由程序语言中的第三方开源工具包实现，如Python中的第三方库Gensim和scikit-learn，以及R中的mallet程序包。

但在具体操作中，主题数目需要提前设置，并且不同的主题数会影响最终的分析结果：主题数目设置太少会把语义不相关的词汇合并到同一主题中，而主题数太多则会把语义相似的词汇分散到不同主题中（何琳等 2020）。理想的状态是文档中的单词出现在尽可能少的主题中，而每个主题包含尽可能少的单词（Törnberg & Törnberg 2016a；刘文字、胡颖 2020）。但是实际研究中有时还需要借助研究人员的经验和对语料的了解，反复设置不同数量的主题进行比较与权衡，以确定最佳主题数目。换言之，在批评话语分析中，最佳主题数目并不一定是统计学意义上的最佳，而是取决于主题建模及其数目能否回答研究问题或者实现研究目的，并且如果语料库中的文本体裁一致且话题统一，则可以选择较少的主题数目（Jacobs & Tschötschel 2019）。

3 北京冬奥会话语研究

3.1 研究语料

本研究采用Factiva新闻及商业数据库,以Beijing 2022 Winter Olympics、2022 Winter Olympics和Beijing Olympic Winter Games为检索词,收集了自北京冬奥会申办成功以来,即2015年7月31日至2021年9月1日之间来自境外媒体的英语新闻报道。为保证研究语料的相关性,我们将Beijing、2022和Olympic Winter Games出现次数均小于2的新闻报道删除。经过清理及统计,最终获得英文新闻报道484篇,其来源包括《纽约时报》《泰晤士报》《俄罗斯卫星报》和路透社、美联社、美国有线电视新闻网、法新社以及《南华早报》等多家境外媒体。这些语篇组成了北京冬奥会新闻报道语料库(简称北京冬奥会语料库),总形符数为328,474,总类符数为13,224。

3.2 研究步骤

本研究按照以下4个步骤进行分析。

首先,清理语料。去除停用词,并对语料库其余形符进行词形还原。

其次,运用Python中的scikit-learn程序包,基于LDA模型对北京冬奥会语料库进行主题挖掘。经过多次调试,我们最终将主题数目确定为3,每个主题下的关键词数目为15,此时获得的主题较为明晰且具有独特性和代表性,效果较好。

再次,选取每个主题中概率权重和频数均相对较高的单词作为节点词,运用AntConc 4.0.4 (Anthony 2022)统计其高频搭配词或共现型式,通过观察其扩展语境,确定语义趋向和语义韵。此外,LDA主题分析中的关键词为词元(lemma)。而在搭配分析中,由于同一词元不同词形的搭配词、用法以及意义存在差别(Sinclair 2004),因此搭配分析中的节点词以词形为基础进行分析。

最后,结合节点词的搭配词、语义韵和词语索引,分析北京冬奥会的话语建构和媒体的态度意义。其中,语义韵的抽象层级采用Sinclair(2004)提出的细微颗粒度法。

4 研究结果与讨论

4.1 北京冬奥会语料库主题分布

表1呈现了北京冬奥会语料库LDA主题建模的结果,共3个主题,每个主题由15个关键词组成。每个关键词后的数据代表关键词在此主题下的概率权重,关键词根据概率权重降序排序。换言之,括号内的数值越大,则关键词在此主题下

出现的概率越大。

表1 “北京冬奥会语料库”主题建模结果

序号	主题	关键词及其概率分布
1	冬奥会 主办权	Beijing (1,625.318,4), games (1,579.402,6), winter (1,434.554,9), Olympic (1,405.580,4), Olympics (1,161.076,8), IOC (1,074.191,1), host (1,015.028,5), China (963.897,8), say (892.576,2), city (753.912), committee (659.306,2), right (648.495), international (542.907,8), summer (532.593,2), bid (505.163,8)
2	赛事准 备工作	Beijing (614.157,2), snow (503.636,6), event (482.208,9), ski (460.696,7), China (400.539,1), world (353.496,9), winter (332.629,4), Chinese (323.562,2), Olympics (308.257,7), venue (302.592,4), year (285.707,4), sport (255.626,6), say (248.445,1), ice (247.754,2), skiing (243.505,6)
3	政府举措	China (914.5631), say (909.9786), year (502.1264), sport (492.9639), Beijing (459.5244), Chinese (387.2854), game (370.726), new (311.5527), country (309.2078), people (298.9593), million (297.2854), go (292.8931), player (278.9928), billion (277.3076), high (245.9475)

根据表1，我们可以看到，境外媒体对于北京冬奥会的报道主题可以概括为以下3类。（1）冬奥会举办权，如Beijing、host、bid、Games、winter等词。值得一提的是，此主题下的关键词如summer，也体现了境外媒体对于北京成为史上第一个“双奥之城”的关注。（2）赛事准备工作，如snow、event、ski、venue等。（3）政府举措，如sport、million、billion等，报道了政府的资金支持。

LDA主题建模能够避免人为干预，自动实现主题聚类，在宏观层面呈现北京冬奥会语料库的有关主题。接下来，我们需要进一步观察关键词的高频搭配词和共现语境，以探究媒体的话语建构及态度立场。

4.2 主题关键词的扩展意义单位分析

我们在每个主题下各选取一个概率权重和频数均较高的关键词，分别为host、snow和sport。经过统计每个词元相对应的不同词形频数，我们选取每个词元中词形频数最高的作为节点词进行扩展意义单位分析，即host（动词）和snow（名词）。

4.2.1 host

统计结果发现，词元HOST在北京冬奥会语料库中共出现了1168次，其中作为名词出现了344次，动词802次。作为动词时，动词原形host在所有词形中出现

频数最高,共444次,占有动词词形总频数的55.36%。此时,host主要与名词搭配,组成型式“host+N”。

我们统计了与型式“host+N”共现5次以上的高频搭配词或者词组,并总结了其语义趋向,具体见表2。

表2 型式“host+N”的语义趋向及高频搭配词(组)

序号	语义趋向	频数	高频搭配词(组)
1	2022冬奥会	209	the 2022 Winter Olympics/Olympic Games/Games (111), the 2022 Olympics/Olympic Winter Games (27), the Winter Games/Olympics (26), the Games (31), the world's greatest sporting events/ most prestigious sports event (6), another games/Olympics (8)
2	夏季和冬季奥运会(“双奥”)	31	both summer and winter Olympic Games/Olympics (21), both a summer and a winter Olympics/both versions of the games (10)
3	冬奥赛事	18	indoor events (18)

通过表2我们可以看到,型式“host+N”的高频搭配词根据语义趋向大致可分为三类。第一类是表示2022冬奥会的搭配词或词组,其中共现次数最多的为the 2022 Winter Olympics/Olympic Games/Games,出现了111次,占动词host频数的三分之一。其次为the 2022 Olympics/Olympic Winter Games(27次)。境外媒体也倾向于使用简化的the Winter Games/Olympics(26次)和the Games(31次)指代2022年奥运会。与前述词组共现时,境外媒体主要报道北京获得了2022年冬奥会主办权,如例(1)所示。

(1) The International Olympic Committee has selected Beijing to **host** the 2022 Winter Olympics.

例(1)报道了北京被选为2022冬奥会举办城市这一事实。经观察“host+N_{冬奥会}”的词语索引发现,该型式大多客观报道了这一事实,并未发现较为明显的语义韵趋势。

与“host+N”频繁共现的另一类语义趋向为表示夏季和冬季奥运会的词语序列,如 both (a) summer and winter Olympic Games/Olympics(21次)、another games(19次)。当与这类词组共现时,媒体报道集中在北京将成为历史上第一个举办过

夏季和冬季奥运会的国家。表3是随机抽取的5行词语索引。

表3 “host+N_{双奥}”的词语索引

1	Chinese capital becomes the <u>first city</u> in the world to	host	both summer and winter Olympic games
2	this”, he added. Beijing thus becomes the <u>first city</u> to	host	both summer and winter Olympic Games.
3	The capital will be the <u>only city</u> in sporting history to	host	both summer and winter Olympics. Today,
4	...The <u>first time</u> in Olympic history that a city will	host	both a summer and a winter Olympic years. In
5	...up the Chinese capital to become the <u>first city</u> to	host	both versions of the Games in more than a cent

型式“host+N_{双奥}”的词语索引显示，当host与both summer and winter Olympic games等词组搭配时，Beijing常被描述为the first city in the world（索引1），the only city in sporting history（索引3）以及the first time in Olympic history（索引4）等。可见，境外媒体不仅关注北京取得了2022年奥运会主办权，更对这件事情带来的影响进行了评价，将北京视为历史上获得夏季和冬季奥运会主办权的第一城，刻画了北京迄今为止首座“双奥之城”的形象。从语义韵角度而言，型式“host+N_{双奥}”表达了“赞许/肯定”的语义韵。

表示“冬奥赛事”语义趋向的高频搭配词组仅有indoor events一个，与“host+N”共现了18次，主要描述举办某些冬奥会赛事的场馆，态度较为客观。

简言之，动词host的高频搭配词（组）和共现语境表明，北京冬奥会相关的境外媒体报道的焦点之一为北京取得了冬奥会主办权，并成为史上首个举办夏季和冬季奥运会的国家。另外，型式“host+N”在描述北京成为“双奥”之城时呈现出“赞许/肯定”的语义韵趋势。

4.2.2 snow

词元SNOW在北京冬奥会语料库中的频数是504，其中名词snow最为高频，共出现了425次。名词snow倾向于与形容词搭配，构成型式“ADJ+snow”，其频数为213，占名词snow频数的50.12%。与“ADJ+snow”共现的高频形容词为natural（63次）、artificial（52次）、man-made（24次）、real（13次）、fake（10次）。这些高频搭配词都表达了“造雪方式”这一语义趋向，聚焦于冬奥会雪上项目中雪的制造方式，即自然雪（natural/real snow）和人造雪（artificial/man-made/fake snow）。

观察词语序列natural snow（63例）和real snow（13例）的扩展语境，我们

发现, 共有58例与北京冬奥会有关。扩展语境中与 *natural snow* 和 *real snow* 共现次数较高的词或词组有 *lack of* (27次)、*little* (9次)、*hard-pressed* (3次), 以及其他表达否定含义的词汇3例, 共计42例, 约占所有与北京冬奥会相关实例的72.41%。这些实例指出, 北京作为冬奥会的主办城市, 缺乏雪上项目所需的自然雪, 表达了“缺乏/不足”的语义韵。具体如例(2)和例(3)所示。

(2) (a) The Chinese capital has been picked to host the 2022 Winter Olympics, (b) despite the fact that it has little **natural snow**.

(3) Beijing had been considered the overwhelming favorite but was criticized for a lack of **natural snow**.

例(2)中, 小句(a)报道了中国首都即北京将举办2022年冬季奥运会, 但小句(b)作为转折, 认为北京几乎没有自然降雪, 并认为这是事实(fact)。相似地, 例(3)首先指出北京以压倒性优势获得了主办权, 但“被批评缺乏自然降雪”。

Artificial/man-made/fake snow 作为节点词组时, 共有86例, 其中76例报道了北京冬奥会雪上项目准备工作。其扩展语境中的高频共现词组有 *rely on* (28次) 和 *reliance on* (9次) 等, 表示北京冬奥会的雪上项目将依赖人工造雪。此外, 在这些例证中, 有27例与 *heavily* (13次)、*completely* (6次)、*entirely* (3次)、*totally* (3次) 等增强语共现, 以强化北京冬奥会依赖人工降雪的程度和印象, 具体见表4.5条词语索引。

表4 *rely/reliance on artificial/man-made/fake snow* 的词语索引

1	means they have to rely <u>completely</u> on	artificial snow	. Distance between Beijing and mountain
2	and the need to rely almost <u>entirely</u> on	artificial snow	. China's Sports Minister Liu Peng
3	it will have to rely almost <u>totally</u> on	artificial snow	. Q: So why did the IOC
4	ues in China will be <u>heavily</u> reliant on	fake snow	Beijing bid leaders insisted they have
5	China's mountain venues rely <u>heavily</u> on	man-made snow	, which was considered one of the bid

表4词语索引表明, 境外媒体在北京冬奥会相关报道中, 有意运用增强语刻画北京依赖人工造雪以满足冰雪项目的要求, 暗示了北京在主办冬奥会上并没有优势, 表达了较为明显的消极态度意义。因此, *artificial/man-made/fake snow* 在语境中构筑了显性的“非真实”语义韵。前人研究也发现, 西方主流媒体如《纽约时报》、BBC等, 在报道北京冬奥会的体育事件中, 质疑北京使用人工造雪的行为

(刘静轩等 2022)。

概言之, snow 的高频搭配词或词语序列倾向于表达消极的语义韵, 即“缺乏/不足”和“非真实”。这表明, 境外媒体试图以北京的降雪条件和自然气候为借口, 批评北京举办冬奥会的不足, 质疑北京是否真正有资格获取冬奥会主办权。

4.2.3 sports

词元SPORT在北京冬奥会语料库中共出现了1,006次, 其中名词复数形式sports出现了853次, 占总频数的84.79%。经观察, 词语序列winter sports的频数为326次, 占sports频数的38.22%, 远超其他词语序列。因此, 我们将以winter sports作为节点词组, 考察其共现语境。

winter sports在北京冬奥会语料库中的搭配词可以分为三个语义组。第一, 精神文化类(46例), 高频搭配词有tradition、destination、culture。第二, 经济活动类(42例), 高频搭配词有market、industry、development; 第三, 推广、发展等行为(78例), 高频搭配词有promote、grow、develop、participate、popularize。

观察词语索引可见, 精神文化类搭配词所在的词语索引中, winter sports经常与表示不足或否定的词汇共现(如lack of、far from、have no等), 借以批评北京缺乏冬季运动的传统或文化。我们抽取了5条相关的词语索引(见表5)。

表5 winter sports 与精神文化类搭配词的词语索引

1	<u>sacrificing</u> some ... of the atmosphere of a	winter sports destination	. Beijing <u>is unlikely to be</u> blanketed
2	. After all, China <u>doesn't have</u> much of a	winter sports tradition	– it won its first Olympic gold
3	home. While China is <u>far from</u> being a	winter sports power	, Beijing, along with the neighbori
4	to its lack of snow, Beijing <u>has no</u> real	winter sports culture	, two things Almaty has in abundance
5	with plenty of resources but virtually <u>zero</u>	winter sports history	or send them to Central Asia for

表5显示, 外媒对于北京举办冬奥会的质疑在于中国或北京没有冬季运动文化传统, 如China doesn't have much of a winter sports tradition(索引2), Beijing has no real winter sports culture(索引4)或zero winter sports history(索引5)。由此, 型式“winter sports N_{精神文化}”在语境中构筑了“不足/缺乏”的语义韵。在46例词语索引中, 有38例(82.6%)表达了消极语义韵, 体现了境外媒体对北京冬

季运动的普及程度持批评态度。

当 *winter sports* 与表示经济活动/组织和发展、推广类的词语搭配时，通常表示中国政府大力发展冰雪产业、拓展冰雪市场的措施和决心，以及北京冬奥会对中国发展冰雪运动的促进作用。我们随机选取推动、发展类搭配词的5行词语索引，显示如表6所示。

表6 *winter sports* 与推动、发展类搭配词的词语索引

1	The ambition of the government (China) <u>to develop</u>	winter sports	is going to create a huge momentum
2	Beijing would use the games <u>to encourage interest in</u>	winter sports	and boost tourism in in a region that
3	further enhance the tremendous potential <u>to grow</u>	winter sports	in our country, in Asia, and around
4	Games is our desire <u>to popularize and develop</u>	winter sports	through hosting the games,” he
5	month announced a \$30 million program <u>to promote</u>	winter sports	such as luge, bobsledding and

表6展示了中国政府对于冰雪运动的重视，不仅有强烈的抱负（*ambition*）和愿望（*desire*）发展冰雪运动，更有实际的行动和措施切实普及冰雪运动，如资金投入和发动群众。这都是中国政府为了2022年北京冬奥会所做出的积极努力。可见，当 *winter sports* 与经济活动类和推广、发展类词语共现时，表达的态度意义也较为积极，展现了“努力推动”的语义韵。

整体而言，境外媒体提到北京冬奥会有关 *winter sports* 的主题时，主要倾向于表达两种态度意义。一种是消极态度，认为北京乃至中国缺乏冰雪运动的传统和群众基础。另一种是积极态度，报道了中国政府为发展冰雪运动产业、推广冰雪运动所采取的措施，描述了中国政府积极行动和努力的一面。

5 结论

本研究采用LDA主题建模，分析了境外媒体有关北京冬奥会英文报道的主题以及基于主题关键词所体现的北京冬奥会话语意义建构。LDA主题建模结果显示，北京冬奥会英文报道聚焦冬奥会主办权、赛事准备工作以及政府举措三大主题。通过深入观察各主题中关键词的扩展语境并统计搭配词，我们发现，各个主题的话语建构呈现出多种不同侧面的态度意义。一方面，境外媒体肯定了北京作为历

史上首个获得夏季和冬季奥运会主办权的“双奥之城”的地位。同时，境外媒体也对中国政府推广冰雪运动、发展冰雪产业的决心以及所采取的措施表达了较为积极的态度。另一方面，对于北京冬奥会的赛事场馆和项目准备工作，境外媒体批评了北京缺乏自然降雪的气候条件，需要依赖人工造雪，并且认为北京缺少冰雪运动的文化底蕴和传统，借此质疑北京冬奥会的举办资格。研究结果显示，境外媒体对有关北京冬奥会赛事运动本身的话题较为关注，同时也能客观地肯定中国政府所做出的积极举措与准备工作。这表明我国对于北京冬奥会的对外叙事传播可聚焦到微观层面的冰雪运动和赛事本身。

研究方法方面，本研究显示，在批评话语分析中使用LDA主题建模不仅能够快速获取大规模语料库的主题，而且避免了参照语料库的干扰，研究结果较为客观、可靠。因此，LDA模型在一定程度上克服了传统批评话语分析中的数据任意性、分析主观性等缺陷，适用于从宏观层面对大规模语料库进行主题分析。在此基础上，深入观察主题关键词的共现语境，统计其搭配词、语义趋向和语义韵，对语料进行细微颗粒度分析，能够在微观层面进一步揭示主题的话语建构意义和态度立场。未来的研究可以结合LDA主题建模、语料库方法以及批评话语分析的更多研究路径（如话语-历史分析路径）对相关媒体报道话语进行分析。

注释

- 1 其中涉及较为复杂的数学知识，在此不赘述，有兴趣的读者可以阅读Blei *et al.*（2003）有关这一问题的详细阐释。

参考文献

- ANTHONY L. AntConc (4.0.4) [CP/OL]. 2022. <https://www.laurenceanthony.net/software>.
- ARANDA A, SELE K, ETCHANCHU H, et al. From big data to rich theory: integrating critical discourse analysis with structural topic modeling [J]. *European Management Review*, 2021, (18): 197-214.
- BAKER P, GABRIELATOS C, KHOSRAV INIK M, et al. A useful methodological synergy? Combining critical discourse analysis and corpus linguistics to examine discourses of refugees and asylum seekers in the UK press [J]. *Discourse and Society*, 2008, 19(3): 273-306.
- BLEI D. Probabilistic Topic Models [J]. *Communications of the ACM*, 2012, 55(4): 77-84.
- BLEI D, LAFFERTY J. A correlated topic model of science [J]. *The Annals of Applied Statistics*, 2007, 1(1): 17-35.
- BLEI D, NG A, JORDAN M. Latent dirichlet allocation [J]. *Journal of Machine Learning Research*, 2003, 3(1): 993-1022.
- ENGSTRÖM R, PARADIS C. The in-group and out-groups of the British National Party

- and the UK Independence Party: a corpus-based discourse-historical analysis [J]. *Journal of Language and Politics*, 2015, 14(4): 501-527.
- GRIFFITHS T, STEYVERS M. Finding scientific topics [J]. *Proceedings of the National Academy of Sciences of the United States of America*, 2004, 101(suppl 1): 5228-5235.
- HOFMANN J. Probabilistic latent semantic indexing [J]. *Proceedings of the Twenty-second Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 1999: 50-57.
- JACOBS T, TSCHÖTSCHEL R. Topic models meet discourse analysis: a quantitative tool for a qualitative approach [J]. *International Journal of Social Research Methodology*, 2019, 22(5): 469-485.
- JO W. Possibility of discourse analysis using topic modeling [J]. *Journal of Asian Sociology*, 2019, 48(3): 321-342.
- MOHR J, BOGDANOV, P. Introduction — Topic models: What they are and why they matter [J]. *Poetics*, 2013, 41(6): 545-569.
- SINCLAIR, J. *Trust the Text* [M]. London: Routledge, 2004.
- SONG Y, LEE C-C, HUANG Z. The news prism of nationalism versus globalism: how does the US, UK and Chinese elite press cover ‘China’s rise’? [J]. *Journalism*, 2021, 22(8): 2071-2090.
- TÖRNBERG A, TÖRNBERG P. Combining CDA and topic modeling: analyzing discursive connections between Islamophobia and anti-feminism on an online forum [J]. *Discourse & Society*, 2016a, 27(4): 401-422.
- TÖRNBERG A, TÖRNBERG P. Muslims in social media discourse: combining topic modeling and critical discourse analysis [J]. *Discourse, Context & Media*, 2016b, 13(Part B): 132-142.
- 何琳, 乔粤, 刘雪琪. 春秋时期社会发展的主题挖掘与演变分析——以《左传》为例[J]. *图书情报工作*, 2020 (7): 30-38.
- 刘静轩, 张子轩, 于杰, 等. 多元与偏见: 西方媒体北京冬奥会报道中的中国国家形象话语表征[J]. *武汉体育学院学报*, 2022 (3): 23-29.
- 刘文字, 胡颖. 基于文本挖掘的非传统文本批评话语研究[J]. *天津外国语大学学报*, 2020 (4): 29-41.
- 杨敏, 符小丽. 基于语料库的“历史语篇分析”(DHA)的过程与价值——以美国主流媒体对希拉里邮件门的话语建构为例[J]. *外国语*, 2018 (2): 77-85.
- 赵永刚. 媒体并购话语中的中国企业形象对比研究——一项语料库辅助的话语-历史分析[J]. *解放军外国语学院学报*, 2021 (1): 62-70.

通信地址: 100191 北京市 北京航空航天大学外国语学院

按 语

《语料库语言学》创编十年，刊物渐入正途。

自本期起，本人任主编期间，将不在《语料库语言学》上刊发文章。

许家金

2023年7月

英语“致使-动结”构式 宾补共选机制研究*

中国海洋大学 刘聪颖 李潇辰 中国科学院大学 曹笃鑫

提要：本文以构式“make+宾语+形容词补语”为例，采用构式搭配分析法研究英语“致使-动结”构式中宾语和补语的共选机制。研究发现，在英国国家语料库中，该构式的宾语、补语吸引的词汇具有鲜明的语法和语义特征，同时宾语和补语的词汇间存在共选关系，形成一系列高频搭配。上述共选模式是构式的语法、语义属性及其语用功能互动的产物。这些发现表明，构式各位置的词汇选择具有层次性、概率性，是构式内外部语法、语义和语用因素博弈的结果。

关键词：致使-动结构式、构式搭配分析法、共选

1 引言

英语动结式是英语论元结构构式的一种基本类型，句法结构通常表现为[SUB V(OBJ)XP]，意为“X致使Y变成Z”。其中动词V可分为致使与非致使两类，本文将分别称为致使类动结式与非致使类动结式，XP则可为形容词短语、介词短语或名词短语等，以形容词短语居多。Goldberg(1995:193)及董成如(2014)等研究虽指出了致使类动结式的宾语和补语存在特殊共选关系，却未能系统揭示其机理。本文基于构式搭配分析法探索致使类动结式的宾语和补语共选机制，以期深化对构式内部各位置间共选机制的研究。

2 文献回顾

英语动结式的研究可分为宏观与微观两个视角。前者关注构式整体的归类问题及表意机理，后者则关注构式内部各位置的词汇共选机制。

* 本文为中国海洋大学本科教育教学研究一般项目“海大英语专业学生笔语中介语的构式搭配偏误及对策研究”(2021JY134)的阶段性成果。

李潇辰为本文通讯作者。

刘聪颖：数据收集、数据分析、初稿撰写、字数占比(70%)；

李潇辰：选题构思、研究方法、字数占比(30%)；

曹笃鑫：讨论结论、修改润色。

2.1 动结式研究的宏观视角

Goldberg (1995: 78) 认为动结式是致使-移动构式的扩展隐喻, Boas (2003) 则认为动结式中的结果短语可表示状态和位置变化, 反将后者纳入前者范畴。董成如 (2014) 认为动结式是独立结构, 而非衍生于其他语言结构。随着动结式研究的深入, 其相关类型也在不断扩展。例如, 基于形式, 动结式可分为一/二元, 甚至三元动结式 (Boas 2003; 殷红伶 2011); 根据结果补语类型, 动结式可分为方位动结式 (多为介词性短语) 和性状动结式 (多为形容词性短语) (Wechsler 2001; Goldberg & Jackendoff 2004; 殷红伶 2011); 根据宾语类型, 动结式可分为由动词或构式提供宾语的动结式及假宾语动结式等 (Goldberg & Jackendoff 2004; 殷红伶 2010)。

学界对动结式的表意机理存在争议。传统解释主要有三种。第一种是基于生成语法的解释 (Simpson 1983), 其基本假设为表示状态变化或位置移动的结果短语指向句中 (潜在性) 直接宾语, 称 DOR 限制 (direct object restriction)。第二种解释基于事件结构的词汇映射模式, 将动结式分为动词次事件和结果次事件, 二者具有时间上的依存性和共延性。第三种解释则基于构式语法 (Goldberg 1995: 78), 主张将动结式视为致使-移动构式的隐喻用法。近年来, 新的理论阐释也不断涌现。如王寅 (2009) 提出体验性事件分析法, 指出言语交际需要也可促使语言演化出相应构式, 不必仅依赖句法和隐喻机制。

2.2 动结式研究的微观视角

本类研究多关注动结式中动词与结果补语位置词汇的准入条件、论元表达及动结式的共现限制问题。Goldberg (1995: 193) 探讨了动词位置的准入条件, 发现除致使动结式外, 其余动结式中动词编码的动作必须被解读为直接造成状态变化, 不能存在时间间隔。

不少学者研究探讨了动结式补语的词汇准入条件 (Wechsler 2001; Broccias 2004; 罗思明等 2010)。Wechsler (2001) 基于事件-论元同构体态模式提出能进入英语 Control 结果构式和 ECM (Exceptional case-marking) 结果构式的形容词具有等级性和有界性等特点。罗思明等 (2010) 讨论了汉语动结式补语成分的准入条件, 从语料库和类型学角度补充了 Wechsler 的结论。辛志英、单健 (2019) 则从认知角度将结果属性视为及物性系统中过程融合的结果, 提出结果属性系统的准入条件并将结果属性分为七个子类。上述研究均证实, 能进入补语位置的一般为非派生的等级形容词, 经常编码瞬发或易变的属性或状态。

Goldberg (1995) 提出了一些普遍适用的共现限制, 如二元动结式必须有一个有生 ([+Animate]) “发动者” 论元 (instigator argument), 动词表示的动作必须

被解读为直接造成状态变化,不能存在时间间隔,形容词结果短语则必须表示一个阶段的终点而且结果短语的中心词不能为派生形容词。Wechsler (2001) 发现只有最大终点封闭等级形容词 (maximum end-point closed-scale adjectives) 才能进入持续动词构成的 Control 结果构式,而最小终点封闭等级形容词 (minimum end-point closed-scale adjectives) 则不可。此外,学界普遍赞同结果短语不能与静态动词一起出现。

综上,既往研究从宏观与微观角度探讨了动结式的整体特征及其内部各位置的准入条件。其中微观角度的研究争议较多。问题的根源在于研究方法的局限。第一,虽然 Goldberg & Jackendoff (2004) 及董成如 (2014) 等已发现以 make、drive 等致使义动词为核心的动结式具有特殊的论元结构及准入条件,但前人研究未重视动结构式各子类的异质性。第二,前人研究多聚焦于句法语义层面的准入条件,忽视了语用(功能)层面。第三,前人研究大多基于内省或原始频数,缺乏有效方法量化构式各位置对词语的吸引程度,难以呈现各位置上的词语分布特征。为弥补上述缺憾,本研究聚焦以 make 为核心的动结式,借助构式搭配分析法揭示宾语与补语位置的词汇分布特征,分析二者在句法、语义和语用维度上的准入条件。

3 研究方法

本文旨在探讨“make+宾语+形容词补语”构式(简称目标构式)内部的词汇准入机制。具体而言,本文旨在回答下列3个研究问题。第一,目标构式的宾语和补语位置各自吸引哪些词汇?第二,占据宾语和补语位置的词汇呈现哪些共选模式?第三,目标构式的整体功能如何影响其宾语和补语位置的词汇偏好及共选模式?

本文采取构式搭配分析法(Stefanowitsch & Gries 2003; 房印杰 2018)回答上述问题。该方法通过比较搭配词素与构式间的构式搭配强度(collostructional strength)反映构式意义的认知语义聚类(胡健、张佳易 2012)。本研究用到该方法下辖的共现词素分析法(collexeme analysis)和互为变化的共现词素分析法(covarying collexeme analysis)。前者可呈现词汇在特定构式单一位置上的分布特征,从而揭示该位置的词汇准入模式;后者则可呈现同一构式内部两个位置的词汇共选关系,从而揭示二者词汇偏好的相互影响(Hilpert 2014)。

本文数据来自英国国家语料库(British National Corpus, 简称BNC)。该库库容约1亿词(111,978,070词),收录语料涵盖当代英式英语的各类主要文体,能较为全面地呈现当代英式英语口语和书面语的整体情况。本文使用的是经词形还原和词性赋码的BNC语料。

本研究采用自动与人工分析相结合的方法分析数据，具体步骤如下。

第一，语料预处理及目标构式提取。通过自制程序提取BNC中所有目标构式，记录构式总频数、在宾语和补语位置出现的词及其频数等信息。

第二，目标构式搭配分析。调整上一步所得数据格式，配合脚本 Coll.analysis（Gries 2022）和R进行共现词素分析及互为变化的共现词素分析，筛选具有统计学显著意义（ $p<0.001$ ）的词语以备后续分析。本步操作所得数据可揭示目标构式宾语和补语位置各自吸引的词汇及二者的搭配模式。

第三，分析目标构式宾语和补语吸引词汇的语法及语义特征，理清两个位置的词汇偏好及其背后的词汇准入条件。

第四，分析目标构式宾语和补语的词汇共选模式，从语法和语义角度探究两个位置的词汇偏好及准入条件的相互影响。

第五，遴选典型搭配进行索引行分析，从语用角度探讨构式的整体功能如何影响其宾语和补语位置的词汇偏好及词汇准入条件。

4 数据分析

目标构式在BNC中共出现10,205次。本节将分别分析目标构式宾语和补语各自的词汇偏好，进而分析两个位置上的词汇共选关系。下文数据均为词形还原后的形式。

4.1 宾语和补语的词汇偏好

表1呈现了在共现词素分析法下宾语和补语的词汇偏好结果。通过表格信息可知，两个位置吸引的词汇均呈现鲜明的语法及语义特征。

表1 目标构式宾、补语位置的词汇

宾语位置		补语位置	
共现词素	搭配强度	共现词素	搭配强度
it	36,378.834,13	clear	9,205.359,355
them	2,346.768,242	easy	5,547.649,413
matter	1,037.548,758	difficult	3,643.030,254
thing	942.349,472	possible	2,968.349,074
life	767.817,315	impossible	2,085.725,892

（待续）

(续表)

宾语位置		补语位置	
共现词素	搭配强度	共现词素	搭配强度
me	748.901,353	bad	1,596.672,049
himself	714.347,116	sick	743.082,29
him	696.566,067	aware	574.199,326
you	632.414,648	happy	528.923,598
yourself	545.151,619	available	515.353,068
themselves	479.177,961	angry	483.133,317
myself	445.473,409	plain	364.337,988
herself	422.988,469	redundant	324.846,904
people	173.147,441	comfortable	296.929,252
us	169.770,106	worthwhile	251.798,851
everything	130.472,085	compulsory	237.699,1
itself	125.251,3	nervous	231.606,938
ourselves	72.145,377	vulnerable	224.996,716
something	66.008,558	ill	190.200,82
yourselves	55.367,055	uneasy	183.373,136

表1中位于宾语位置的词语可归为名词与代词两大类,每类下辖数个语义场(图1)。除thing、matter等少数泛指词外,宾语位置倾向于吸引指代金钱、人及地点等具体事物的名词和代词。这种分布规律与构式整体语义有关。动结式编码的事件中,宾语对应事件经历者,其原始状态因经历事件而发生改变。现实生活中,具体的人、事、物更容易发生状态变化,因而更容易充当宾语。

目标构式的补语倾向于吸引表示事物性质或特征的等级形容词(见图2)。表明晰性、难易度、可及性及情绪状态等可变属性的子类(Dixon 2005: 84)更频繁地出现在补语位置,而表尺寸、年龄和颜色等事物固有属性的形容词(类属形容词)则比较罕见。上述词汇偏好与构式整体功能相关。动结式旨在说明事物的状态变化而非静态特征,因此其补语更吸引表述可变属性的形容词。此外,目标构式除具有动结义,还继承了make自带的致使义(Goldberg & Jackendoff 2004)。这使得homeless和unlawful等少数类属形容词也可出现在补语位置。

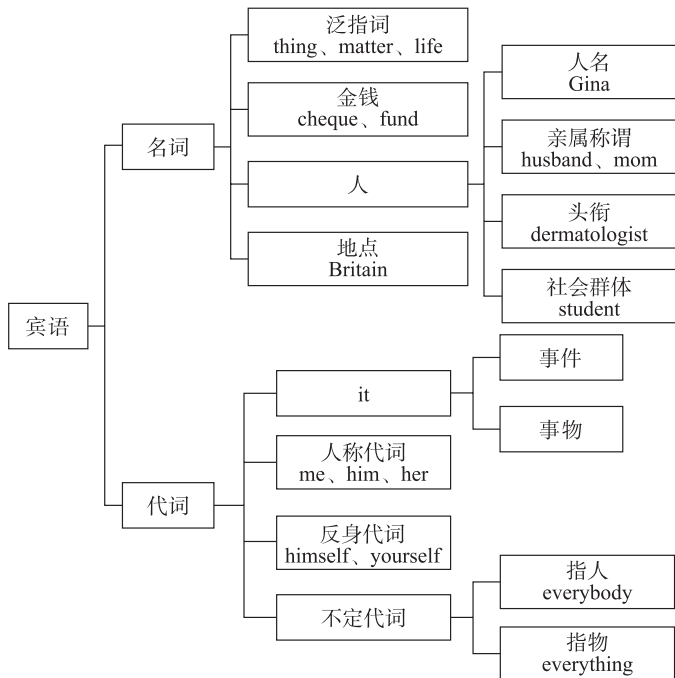


图1 宾语词汇的语法和语义属性归类

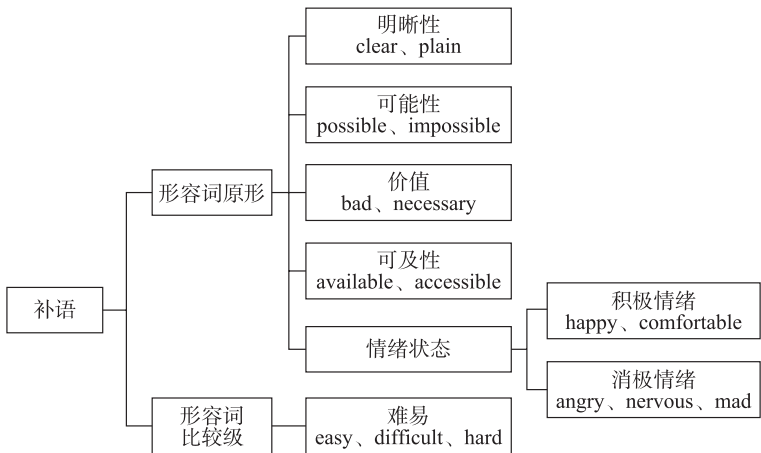


图2 补语词汇的语义归类

4.2 宾语和补语的词汇共选

表2呈现了在互为变化的共现词素分析法下的目标构式宾语和补语词汇共选情况。表中数据按构式搭配强度排列。例如，表中第一项 make it clear 构式搭配强度为无限（infinite），表明 it 与 clear 之间具有极高的吸引力。

表2 目标构式宾、补语词汇共选情况

搭配组合	搭配强度	搭配组合	搭配强度
make it clear	Inf.	make life easy	18.055,5
make matter bad	228.499,8	make life difficult	18.039,6
make it possible	98.334,3	make thing easy	17.797,6
make me sick	49.849,7	make people aware	176,498
make it easy	47.538,8	make me mad	17.368,6
make thing bad	46.146,9	make information available	17.027,6
make it impossible	44.635,6	make you sick	16.524,8
make it difficult	40.617,9	make husband jealous	16.322,1
make you happy	32.720,9	make it hard	16.211,6
make cheque payable	27.216,6	make facility available	14.455,2
make fund available	22.370,9	make people redundant	14.289,1
make yourself comfortable	20.926,8	make life miserable	13.972,7
make me angry	19.918,1	make me nervous	12.741,4
make it plain	19.111,2
make himself available	18.063,4		

由表2可知，宾语和补语词汇共选模式同宾语的语义和语法特征有关。首先可按照宾语的语义类型分为“宾语指事物”和“宾语指人”两大类。两类宾语搭配的形容词补语具有系统性差异。进而可按照宾语的语法特征进一步细分，将两大类分别切分为“名词作宾语”与“代词作宾语”的情况。

宾语指事物时，补语多为具有评估判断功能、等级性、修饰性和主观性等特点的评价形容词。指事物的代词性宾语均为it，其补语常为表示可能性、难易度、必要性和可理解性的形容词原形（见图3）。构式整体功能是评价it所指情况，其语义通常为“使某事变得可能/简单/困难”等。此时it通常指某个事件，其所指有时只能基于语境大致判断。如例（1）和（2）中构式的意思均是“把某事阐述清楚”（使某事变得清晰）。根据语境判断，例（1）中it应指其后从句“he was deeply hurt”，而例（2）中则指前一句“I fight on, I fight to win”。

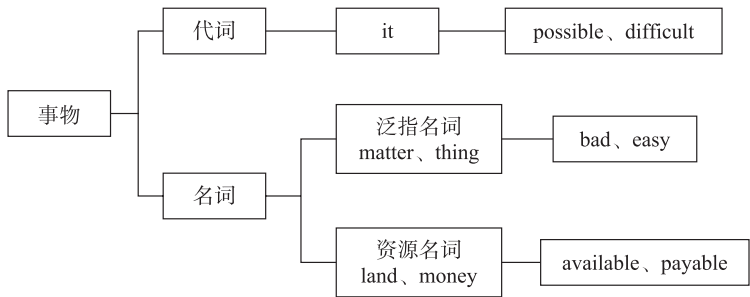


图3 宾语指事物时宾、补语的共选情况

(1) A source close to the player last night **made it clear** that he was deeply hurt and being dropped from the match.

(2) I fight on, I fight to win. Wakeham had **made it clear** when he took over from Younger as her team manager.

指事物的名词性宾语可分为两类。第一类为thing、life和matter等泛指名词。其中matter和thing常为复数形式，泛指“事态”“情况”或“局面”等。life则一般为单数形式，泛指“人生的某阶段”，常略带戏谑与夸张。如例(3)和(4)中，make matters worse用于评价“女医生让他住院”一事对整体局面的影响，而make life easier则用于评价it所指对“(我的)生活”的影响。两例中的评价均由言者而非句子主语做出。类似代词性宾语it，此类名词性宾语的补语也多为评价形容词。不同的是，此类补语常为形容词比较级，如例(3)中的worse和例(4)中的easier。构式整体功能为定性评价某因素对局面、事态与生活的影响。第二类名词性宾语是land、finance、money及cheque等指称土地、金钱或人力等资源的名词。补语位置多为available和payable等表示可及性的形容词。构式整体意为“使得(土地、金钱等资源)可用”，如例(5)和例(6)。

(3) A woman doctor, to **make matters worse**, had had him into hospital.

(4) It **makes life easier** for me, and they don't get overworked, so it's wonderful.

(5) Its main priority is to **make land available** for house building in the private sector.

(6) Please **make cheques payable** to Lifestyle Promotions and allow 14 days for delivery.

指人的宾语以代词为主（名词宾语仅有 husband 和 people）。其补语多为具有明显积极或消极情感倾向的形容词（见图4）。宾语为人称代词时，补语整体偏向消极语义韵，主语一般为非第一人称，可见是非自身原因致使宾语产生了负面情绪或进入消极状态，如例（7）。与之相反，宾语为反身代词时，补语整体偏向积极语义韵，主语一般为第一人称，即主语通过实施某种行为使自己产生了正面情绪或进入积极状态，如例（8）和例（9）。此外，此类中存在较多具有习语性的固定搭配，如例（8）中的 make oneself comfortable（“别客气”）与例（9）中的 make oneself useful（“帮忙”）等。

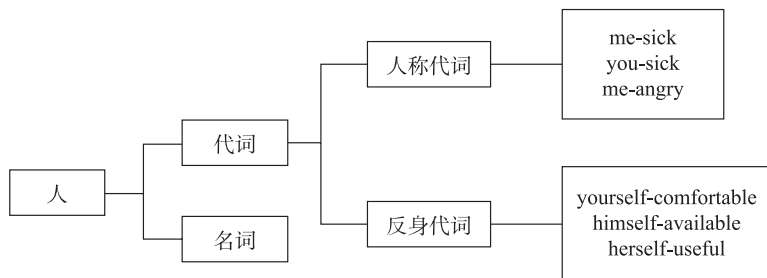


图4 宾语指人时宾、补语的共选情况

(7) Now she'd *made him angry* again.

(8) Come into the lounge and *make yourself comfortable*.

(9) Perhaps, if she *made herself useful*, he might decide she could stay —
for a while at least.

5 讨论

上节的分析发现，目标构式的宾、补语位置各自存在明显的词汇偏好，同时两者间存在相互制约的共选关系（卫乃兴 2011：50）。本节旨在分析上述发现背后的机制。

语法及语义因素可解释目标构式宾语和补语位置的部分词汇偏好。例如，语法规则要求宾语由名词或代词充当。又如，补语用于说明变化后的状态，因此更吸引品质而非类属形容词。然而单凭语法及语义因素难以解释宾、补语间的共选模式。例如，为何指物宾语特别吸引评价性补语而指人宾语则不倾向于吸引评价性补语？

我们认为，这些词汇偏好的根源在语用层面。目标构式在几类典型语境中呈现特定功能。构式功能同时影响宾语和补语的词汇选择，造成两个位置对特定语

义场的偏好。探讨词汇准入条件时需分别考察这些功能子类，理清它们与构式整体词汇准入条件的异同。语料分析显示目标构式具有3类主要功能。第一类功能是“评述某种行为的后果”，其语义结构可表示为“[某种行为]使[某件事情]产生[某种状态]”，如make it clear可解释为“主语通过其言语行为把某事表述清楚”。该子类的宾语多为指代事态的代词或名词，补语则多为表述明晰性、难易度及可能性的形容词。宾语为it时，补语多为形容词原形；宾语为life等泛指事态的名词时，补语则多为比较级。第二类功能是“评述某种资源的可及性或价值变化”，其语义结构为“[某种因素]致使[某种资源]的[可及性或价值]变得[更高或更低]”，如make money available。该子类的宾语对应“某种资源”这一变量，吸引指称相应资源的名词，补语则对应“可及性或价值”及“更高或更低”两个变量，吸引表示价值与可及性的形容词。第三类功能是“评述某人情绪状态的变化”，其语义结构为“[某种因素]致使[某人]产生[某种情感状态]”。该子类的宾语对应“某人”，一般为代词，补语对应“某种情感状态”，吸引表示情感的形容词。当补语为消极情感状态时，该构式的主语多为外在因素，而宾语多为人称代词。当补语为积极情感状态时，主语多为内在因素，宾语则多为反身代词。

语用因素与语法、语义因素的作用方式不同，呈现三方面的辩证统一关系。第一，语用因素具有主观性，而语法、语义因素具有客观性。语用因素体现在言者的自主选择中。英语中同一“致使-动结”事件可由多种方法表述，如例（10）和例（11）。言者可自主选择目标构式，进而决定其宾语和补语位置的词语。语法、语义因素则由构式及词语自身的语法和语义属性决定，独立于言者意图，因而具有客观性。第二，语用因素不具有强制性，而语义和语法因素具有强制性。英语并不强制言者选取哪一种特定方式表述致使-动结事件。但无论选择哪一种，必须遵循其附带的语法、语义规则。例如，言者可以在例（10）和例（11）间任选其一。但如果选择第一种表述方式，则make的宾语必须是名词性成分，补语则必须由形容词充当。第三，语用因素的作用方向是自顶向下的，而语义和语法因素的作用方向则是自底向上的。语用因素的制约是言者意图和交际语境对具体表述形式的影响，其影响的方向是从宏观到微观。语义和语法因素则是构式及词汇自身属性对言者意图的反向制约，其影响的方向是从微观到宏观。总之，语用、语义及语法因素使得目标构式成为有机整体。三者合力塑造了构式内部各位置的词汇偏好，协调了各位置间的共选模式，使得构式整体具有鲜明的功能特征。

（10）My failure makes me sad.

（11）I am sad because I failed.

6 结语

本研究以 *make* 为例, 使用构式搭配分析法探讨了英语致使-动结构式宾语和补语的词汇偏好及其共选模式, 进而探究其背后机制, 揭示出目标构式整体及其各典型功能子类的词汇偏好和背后的词汇准入条件。从这些发现中可归纳出如下 3 点启示。

第一, 构式的词汇准入条件具有概率性。Goldberg (1995: 193) 及 Wechsler (2001) 等研究均采用规定性的语言表述其发现, 即只有满足某条件的词才可进入某构式的某位置。本研究则发现准入条件是兼具刚性规则与柔性偏好的概率连续统, 构式的特定位置倾向于吸引具有某种特征的词汇。前人研究仅关注刚性规则却忽视柔性偏好, 对准入条件的认识存在片面性。本研究则兼顾两者, 因而更加全面均衡。

第二, 构式的词汇准入条件具有层次性。前人研究发现通常默认词汇的准入条件是恒定的, 即构式每个实例的准入条件完全一致。本研究则发现抽象构式与其具体实例的词汇准入条件不同, 其间可切分出几个层次。越贴近具体实例, 准入条件就越精细。这是因为具体构式一方面继承了抽象构式的词汇准入条件, 另一方面又反映出特定语境的制约。

第三, 构式的词汇准入条件是语法、语义及语用三层面互动的结果。前人研究发现的词汇准入条件多集中在语法-语义层面, 可表述为具有某种语法/语义特征的词语可以进入某构式的某位置。本研究则揭示出语用因素对词汇准入条件的巨大影响。

本研究仍存在一些不足之处。第一, 囿于篇幅, 本研究仅探讨了构式各部分对词汇的吸引倾向, 未探讨其对词汇的排斥倾向, 而后者也与构式的词汇准入机制密切相关。第二, 本研究的规模有限。个案研究虽能揭示构式词汇准入机制的部分规律, 但这些规律的普适性有待大规模研究的检验。未来研究可兼顾构式对词汇的吸引与排斥倾向, 并将研究视野扩展到多种语言中的多类典型构式。

参考文献

- BOAS H. A constructional approach to resultatives [M]. Stanford: CSLI Publications, 2003.
- BROCCIAS C. The cognitive basis of adjectival and adverbial resultative constructions [J]. *Annual Review of Cognitive Linguistics*, 2004, 2(1): 103-126.
- DIXON R. A semantic approach to English grammar [M]. Oxford: Oxford University Press, 2005.
- GOLDBERG A. Construction: a construction grammar approach to argument structure

- [M]. Chicago: The University of Chicago Press, 1995.
- GOLDBERG A, JACKENDOFF R. The English resultative as a family of constructions [J]. *Language*, 2004, 80(3): 532-568.
- GRIES S. Coll. Analysis 3.5. A script for R to compute perform collostructional analyses [EB/OL]. (2022-08-21) [2023-03-21]. <https://stgries.info/teaching/groningen/coll.analysis.r>.
- HILPERT M. Collostructional analysis: measuring associations between constructions and lexical elements [C]//GLYNN D, ROBINSON J. Polysemy and synonymy: corpus methods and applications in cognitive linguistics. Amsterdam: John Benjamins, 2014: 391-404.
- SIMPSON J. Resultatives [C]//LEVIN B, HOVAV M, ZAENEN A. Papers in lexical-functional grammar. Bloomington: Indiana University Linguistics Club, 1983.
- STEFANOWITSCH A, GRIES S. Collostructions: Investigating the interaction between words and constructions [J]. *International Journal of Corpus Linguistics*, 2003, 8(2): 209-243.
- WECHSLER S. An analysis of English resultatives under the event-argument homomorphism model of telicity [R]. Presented at the 3rd Workshop on Text Structure, University of Texas, Austin, USA, 2001.
- 董成如. 认知语法框架下动结式的形成和论元实现[J]. *现代外语*, 2014 (5): 608-617.
- 房印杰. 搭配构式分析——应用与发展[J]. *现代外语*, 2018 (3): 425-435.
- 胡健, 张佳易. 认知语言学与语料库语言学的结合: 构式搭配分析法[J]. *外国语*, 2012 (4): 61-69.
- 罗思明, 王文斌, 洪明. 英汉结果构式R (AP) 制约的语料库与类型学研究[J]. *外语教学与研究*, 2010 (4): 268-274.
- 王寅. 动结构式的体验性事件结构分析[J]. *外语教学与研究*, 2009 (5): 345-350.
- 卫乃兴. 词语学要义[M]. 上海: 上海外语教育出版社, 2011.
- 辛志英, 单健. 英语及物性系统中结果属性的入列条件[J]. *现代外语*, 2019 (6): 731-742.
- 殷红伶. 英语动结式的语义结构问题[J]. *解放军外国语学院学报*, 2010 (6): 15-18.
- 殷红伶. 英汉动结式语义结构研究[M]. 南京: 东南大学出版社, 2011.

通信地址: 266100 山东省青岛市 中国海洋大学外国语学院 (刘聪颖、李潇辰)
100049 北京市 中国科学院大学外语系 (曹笃鑫)

变异语言学视角下英语并列二项式词序多元定量研究*

大连海事大学 林奕冰 孟庆楠

提要：本研究采用基于语料库的变异语言学视角及多因素分析的研究方法，借助英国国家语料库和R统计软件，对名词和形容词并列二项式词序的制约因素进行定量研究。结果表明：制约名词和形容词并列二项式词序的主要因素不尽相同。在绝大多数情况下，音节数量和尾音响亮度分别对名词和形容词并列二项式的词序起到决定性作用。

关键词：英语并列二项式、词序制约因素、可逆性、多元定量研究、随机森林分析

1 引言

在英语实词短语的使用方面，并列二项式（binomials）是必不可少且常见的语言形式。并列二项式，又称双项式并列短语，来自同一词类、隶属于同一句法层级的词语并列组合而成，并由连词等词汇手段进行连接（Malkiel 1959）。并列二项式具有悦耳且易于记忆的特性，有助于提升语言表达的准确性、流利度和趣味性。如果说话者使用并列二项式的劣势词序，听者往往需要更长的反应时间，从而造成沟通的障碍。因此，对并列二项式词序的制约因素进行定量研究，有助于深入了解其词序架构的特点以及词序不可逆的原因，对非英语母语使用者提升语言表达能力亦具有重要意义。

语言学家对并列二项式的研究已经取得了丰硕的成果。国外对英语并列二项式的研究以Mollin（2014）基于语料库的研究为代表，国内学者则侧重于对汉语并列二项式的研究，如对现代汉语中名词性并列结构排序原则的定性研究（廖秋忠 1992），以及对英汉并列二项式词序制约因素的差异研究（刘世英 2015）。但是以上研究均未深入探讨不同词类的并列二项式在制约因素方面的差异。

英语母语者在快餐厅点餐时多会使用fish and chips而非chips and fish。探寻词序差异的原因，能够为人们理解词序选择的偏好提供新的视角。本研究着眼于由

* 本研究系辽宁省社会科学规划基金青年项目“基于原美国杨百翰大学系列语料库的英语句式交替现象研究”（L21CYY004）的阶段性研究成果。

孟庆楠为本文通讯作者。

作者贡献：

林奕冰：数据收集、数据分析、讨论结论、初稿撰写、字数占比（80%）；

孟庆楠：选题构思、研究方法、字数占比（20%）、修改润色。

连词and连接的英语并列二项式,即存在word₁ and word₂与word₂ and word₁两种组合方式,且表达相近含义的短语结构。通过考察不同因素对英语并列二项式词序的制约力,着重探究对于不同词类(名词和形容词)的英语并列二项式而言,这些因素的制约力排序是否存在差异。

2 研究背景与研究问题

多年来,语言学家一直试图考察并列二项式词序的制约因素,通过定量研究评估其有效性,并探究这些制约因素与可逆性之间的关系。Pinker & Birdsong (1979)通过心理学实验,选择gligy and glagy、boof and kaboof等并列二项式作为语料,发现语音因素对词序选择的偏好有所影响。然而,尽管用无实际意义的成对词语成分作为实验语料能够消除语义因素对实验结果的影响,但该实验并未反映出真实的语言加工过程。

Cooper & Ross (1975)提出用“我优先原则”(Me First Principle)来标注心理标记性,即位于and之前的单词通常是用于描述典型说话者的词,例如成年人、男性、对“我”而言积极的事物。Kelly *et al.* (1986)的研究进一步表明,“我”的认知偏好是词序差异的原因之一。当研究者在受试者面前播放并列二项式的劣势词序时,大多数受试者在回忆时仍倾向于以优势词序进行复述。这些前人研究为后续研究者将语料库与心理实验相结合,以便对词序选择的偏好进行合理描述提供了有益的借鉴。

Fenk-Oczlon (1989)通过测试词频因素、语义因素、音节数量、元音质量和开头辅音数量对并列二项式词序的制约力,发现词频因素是影响并列二项式词序选择的决定性因素。然而,该研究得出的结论仅仅基于二分法,忽略了制约因素对并列二项式词序没有影响的情况。Mollin (2014)对Fenk-Oczlon (1989)的数据进行了重新分析,发现语义因素在制约力排序中优先级较高,由此说明了在标注中灵活使用二分法和三分法的重要性。不过,以上研究均以制约因素作为变量,目前尚未有学者就不同词类、语域的并列二项式在制约因素的制约力方面开展研究。

基于此,本研究将重点探讨下面三个问题:(1)不同制约因素对英语并列二项式词序的制约力有何差异?(2)同一制约因素对名词和形容词并列二项式词序的制约力是否相同?(3)起决定性作用的制约因素是否对英语并列二项式的词序选择具有普适性?

3 理论框架及研究方法

本研究采用基于语料库的变异语言学(corpus-based variationist linguistics)研究范式,其基本假设:语言是异质有序且动态变化的,与周围的社会环境有着千丝万

缕的联系。语言变异是语言存在的唯一形式，故语言学研究应注重收集和分析实际使用中的自然语言材料，重视实际生活中活的言语，特别是语言的口头表达（田莉、田贵森 2017）。基于概率语法观（Bresnan 2007），语言使用者对不同语言变体的选择会受到诸多语言学内外部因素的影响。这使得非英语母语使用者在不清楚某个英语并列二项式优势词序的情况下，在特定的语境中亦有可能准确选用优势词序。

基于上述理论框架，本研究采用了多因素分析的研究方法。与传统的基于语言使用频数的描述性方法相比，这一新兴的研究方法并不把对各种语言变体形符频数的统计放在首位，而是通过大型语料库的检索，随机抽取适量语料，对英语并列二项式词序的制约因素进行人工标注。语料的手工标注极为耗时，但是这也最能体现研究的语言学价值（许家金 2020）。根据房印杰（2016）的观点，多因素分析的重要价值在于其能够从海量制约因素中客观、准确地剥离出对研究对象具有显著影响的因素及因素间的交互效应，从而为进一步的理论阐释奠定坚实的基础。

4 研究设计

为了快速提取相关语料、标注变量，本研究选用离线版本的英国国家语料库（British National Corpus，简称BNC）。BNC语料库是一个通用语料库，拥有超过一亿词的容量，包括书面语（9,000万词）和口语（1,000万词）。该语料库是语料库语言学发展的产物，能够为语言学家研究语言提供真实的材料。根据笔者的初步统计，BNC语料库共收录了超过70万个并列二项式。

笔者通过Python编程，运用正则表达式来提取并列二项式。出于可操作性的考虑，又考虑到语料库中书面语和口语的语料比例大致为9:1，因此笔者按照9:1的比例从两个子库中分别抽取并列二项式，并删除了word₁和word₂相同的并列二项式、协调短语以及含有两个以上词语成分的并列二项式，最终分别筛选出符合要求且频数排序在前600的名词和形容词并列二项式，同时通过替换正则表达式中word₁和word₂的序列，得出优势词序和劣势词序的频数。正则表达式如下：

<w N..>[a-z]+ <w C..>and <w N..>[a-z]+（用于提取名词并列二项式）

<w AJ..>[a-z]+ <w C..>and <w AJ..>[a-z]+（用于提取形容词并列二项式）

在对制约因素进行选取与标注的过程中，笔者参考了前人对并列二项式可逆性的相关研究，并借鉴了Benor & Levy（2006）和Mollin（2014）对相关变量的分类及标注体系，最终选取了10个可能影响并列二项式词序的制约因素，分为语义因素、韵律语音因素、非韵律语音因素以及其他因素4类。其中语义因素又细分为象似性、心理标记性、权力和形式标记性四个子类，非韵律语音因素细分为

音核元音长度、首音响亮度、尾音响亮度3个子类，韵律语音因素用音节数量来度量，如表1所示。

在标注词频和象似性等制约因素时，通过二分法给出“*Yes*”或“*No*”的标注。当标注音节数量、首音响亮度、尾音响亮度、音核元音长度等制约因素时，针对制约因素对并列二项式词序没有影响的情况，本研究采用三分法，引入标注符号“*NA*”，意为“无效”。

表1 并列二项式词序的制约因素及标注依据

类别	名称	水平	标注依据	典型语例
语义因素	象似性	Yes	word ₁ 和word ₂ 是否按所描述事物之间的时间或因果顺序排列	born and bred、trial and error、spring and summer
		No		women and men
	心理标记性	Yes	word ₁ 的含义是否比word ₂ 更典型、更重要或更容易成为信息的焦点	positive and negative、near and far、front and back
		No		bits and pieces
	权力	Yes	word ₁ 与word ₂ 是否按事物对人类社会的重要性呈降序排列	men and women、landlord and tenant, food and water
		No		birds and animals
	形式标记性	Yes	word ₁ 的含义是否比word ₂ 更宽泛、更普遍、结构更简单，并且包含更多的子类	rules and regulations、health and fitness
		No		goods and services
韵律语音因素	音节数量	Yes	word ₁ 的音节数量少于word ₂	urban and industrial、real and imaginary
		NA	word ₁ 的音节数量等于word ₂	women and children
		No	word ₁ 的音节数量多于word ₂	education and training
非韵律语音因素	音核元音长度 ¹	Yes	word ₂ 主要重读音节的元音长度比word ₁ 长	goods and services、husband and wife
		NA	word ₂ 和word ₁ 主要重读音节的元音长度相等	theory and practice
		No	word ₂ 主要重读音节的元音长度比word ₁ 短	ladies and gentlemen

(待续)

(续表)

类别	名称	水平	标注依据	典型语例
非韵律语音因素	首音响亮度 ²	Yes	word ₁ 的第一个音段比 word ₂ 更响亮	here and there
		NA	word ₁ 的第一个音段与 word ₂ 响亮度相同	flora and fauna
		No	word ₂ 的第一个音段比 word ₁ 更响亮	rights and obligations
	尾音响亮度	Yes	word ₂ 的最后一个音段比 word ₁ 更响亮	well and truly
		NA	word ₂ 的最后一个音段与 word ₁ 响亮度相同	teachers and pupils
		No	word ₁ 的最后一个音段比 word ₂ 更响亮	pay and conditions
其他因素	词频	Yes	word ₁ 的频数大于 word ₂ 的频数	facts and figures
		NA	word ₁ 的频数等于 word ₂ 的频数	未发现此类语料
		No	word ₁ 的频数小于 word ₂ 的频数	sheep and cattle
	首字母顺序	Yes	word ₁ 的首字母在字母表中的位置比 word ₂ 的首字母更靠前	boys and girls
		NA	word ₁ 和 word ₂ 的首字母在字母表中的位置相同	shape and size
		No	word ₁ 的首字母在字母表中的位置比 word ₂ 的首字母更靠后	teachers and parents

Malkiel (1959) 根据并列的两个词语成分可逆程度的差异, 首先提出以可逆性曲线来划分并列二项式的种类, 即: 可逆的并列二项式存在两种词序, 而完全不可逆的并列二项式存在一种词序, 这两类二项式分别位于可逆性曲线的两端。例如, 在表达红绿两色时, red and green 是优势词序, 因此在可逆性曲线中, 该并列二项式更靠近完全不可逆的一端。本研究将沿用可逆性曲线这一分析模型, 以

并列二项式在可逆性曲线中所处的位置为依据,探究制约因素对词序的制约力。Mollin (2014) 为将可逆性曲线转化为可观察和测量的具体指标,主张为给定的二项式分配一个可逆度分值 (ir)reversibility score)。具体操作如下:提取一个并列二项式的优势和劣势词序 (即 word₁ and word₂ 和 word₂ and word₁) 各自的频数,并将频数较高的标记为“freq”,频数较低的标记为“revfreq”。通过如下计算公式即可得出可逆度分值:

$$(ir)reversibility\ score = \frac{freq}{freq + revfreq}$$

本研究借用了这一方法,计算从BNC语料库中提取的名词与形容词并列二项式的可逆度分值。该分值用于表示并列二项式优势词序和劣势词序在语料库中的相对比例 (刘世英 2015)。例如,并列二项式 education and training 为优势词序,频数为 373; training and education 为劣势词序,频数为 58。优势词序占 86.54%,即并列二项式 education and training 的可逆度分值为 86.54,表明人们更倾向于使用 education and training 而不是 training and education。对于完全不可逆的并列二项式,如 bits and pieces,只存在一个词序 (revfreq 为 0)。并列二项式词序的可逆度取决于优势词序的频数,优势词序的频数越高,并列二项式的可逆度越低,词序越固化,反之亦然。引入可逆度分值这一参数有利于评估并列二项式的可逆性。

5 数据分析与讨论

对所有变量进行标注后,本研究通过成功率排序和随机森林分析进行交叉验证,以此综合衡量制约因素对名词和形容词并列二项式词序的制约力异同。

5.1 制约因素对并列二项式词序影响的成功率及排序

本研究中,N值为受某因素影响 (水平为“yes”) 的并列二项式总频数,固定值为受该因素影响的各个二项式的可逆度分值之和,成功率为固定值除以N值得到的商。成功率由每个因素成功指示并列二项式实际优先顺序的百分数来表示。以象似性因素为例,在600个名词并列二项式中,共有161个并列二项式受象似性影响,故N值为161;将这161个并列二项式的可逆度分值相加,总和即为固定值13,607.84;用固定值除以N值,商即为成功率84.52。笔者根据因素类型 (语义、韵律语音、非韵律语音和其他) 进行分组,并分别计算出10个制约因素对应的N值、固定值和成功率,结果如表2所示。

表2 制约因素影响并列二项式词序的成功率及排序

制约因素		名词并列二项式				形容词并列二项式			
		N值	固定值	成功率 (%)	成功率 排序	N值	固定值	成功率 (%)	成功率 排序
语义 因素	象似性	161	13,607.84	84.52	3	92	6,267.62	68.13	10
	心理标 记性	184	14,911.17	81.04	7	111	8,860.14	79.82	2
	权力	241	19,978.18	82.90	4	137	10,360.54	75.62	6
	形式标 记性	176	14,053.48	79.85	9	186	14,468.32	77.79	4
韵律语 音因素	音节 数量	262	22,562.74	86.12	1	243	19,238.06	79.17	3
非韵律 语音因 素	音核元 音长度	119	9,792.87	82.29	6	122	8,906.69	73.01	7
	首音响 亮度	211	17,396.31	82.45	5	203	14,453.31	71.20	9
	尾音响 亮度	214	18,250.16	85.28	2	182	14,678.37	80.65	1
其他 因素	词频	368	29,486.66	80.13	8	368	28,316.97	76.95	5
	首字母 顺序	231	18,359.42	79.48	10	257	18,601.54	72.38	8

通过观察表2可知，制约因素对名词和形容词并列二项式词序的影响不尽相同。通过比较成功率发现，音节数量是影响名词并列二项式词序的决定性因素，而尾音响亮度则是影响形容词并列二项式词序的决定性因素。在综合分析这些制约因素对名词和形容词并列二项式词序的制约力后发现，对于名词并列二项式，音节数量（86.12%）排在首位，尾音响亮度（85.28%）、象似性（84.52%）次之，权力（82.90%）、首音响亮度（82.45%）、音核元音长度（82.29%）的成功率差距较小，这说明它们对名词并列二项式词序的制约力相差无几。心理标记性（81.04%）、词频（80.13%）、形式标记性（79.85%）、首字母顺序（79.48%）排在末位。对于形容词并列二项式，尾音响亮度（80.65%）为决定性因素，随后是心理标记性（79.82%）和音节数量（79.17%）。形式标记性（77.79%）和词频（76.95%）居中。权力（75.62%）、音核元音长度（73.01%）、首字母顺序

(72.38%)、首音响亮度(71.20%)以及象似性(68.13%)的成功率均相对较低。

对于受多个制约因素影响的并列二项式,成功率较低的因素对词序的制约力很难被察觉。不过,笔者认为这些因素的有效性也应当得到检验。尽管如音核元音长度等非韵律语音因素只在少量例子中起作用,但这些因素作为英语并列二项式的潜在决定因素有着悠久的历史,并多次被证实可以影响人们对无实际意义的成对词语成分的词序选择偏好。因此,只要成功率较高的因素不影响并列二项式的词序,这些因素就很可能成为决定性因素。成功率排序为9或10的制约因素则不再被视为影响因素,因为它们可能是主要制约因素如尾音响亮度的副产物。与其他因素相比,词频因素并不体现在并列二项式的两个词语成分本身上,但该因素一直是语言产生和处理的决定性因素(Ellis 2002),并被认为是词汇化背后的驱动力(Bybee 2010)。整体来看,制约因素对名词并列二项式词序的制约力普遍高于形容词并列二项式,而音节数量和尾音响亮度对两种词类并列二项式的制约力均相对较高。

5.2 并列二项式词序制约因素的随机森林分析

本研究通过R语言实现随机森林分析这一探索性统计分析方法。根据李欣海(2013)的观点,在构建分类树时,随机森林会随机地在因变量中选择 n 个观测值,同时从 k 个自变量中选择部分变量确定分类树节点,因此每次构建的分类树都可能不同。在随机生成的几百个甚至几千个分类树中,重复程度最高的树将作为最终结果(Breiman 2001)。与决策树分析相比,随机森林分析被广泛应用于数据分类和非参数回归分析(董师师、黄哲学 2013),可以通过点状图对各种变量的相对重要性进行可视化呈现(Szmracsanyi *et al.* 2016),尤其适合处理样本量较小但预测变量较多的数据。通过在R软件中加载{party}程序包,使用cforest和varimp函数可基于条件推断决策树构建随机森林模型。结果如图1所示。

在本研究中,图1的纵轴展示了研究选取的所有制约因素,各点对应的横坐标代表每个预测变量的相对重要性。制约因素的制约力通过各点与虚线之间的距离来体现,点离虚线越远,说明制约因素的制约力越强。

通过随机森林分析可以发现5.1中的结论与图1基本一致。对于名词并列二项式而言,音节数量的成功率最高(86.12%),在图1中距离虚线最远,相对重要性最强;首字母顺序的成功率最低(79.48%),在图1中对应的横坐标位于虚线左侧。成功率相对较低的限制因素(心理标记性、词频、形式标记性)的横坐标值靠近虚线,这表明这些因素对名词并列二项式几乎没有制约力。成功率相对较高的限制因素(音核元音长度、首音响亮度、尾音响亮度)更为重要。对于形容词并列二项式而言,尾音响亮度、音节数量、心理标记性和词频这4个限制因素的成功率均超过85.00%。虽然随机森林分析的结果与成功率的排序并不完全一致,

但显然这4个因素的横坐标值仍处于中位。相比之下，其他6个因素的成功率较低，相对重要性接近于零，这表明尽管一些并列二项式的词序符合这些因素的标注依据，但这些因素并未产生影响。

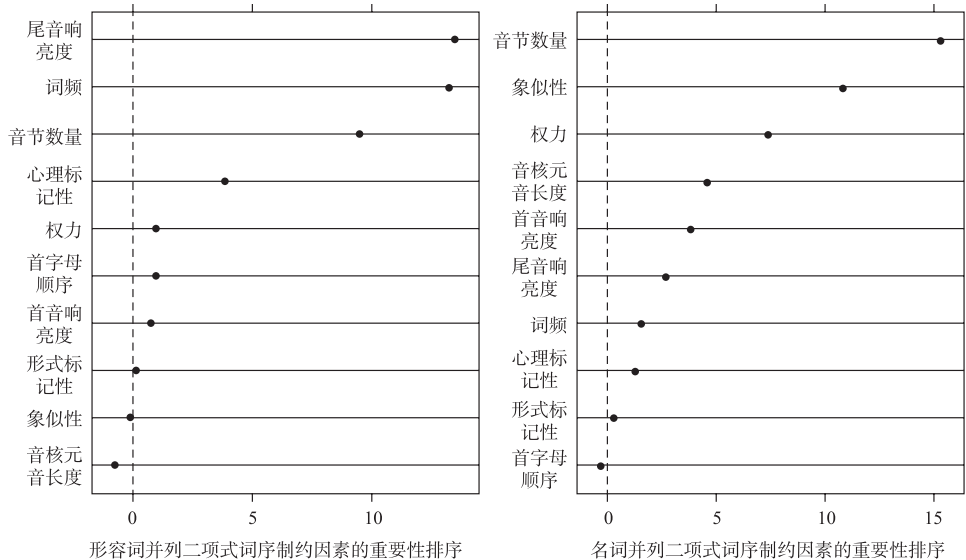


图1 并列二项式词序制约因素的随机森林分析

随机森林分析的结果基本证实了表2中的结论。但是，这一排序规律不一定适用于所有并列二项式。在大多数情况下，该规律体现了人们对词序选择的偏好，增强了英语使用者对并列二项式可逆性程度的直觉判断，同时为语言的规范化提供了帮助。少数情况下，一些不符合规律的并列二项式也存在，现举例说明如下。

在并列二项式 *bride and groom* 中，*groom* 本应作为第一个词语成分，但从心理标记性的角度来看，新娘在婚礼中更为夺人眼球，因此 *bride* 作为第一个词语成分的情况更为普遍。此时心理标记性优先于权力。

在并列二项式 *beginning and end* 中，从音节数量的角度看，*end* 的音节数更少，本应作为第一个词语成分，但从象似性的角度看，*beginning* 先发生，因此 *beginning* 作为第一个词语成分的情况更为普遍。此时象似性优先于音节数量。

由此可见，心理语言学能够为不符合规律的并列二项式的词序偏好作出解释。对于完全不可逆的并列二项式，如 *bits and pieces*、*by and large*，可逆度分值为0，但这些并列二项式也会由于口误、非英语母语者的一知半解，或是小说和媒体写作中精心设计的倒装等原因而出现倒置。不过，笔者认为，在心理上，这些并列二项式对英语母语者而言是完全不可逆的，基于语料库的可逆度分值已在很大程度上佐证了这一点。可逆度分值是语言生成的一个表象，还是心理学意义上的真实概念，这一点有待作为未来的研究主题。

6 结语

本研究基于变异语言学视角,借助BNC语料库和R统计软件,通过随机森林分析的方法探究了名词和形容词并列二项式的可逆性及制约因素的制约力在两种词类上的异同。结果表明:名词和形容词并列二项式的词序受到不同制约因素的影响。音节数量对名词并列二项式的词序起决定性作用,而尾音响亮度则对形容词并列二项式的词序起决定性作用。

本研究在研究工具和方法方面还存在一些不足。首先,本研究语料数量较少,原始频数的代表性存在着一定的局限性。不同地理环境和社会文化对事物价值的定义存在差异,使得部分语义因素(如心理标记性、权力)的标注结果缺乏普遍性。后续研究可以考虑采用大型语料库对本研究所得出的结论进行交叉验证,以期减小偏差。其次,本研究仅从整体上论述了制约因素在名词和形容词并列二项式之间的区别与联系,并未对书面语和口语语料进行区分,也未对不同语域的并列二项式(如小说、新闻、杂志)进行分类探讨。以上缺陷,还有待后续研究进一步完善。

注释

- 1 Mollin (2014) 将长元音 /i:, u:, ɜ:, ɔ:, ɑ:/ 和二合元音 /eɪ, aɪ, ɔɪ, aʊ, əʊ, ɪə, eə, ʊə/ 标注为长的元音;将短元音 /ɪ, ʊ, e, æ, ʌ, ɒ/ 标注为短的元音。
- 2 为了标记首音响亮度、尾音响亮度,本研究引入了响亮度层级。然而,关于响亮度层级的排序至今仍存在争议。到目前为止,以Parker (2011) 为代表的主流观点是:元音>滑音>流音>鼻音>阻塞音,故本研究采用了该响亮度层级。

参考文献

- BENOR S, LEVY R. The chicken or the egg? A probabilistic analysis of English binomials [J]. *Language*, 2006, 82(2): 233-277.
- BREIMAN L. Random forests [J]. *Machine Learning*, 2001, 45(1): 5-32.
- BRESNAN J. Is syntactic knowledge probabilistic? Experiments with the English dative alternation [C]//FEATHERSTON S, STERNEFELD W. *Linguistics in search of its evidential base*. Berlin: Mouton de Gruyter, 2007: 75-96.
- BYBEE J. *Language, usage and cognition* [M]. Cambridge: Cambridge University Press, 2010.
- COOPER W, ROSS J. Word order [C]//GROSSMAN R, SAN L, VANCE T. *Papers from the parasession on functionalism*. Chicago: Chicago Linguistic Society, 1975: 63-111.

- ELLIS N. Frequency effects in language processing [J]. *Studies in Second Language Acquisition*. 2002, 24(2): 143-188.
- FENK-OCZLON G. Word frequency and word order in freezes [J]. *Linguistics*, 1989, 27(1): 517-556.
- KELLY M, BOCK J, KEIL F. Prototypicality in a linguistic context: Effects on sentence structure [J]. *Journal of Memory and Language*, 1986, 25(1): 59-74.
- MALKIEL Y. Studies in irreversible binomials [J]. *Lingua*, 1959, 8: 113-160.
- MOLLIN S. The (Ir)reversibility of English binomials [M]. Amsterdam: John Benjamins, 2014.
- PARKER S. Sonority [C]//OOSTERNDORP M, EWEN C, HUME E. *The Blackwell Companion to Phonology*. New Jersey: Blackwell, 2011: 12-14.
- PINKER S, BIRDSOON D. Speakers' sensitivity to rules of frozen word order [J]. *Journal of Verbal Learning and Verbal Behavior*, 1979, 18(4): 497-508.
- SZMRECSANYI B, GRAFMILLER J, HELLER B, RÖTHLISBERGER M. Around the world in three alternations: Modeling syntactic variation in varieties of English [J]. *English World-Wide*, 2016, 37(2): 109-137.
- 董师师, 黄哲学. 随机森林理论浅析[J]. *集成技术*, 2013 (1): 1-7.
- 房印杰. 语言学研究中的多因素分析[J]. *语料库语言学*, 2016 (1): 82-92.
- 李欣海. 随机森林模型在分类与回归分析中的应用[J]. *应用昆虫学报*, 2013 (4): 1190-1197.
- 廖秋忠. 现代汉语并列名词性成分的顺序[J]. *中国语文*, 1992 (3): 161-173.
- 刘世英. 英汉并列二项式词序制约因素的语料库研究: 可逆性与预测成功率[J]. *外语教学*, 2015 (3): 39-43.
- 田莉, 田贵森. 变异社会语言学的研究方法论[J]. *外语学刊*, 2017 (1): 25-30.
- 许家金. 多因素语境共选: 语料库语言学新进展[J]. *外语与外语教学*, 2020 (3): 1-21.

通信地址: 116026 辽宁省大连市 大连海事大学外国语学院

计算社会科学、文化组学与语言学*

浙江大学 邵 斌 李雨飞

提要：近些年，计算社会科学和文化组学研究大有融合的趋势，上述两大新兴领域又与语言学，尤其是语料库语言学息息相关。本文对近年来文化组学在计算社会科学中的应用做了述评，指出语言学是文化组学及计算社会科学的基础，并探讨了语言学者参与其他社会科学研究的可能性。本文认为，未来的人文社会科学研究会进一步呈现学科交叉和数据驱动的特点，其发展离不开语言学者的参与，同时语言学者也有能力借助海量的语言数据探索其他人文和社会科学的规律。

关键词：计算社会科学、文化组学、语言学、语料库

1 引言

2009年，以David Lazer为首的15位学者在《科学》杂志上发表了《计算社会科学》(Computational Social Science)一文，标志着“计算社会科学”这一新兴学科的诞生。计算社会科学是收集和分析海量数据以揭示个体或集体行为模式的新兴领域(Lazer *et al.* 2009)。其核心内容是对人类社会发展中的各类信息进行自动化处理，透过行为分析、媒体分析、网络分析和对现实社会的典型化事实的分析，运用代码、算法、程序、建模、模拟等数字化手段，对个体行为特征与社会运行规律及二者相互作用进行了深入的观察和探讨(王国成 2020)。目前，计算社会科学的研究领域已涉及社会学、心理学、教育学、政治学、经济学、管理学、公共卫生学等各个社会科学学科，成为跨学科研究的前沿热点。

与此同时，作为人文社会科学领域一个新的研究范式，“文化组学”(culturomics)也迅猛发展。文化组学的概念源自Michel *et al.* (2011)在《科学》杂志上发表的《使用百万数字化书籍的文化定量分析》一文，其定义是通过大规模数据文本的量化分析对人类文化及其发展趋势的研究。Culturomics一词为

* 本文系国家社科基金重点项目“基于语料库的汉英动名兼类词历时演变对比研究(1919—2019)”(20AYY001)的阶段性研究成果。邵斌为本文通讯作者。

作者贡献：

邵斌：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；

李雨飞：数据收集、数据分析、初稿撰写、字数占比(40%)。

culture（文化）和genomics（基因组学）的缩合。该概念背后的理念是：大规模基因组的DNA序列能够揭示生命信息，与之同理，作为人类文化承载物的词语，其海量数据的累积也能揭示人类文化的特点。

Michel *et al.*（2011）的研究以“谷歌图书语料库”为基础，该语料库收录自1500年以来包含英、法、德、西、俄、汉和希伯来语等7种语言的3,000万册电子图书文本，总计达5千亿词。因此，谷歌图书语料库不仅是“大数据”，更是“长数据”（long data），可对各种演变现象进行定量研究。研究人员开发了“谷歌图书词频查看器”（Google books Ngram Viewer），以曲线图形式呈现语料库中单词或词组的使用频率变化。之后，Aiden & Michel（2013）所著的《可视化未来：数据透视下的人文大趋势》又进一步详细论述了文化组学的上述应用。文化组学概念的提出虽只有十余年，但国外已有数以百计的著述探索其应用。近些年，国内开展了相关研究，如陈云松（2015）考察了大数据中的百年社会学；陈云松等（2015a）研究了500年来中国世界文化遗产的国际知名度变迁；陈云松等（2015b）探讨了近300年中国城市的国际知名度的变化；龚为纲、罗教讲（2015）以丝绸、瓷器与茶叶的文化影响力为中心考察了19世纪“海上丝绸”；邵斌（2017）论证了浙江文化关键词在英语世界的影响力；王国燕、沈佳斐（2018）对百年传播学史进行研究等。但总体而言，国内研究仍处于起步阶段。

事实上，文化组学与计算社会科学的研究范围有诸多重合之处，两者都借助大数据，通过数据驱动的方法对人类社会和人类行为进行考察。当然，两者的侧重点稍有不同，一是文化组学更侧重考察人文学科的话题，如文学、历史、哲学、语言、文化、艺术等领域，而计算社会科学更侧重考察社会科学领域，如经济学、管理学、法学、教育学等学科。二是文化组学研究更强调历时演变研究，其命名来源“组学”便是强调基因的代代相传性，故其本质就有考察演变的要求，而当下计算社会科学研究多考察人类社会现状，更多为共时研究。

可喜的是，近些年文化组学的范围已不再局限于人文领域，而是对社会科学各个领域的演变现象进行广泛研究，计算社会科学的考察也加强从历时维度观察社会科学中的演变现象。换言之，文化组学与计算社会科学出现了交叉和融合。邵斌、王文斌（2018）对文化组学在人文学科中的研究做了综述，但对文化组学在社会科学领域的应用尚乏介绍。此外，前人很少谈及语言学在文化组学和计算社会科学研究中的作用。事实上，文化组学是基于词汇计量的研究，计算社会科学也需以语言数据为基础。西方的语言学者如今也借用语料库语言学和计算语言学的方法研究社会科学领域的话题，拓宽了语言学的研究范围，这些研究在国内语言学界也未引起足够重视。

有鉴于此，本文拟从以下两方面展开综述和阐释。一是考察文化组学在社会科学领域的应用，二是论证语言学在文化组学和计算社会科学领域中的重要作用，

探讨语言学进入社会科学研究的可能性，以期对语言学的跨学科研究提供启示。

2 文化组学在社会科学领域的应用

自文化组学诞生之初，就未将自己局限于文化和人文学科领域。Michel *et al.* (2011) 所开展的研究个案中，预测英语词汇总量以及探讨不规则动词的演化规律属于语言本体研究，可归为人文学科，但对人的名气的测算以及出版审查制度的追踪则属于社会科学领域。Leetaru (2011) 已经将研究话题拓展至预测社会形势和挖掘文本背后的地理空间（如本·拉登的藏身之处）等社科领域。归根结底，文化组学研究以词为起点研究人类行为的变化，其本质是基于语言数据发现人类行为的一般模式及其演变规律，故不必局限人文领域。下文对文化组学在社会科学领域的应用做一综述。

在社会学领域，文化组学可用于考察横亘百年乃至千年的社会行动、人际互动和社会结构的线性流动。Schich *et al.* (2014) 基于过去2000年来15万名杰出人物的出生和死亡地点数据，从宏观的角度归纳了这些人物的空间迁徙模式，绘制了2000年来欧洲和北美的名人迁徙图。Chen & Yan (2016) 利用美国英语谷歌图书探索20世纪书籍中社会阶层的公众关注与经济发展之间的关联性。研究发现，美国社会的公众对社会阶级的关注程度受到“经济悲惨指数” (economic misery index) 的影响，但与收入不平等（即基尼系数）的关联性并不强。还有学者利用文化组学探究特定社会现象背后的原因。Weiss & Hoegl (2015) 以teamwork相关词语作为检索词考察美国社会中“团队”概念的传播情况。结果表明，团队概念在过去一个世纪里呈现波动状态，大规模的重要事件，如战争、技术革新以及机构组织变化，会对团队精神的社会扩散产生影响。文化组学在社会学各个领域的应用蓬勃发展，以数据挖掘方式探讨了现有社会学研究未曾触及的领域，催生了更多具有科学性的社会研究模式。

在心理学领域，文化组学展示了社会集体认知、情绪、行为习惯的演变。比如，Scheffer *et al.* (2021) 利用谷歌图书词频查看器、《纽约时报》数据库和谷歌搜索关键词考察了美国公众价值观的变迁。检索发现，19世纪中期之后，与理性主义相关的词汇，如“确定” (determine) 和“结论” (conclusion)，呈现上升趋势，而与人类经验相关的词汇，如“感觉” (feel) 和“相信” (believe)，呈现下降趋势。但自1980年之后，这种模式发生了逆转，理性词语下降，感性词语上升。由此可见，社会价值观在近几十年内发生了转变，从集体主义转向个人主义，从理性导向转为情感导向。Bollen *et al.* (2021) 对“认知扭曲” (cognitive distortion) 现象的流行进行了定量分析。其研究考察了英语、西班牙语和德语的1,400万书籍，发现在过去20年里，有关认知扭曲的书写大幅上升。社会集体近年来的焦虑和抑郁程度激增，这与社会经济变化、科技发展和社交媒体的发达

有着密切的联系。此外，文化组学还被用来探索集体记忆与遗忘机制，为集体心理学研究的宏大板块贡献了重要的拼图。比如，West *et al.* (2021) 系统性地分析了社交媒体和新闻对公众人物集体记忆产生的影响。通过建立近5年在线新闻和社交媒体推文的综合数据集，该研究追踪了数千名公众人物在其去世后的一年中被提及的情况。研究表明，新闻和社交媒体在集体记忆形成过程中扮演着独特角色，且对不同类型的公众人物触发的效应有所区别。在自然语言与集体认知层面，文化组学也逐渐成为心理学研究的有力工具。Dodds *et al.* (2015) 测量了语言中的情感。该研究发现，在10种语言的24个语料库中，常用词表现出明显的“语言积极倾向”(linguistic positivity bias)，这从大数据视角证实了“波丽安娜假说”(Pollyanna hypothesis)，即人总是看重和追求好的一面，摒弃坏的一面，因此倾向于把积极的词当作无标记，把消极的词当作有标记。在此基础上，Iliev *et al.* (2017) 对“语言积极倾向”进行了更为深入的研究，从时间维度(1800—2000年)探索了其起源，发现语言积极倾向与客观环境和主观情绪的波动有关，不能简单地认为是认知偏见的产物。文化组学与心理学研究的联系日益密切，涉及知觉、认知、情绪、思维、人格、行为习惯和人际关系等多个方面，且常与家庭、教育、健康和社会等发生关联。

在政治学领域，文化组学为探究政治行为和政治决策提供了独特视角，通过语言分析以小见大，展现政治发展趋势。Jordan *et al.* (2019) 分析了政治领导人和文化机构的语言特征，考察了政治和文化发展的总体趋势。研究聚焦于“分析思维”(analytic thinking) 和“信心”(confidence) 这两个重要的心理学维度。结果表明，在过去一个世纪里，政治领导人与公众沟通的过程呈现分析思维持续下降，而信心持续上升的趋势。特朗普的语言使用便是典型代表。在经济学和管理学领域，文化组学研究从宏观角度梳理学科发展脉络，将特殊事件连点成线，绘制出更为完整的学科发展地图。Westley (2014) 借助词频查看器考察了奥地利经济学派的发展史，特别是该经济学派里的事件、观点以及人物的影响力变化。比如，对比“moral sciences”(道德科学) 和“economics”(经济学) 两词的词频曲线图可以发现，1871年经济学家卡尔·门格尔的《国民经济学原理》一书出版后不久，经济学就从传统的道德科学中分离出来，成为独立的学科。学者们还运用文化组学考察管理学思想的演变规律和发展趋势，探究不同流派兴衰的本源，通过社会发展需求预测未来学科动向。比如，Kumar & Sahu (2010) 借助词频查看器考察了营销学的发展史。结果显示，营销学历经诸多阶段，先是以生产为导向，之后以销售为导向，再以营销为导向，最后发展到现代社会营销学阶段。Warner (2016) 也以词频查看器探索20世纪管理学思想的生命周期。词频曲线显示，科学管理学说兴盛于20世纪20年代，人际关系学说兴盛于20世纪60年代，人力资源管理学说则兴盛于90年代末和21世纪初。由此可见，文化组学方法能清晰呈现管理思想的演变。

综上所述，文化组学已经应用于社会科学的各个领域，如社会学、心理学、政治学、经济学和管理学等。文化组学推进了社会科学领域的数字化进程，预示着数据驱动的跨学科研究已成为当前社会科学研究的重要特征。简言之，文化组学给计算社会科学增加了历时维度，为洞察社会科学的变化规律提供了新视角。文化组学中跨语言、跨时期的语言数据可用来对比、识别、发现和验证社会科学中的问题，同时，文化组学作为一种强调考察历时演变的实证方法，促进了计算社会科学历时研究的发展。

3 语言学与文化组学及计算社会科学的相通性

搜寻和分析模式是人文研究的中心任务（Youngman & Carmichael 2014）。复杂性科学专家 Steven Johnson（2010：127）甚至认为，“人类在本质上就是模式的认知者”。模式认知不仅是人类的一项伟大技能，而且是我们神经回路不可或缺的部分。Fisher（2009）认为，有两种方法可以寻找到模式，一种是通过想象，一种是通过统计。前者是传统的人文主义学者所秉持的，而后者正是数据专家所擅长的，两者相辅相成，并行不悖。在大数据时代，后者因具有实证性、可验证性以及可学习性，所发挥的作用日益重要。

纵观文化组学在社会科学中的应用研究，其研究的着眼点往往是概念的变迁，而概念需通过语言，尤其是词汇加以表征，因此文化组学研究与语言学息息相关。甚至可以说，文化组学就是以词为起点去研究人文和社会科学的演变现象，进而发现人类行为的一般规律。比如，上文谈及的 Weiss & Hoegl（2015）考察美国社会中“团队”概念传播的个案以及 Scheffer *et al.*（2021）考察美国公众价值观变迁的个案，都需要先将概念转化为可检索的词，通过词语频率的变化来考察概念的演变，进而探索人类行为和人类社会的演变。文化组学和计算社会科学研究的基础是语言学，尤其是词与概念、词的形式与意义之间错综复杂的关系。绝大多数文化组学和计算社会科学的研究都需考虑并处理词语的多义（polysemy）和同义（synonymy）以及一形多义（semasiology）和一义多形（onomasiology）等问题，否则其研究很可能在源头上就会出现偏差。而上述问题在语言学中已有较为成熟的研究，其研究成果可为文化组学和计算社会科学研究所用。比如，考察金融领域的概念变迁，如果仅仅把 bank 作为“银行”的对应词，忽略 bank 的同形异义性（如作为“河堤”的意义），就会导致结果出现偏差。反过来，研究“团队”概念变迁时，则必须考虑一义多形的问题，比如该概念除了 teamwork 之外，是否还可由其他词语来表达？仅仅检索 teamwork 是否存在陈述不足的缺陷？

文化组学和计算社会科学研究也需要考虑语类和语域以及语料库的代表性。比如，West *et al.*（2021）考察的是新闻和社交媒体对公众人物的集体记忆。这其中就涉及新闻语类和社交媒体语类的构成，比如哪些报纸可作为新闻语类的代

表? 哪些社交媒体可作为媒体语类的代表? 语类和语域研究是语言研究的重点, 其研究方法和成果也可作为文化组学和计算社会科学研究提供借鉴。此外, 语料库的代表性是语料库建设中的关键, 从早期的布朗语料库到BNC语料库, 再到当下在线的COCA语料库等, 语言学者在语料库建设方面已经积累了丰富的经验, 开发了许多有用的工具, 探索出诸多可行的方法, 这也可作为文化组学和计算社会科学提供借鉴。

文化组学和计算社会科学研究与语料库语言学所用的方法有诸多共通之处。首先, 上述研究的关键都是语言数据的频率及其变化。比如, Chan *et al.* (2019) 考察过去200年科学家的名气变化, 便是基于科学家名字频率变化进行的研究, 这与语料库语言学家考察词语或语法结构的变化并无二致。其次, 上述研究都会运用复杂的统计方法, 且方法多有可相互借鉴之处。比如, Gálvez *et al.* (2019) 考察了过去50年西方电影中性别与智力之间的刻板联系。文中用来衡量相关性的正点互信息 (positive pointwise mutual information) 概念, 便是由语料库语言学家 Kenneth Ward Church 和 Patrick Hanks 在计算词语搭配强度时提出的 (Church & Hanks 1990)。由此可见, 语料库语言学为文化组学和计算社会科学研究提供了诸多工具、路径和方法。

事实上, 一旦语言学者的研究兴趣和话题不仅仅聚焦于语言, 也关注人类行为和人类社会时, 语言学者便自然而然地从语言学领域跨到了社会科学研究领域。这其中的典型代表是英国兰卡斯特大学的社会科学语料库研究中心 (Corpus Approaches to Social Science, 简称CASS) 的学者所做的研究。众所周知, 兰卡斯特大学是语料库语言学的研究重镇, 但最近10多年, 该校的语言学者已将语料库研究方法的应用拓展至各个社会科学领域, 证实了语言学在社会科学研究中的基础地位。他们所涉及的社会科学研究话题如下: (1) 英国报刊中穆斯林及伊斯兰文化的表现, 如 Baker *et al.* (2013); (2) 网络暴力; (3) 英国议会中反对同性恋权利话语的变迁, 如 Baker (2019); (4) 英格兰护卫联盟: 英国右翼民粹主义的网上运动; (5) 隐喻和疾病, 如 Potts & Semino (2019); (6) 英国17世纪乞丐和流浪汉的贫穷者话语, 如 McEnery & Baker (2019); (7) 公司财务信息中的叙事话语; (8) 英国和巴西的气候变化, 如 Deignan *et al.* (2019); (9) 英国公共话语中的社会关怀, 如 Brookes *et al.* (2021); (10) 英国媒体中的海洋形象; (11) 200年来英国的干旱事件; (12) 新冠肺炎疫情所涉的隐喻, 如 Semino (2021)。由上可知, CASS研究中心运用语料库方法考察了人类社会的方方面面, 涉及的学科不仅覆盖政治学、宗教学、社会学、经济学和公共管理等社会科学, 甚至还包括医学和自然灾害等自然科学。语料库语言学家之所以能够跨到社会科学领域进行研究, 其原因一是研究方法的共通性, 二是语言数据是所有人文社科研究的基础, 而研究语言数据正是语言学的基本任务, 这也是语言学可以成为且应该成为其他

人文社会科学研究的基础所在。

综上所述，文化组学和计算社会科学与语言学关系密切，甚至不妨说，语言学是文化组学和计算社会科学研究的基础。文化组学和计算社会科学研究都需要考虑词与义的关系，需重视语域概念，需确保语料库的代表性，而且其研究方法与语料库语言学也有共通之处。语言学者也不必在语言学领域画地为牢，而不妨把触角伸向语言学之外，参与更广泛的社会科学研究，以探索人类社会和人类行为的规律，从而使得语言学在人类社会发展中发挥更大的作用。

4 结语

Aiden & Michel (2013: 8) 指出：大数据会改变人文科学和社会科学的研究范式。基于大数据的文化组学和计算社会科学之提出虽只有十多年，但它对人文社会科学研究已产生广泛的影响。正如 Liberman (2010) 所言，2010 年与 1610 年相仿佛。由于数字文本和话语不断激增和存储，用于分析和计算的新工具层出不穷，21 世纪成为发明望远镜和显微镜的 17 世纪初时代的翻版。如今所能观察到的不同时空及文化语境中的模式，其规模不啻为以往的数百万倍。无论身在何处，借助于此类新工具，即可发现有趣的新兴模式。

本文着重对文化组学在计算社会科学中的应用做了述评，并对语言学与文化组学及计算社会科学的关系进行了探讨，指出语言学在后两者研究中的基础作用，以及语言学者参与其他社会科学的可能性。总而言之，未来的人文社会科学研究会进一步体现出学科交叉性和数据驱动性。其他社会科学研究有必要借鉴语言学的研究成果，语言学者也有必要走出语言学的象牙塔，借助语言数据探索其他人文和社会科学发展的规律，进而对人类社会发展和文明进程做出更大的贡献。

注释

- 1 详见其网站介绍 <https://cass.lancs.ac.uk/cass-briefings/>。

参考文献

- AIDEN E L, MICHEL J. Uncharted: big data as a lens on human culture [M]. New York: Riverhead Books, 2013.
- BAKER P. Fabulosa!: the story of Polari, Britain's secret gay language [M]. London: Reaktion Books, 2019.
- BAKER P K, GABRIELATOS C, MCENERY T. Discourse analysis and media attitudes: the representation of Islam in the British press [M]. Cambridge: Cambridge University Press, 2013.

- BOLLEN J, THIJ M T, BREITHAUPT F, et al. Historical language records reveal a surge of cognitive distortions in recent decades [J]. *PNAS*, 2021, 118(30): e2102061118.
- BROOKES G, MCENER T, MCGLASHAN M, et al. Narrative evaluation in patient feedback [J]. *Narrative Inquiry*, 2021, 32(1): 9–35.
- CHAN B, MISXON F, TORGLER B. Fame in the sciences: a culturomics approach [J]. *Scientometrics*, 2019, 118(2): 605-615.
- CHEN Y, YAN, F. Economic performance and public concerns about social class in twentieth-century books [J]. *Social Science Research*, 2016, 59: 37-51.
- CHURCH K W, HANKS P. Word association norms, mutual information, and lexicography [J]. *Computational Linguistics*, 1990, 16(1): 22–29.
- DEIGNAN A, SEMINO E, PAUL S S. Metaphors of climate science in three genres: research articles, educational texts, and secondary school student talk [J]. *Applied Linguistics*, 2019, 40(2): 379–403.
- DODDS P N, CLARK E, DESU S, et al. Human language reveals a universal positivity bias [J]. *PNAS*, 2015, 112(8): 2389–2394.
- FISHER L. *The perfect swarm: the science of complexity in everyday life* [M]. New York: Basic Books, 2009.
- GALVEZ R H, TIDDENBERG V, ALTSZYLER E. Half a century of stereotyping associations between gender and intellectual ability in films [J]. *Sex Roles: A Journal of Research*, 2019, 81(9-10):643–654.
- ILIEVE R, HOOVER J, DEHGHANI M, et al. Linguistic positivity in historical texts reflects dynamic environmental and psychological factors. [J]. *PNAS*, 2017, 113(49), E7871–E7879.
- JOHNSON S. *Emergence: the connected lives of ants, brains, cities, and software* [M]. New York: Scribner, 2010.
- JORDAN K, JOST J T, PENNEBAKER J W, et al. Examining long-term trends in politics and culture through language of political leaders and cultural institutions. [J]. *PNAS*, 2019, 116(9): 3476–3481.
- KUMAR N, SAHU M. The evolution of marketing history: a peek through Google Ngram Viewer [J]. *Asian Journal of Management Research*, 2010, 2: 415–426.
- LAZER D, PENTLAND A, ADAMIC L, et al. Computational social science [J]. *Science*, 2009, 323: 721-723.
- LEETARU K. Culturomics 2.0: forecasting large-scale human behavior using global news media tone in time and space [J]. *First Monday*, 2011, 16 (9).
- LILERMAN M. More on “culturomics” [EB/OL]. (2010-12-17) [2020-11-13]. <https://>

- languageolog ldc.upenn.edu/nll/?p=2848.
- MCENER T, BAKER H. Language surrounding poverty in early modern England: a corpus-based investigation of how people living in the seventeenth century perceived the criminalised poor [G] // SUHR C, NEVALAINEN T, TAAVITSAINEN I. From data to evidence in English language research. Leiden: Brill, 2019: 225-257.
- MICHEL J, SHEN Y K, AIDEN A P, et al. Quantitative analysis of culture using millions of digitized books [J]. *Science*, 2011, (331):176–182.
- POTTS A, SEMINO E. Cancer as a metaphor [J]. *Metaphor and Symbol*, 2019, 34(2): 81–95.
- SCHEFFER M, LEEMPUT I, WENANSA E, et al. The rise and fall of rationality in language [J]. *PNAS*, 2021, 118(51): 1-8.
- SCHICH M, SONG C, AHN Y, et al. A network framework of cultural history [J]. *Science*, 2014, 345(6196): 558-562.
- SEMINO E. “Not soldiers but fire-fighters”— metaphors and covid-19 [J]. *Health Communication*, 2021, 36(1): 50–58.
- YOUNGMAN P A, CARMICHAEL T. Big data, pattern recognition, and literary studies: n-gramming the railway in nineteenth-century German fiction [G] // ERLIN M, TATLOCK L. Distant readings: topologies of German culture in the long nineteenth century. Rochester, NY: Camden House Press, 2014: 285-300.
- WARNER M. A life cycle of management ideas: a research note [J]. *Journal of General Management*, 2016, 2: 17–29.
- WEISS M, HOEGL M. The history of teamwork’s societal diffusion: a multi-method review [J]. *Small Group Research*, 2015, (6): 589–622.
- WEST R, LESKOVEC J, POTTS C. Postmortem memory of public figures in news and social media. [J]. *PNAS*, 2021, 118(38): e2106152118.
- WESTLEY C. Ngrams and the Austrian school [J]. *Quarterly Journal of Austrian Economics*, 2014, 3: 365-397.
- 陈云松. 大数据中的百年社会学——基于百万书籍的文化影响力研究 [J]. *社会学研究*, 2015 (1): 23-48.
- 陈云松, 孙艳, 严飞. 大数据中的中国世界文化遗产: 500年国际知名度分析 [J]. *学术论坛*, 2015a (12): 92-98.
- 陈云松, 吴青熹, 张翼. 近三百年中国城市的国际知名度: 基于大数据的描述与回归 [J]. *社会*, 2015b (5): 60-77.
- 龚为纲, 罗教讲. 大数据视野下的19世纪“海上丝绸”——以丝绸、瓷器与茶叶的文化影响力为中心 [J]. *学术论坛*, 2015 (12): 82-91.

- 邵斌. 浙江文化关键词在英语世界的影响力研究——基于文化组学的视角[J]. 浙江学刊, 2017(2): 201-207.
- 邵斌, 王文斌. 文化组学: 大数据时代的人类文化研究[J]. 外语教学理论与实践, 2018(2): 18-23.
- 王国成. 计算社会科学: 发展现状与前景展望[N]. 中国社会科学报, 2020-08-18(4).
- 王国燕, 沈佳斐. 基于 Google Books 的百年传播学史的文化组学研究[J]. 现代传播, 2018(5): 80-85.

通信地址: 310058 浙江省杭州市 浙江大学外国语学院

中外跨国公司英文社会责任报告情感分析对比^{*}

西安外国语大学 宋天祎 黄立波

提要：本文建立了华为公司与英国电信（BT）公司英文企业社会责任报告语料库（CSR报告语料库），借助LIWC-22情感词典对两家公司的社会责任报告进行情感分析对比考察，并从批评话语分析视角阐释差异背后的动因。研究发现，差异主要体现在以下几个方面。（1）华为公司报告倾向于阐述具体事件实例来支撑观点以缩短与读者的情感距离，并注重与合作伙伴长期关系的连结与维护。BT公司报告则更多展现其负责任的情感立场以及与外部社会战略共赢的情感态度。（2）BT公司报告中的情感倾向较为中立，华为公司报告则呈现出由正面情感信息偏多到正、负情感均衡的发展趋势。（3）BT公司报告中阐述了较多独具特色且丰富多元的活动，与外部构建了多维全面的情感关系。华为公司报告则更多体现出了疫情时代下的情感关怀。本文对商务英语写作教学有一定启示，也可为我国跨国公司企业形象的塑造与海外传播提供参考。

关键词：情感分析、英文企业社会责任报告、中外跨国公司、语料库

1 引言

学界有必要对企业公开披露的报告进行研究，以便更深入地理解投资者是如何从中解读信息并做出决策的（Hajek *et al.* 2014；谢德仁、林乐 2015）。企业社会责任报告（Corporate Social Responsibility Report，简称CSR报告）是企业披露经济、社会和环境等方面信息的重要商业报告，承载着加强企业内外部沟通交流的使命。华为公司是我国通信行业具有代表性的高科技跨国企业，自2008年起面向全球发布不同语种的CSR报告，2010年起聘请第三方审计机构对报告进行审验以保证其真实性与有效性，是我国具有社会责任担当的模范企业。英国电信公司（简称BT公司）是通信行业在社会责任领域的全球领先者，很多客户因该公司在社会责任方面的政策和成绩关注它并加大投资力度。两家公司在CSR报告研究领

^{*} 本文系国家社科基金重大项目“围绕汉语的超大型多语汉外平行语料库集群研制与应用研究”（21&ZD290）和西安外国语大学研究生科研基金项目（2021BS005）的阶段性成果。

黄立波为本文通讯作者。

作者贡献：

宋天祎：数据收集、数据分析、初稿撰写、字数占比（70%）；

黄立波：选题构思、研究方法、讨论结论、字数占比（30%）、修改润色。

域具有较好的代表性。话语作为一个不断被话语分析者重新阐释的动态意义实体(王琴 2022),借助大数据手段进行话语间的情感关系分析可以在一定程度上克服人为解读的主观偏差,大大增加话语分析的强度与活力(李战子 2022)。本文以华为和BT公司2012—2021年的英文版CSR报告为研究对象,借助情感词典法对中外跨国公司CSR报告进行情感分析对比,并从批评话语分析视角阐释差异背后的意识形态因素,以期为我国跨国公司英文版CSR报告的撰写提供启示,为企业形象的海外塑造与传播提供一定参考。

2 研究现状

近年来,商务话语的情感分析研究主要集中于自然语言处理、情报科学和管理学等领域,且多为针对情感倾向分析进行的算法设计。奚金金等(2013)从文本情感信息的抽取、分类和检索三个方面阐述了情感分析在商品网络评论中的分析与应用。李涵昱等(2017)研究设计了商品评论中商品属性提取与情感倾向性分析算法。张钊炜(2018)采用了支持向量机方法构建情感词库以实现文本情感倾向性分析,提供了自然语言处理与语言学相结合的新方向。另外一些研究则运用大数据挖掘手段分析了年报、电商评价话语等商务话语的整体情感倾向。Fuoli(2012)从评价理论的态度资源和介入资源切入,探究了CSR报告中企业内外部关系的建构与企业形象的塑造。王立非、邵寒(2017)运用Diction和VOSViewer软件对比了中美企业英文年报话语的情感倾向特征,发现两国年报在情感倾向度和主题分布上均存在显著差异。邵珊珊、王立非(2019)建立情感倾向模型对电商英汉评价话语进行情感分析,对比了英汉正、负评价话语使用上的差异。

然而,目前较少有研究运用情感词典法对商务话语中的CSR报告进行情感挖掘与分析。因此,本文拟用LIWC-22情感词典对比考察华为公司和BT公司CSR报告文本中的情感差异,并阐释背后的意识形态动因。研究主要回答以下两个问题:1)根据LIWC-22情感词典的情感特征数据,华为公司与BT公司的CSR报告在情感关系建构上有哪些差异?2)从批评话语分析视角看,这些差异背后的动因是什么?

3 研究设计

3.1 研究语料

笔者自建的语料库包含华为和BT公司2012至2021年共计20份英文CSR报告,建成语料库的相关参数见表1。

表 1 语料库库容

研究语料	文献数量	形符总数
华为公司报告语料	10	180,104
BT 公司报告语料	10	218,269

语料库中的 CSR 报告文本宏观特征统计见表 2。华为公司 CSR 报告的平均篇幅短于 BT 公司报告，大于 6 个字母的单词比例高于 BT 公司报告，说明华为报告的词汇复杂度可能相对较高。而华为报告的代词使用比例相对较低，连词使用比例相对较高，可能与撰写者为非英语本族语者有关，语言把控力不及英语本族语者。使用连词可以更好地实现连贯效果，提升语篇的沟通效率（部寒、王立非 2021）。两家公司报告中冠词和介词的使用比例相当，无显著差异。因此，从宏观语言特征数据可知，两家公司 CSR 报告文本的词汇复杂度与可读性对情感关系建构无显著影响。

表 2 华为和 BT 公司 CSR 报告文本宏观特征统计

公司名称	基本统计信息			语法特征指标			
	篇均词数	篇均每句词数	大于 6 个字母单词比例 (%)	篇均代词比例 (%)	篇均冠词比例 (%)	篇均介词比例 (%)	篇均连词比例 (%)
华为公司	18,010	24.26	37.10	4.97	5.80	13.86	6.38
BT 公司	21,826	23.15	29.29	9.53	5.30	15.14	5.67

3.2 研究工具

LIWC-22 情感词典是由美国德州大学奥斯汀分校心理学系 Pennebaker 教授等人在 2015 版本基础上最新迭代的语言探索与字词计数文本分析工具，内部包含基于 Python 的自然语言处理程序和情感词典（Boyd *et al.* 2022；Pennebaker *et al.* 2022）。前人相关研究（吴育锋等 2018；苏悦等 2019）发现，LIWC-22 情感词典可较为精确地识别文本中的情感和写作者的心理过程。抽样检测发现，LIWC-22 情感词典对情感词汇的评分与人工主观处理各种写作片段节选的情感评分是一致的（张信勇 2015），可以保证研究的信度和效度。

LIWC-22 情感词典可以从语言特性类别（如人称代词、助动词等）、心理特性类别（如正负向情绪词、情感历程词等）、副语言学类别（如停止词、暂定词等）

等 117 个维度提供语言统计参数，笔者选取了其中 16 项最能反映 CSR 报告文本与外部意识形态互动关系的情感元素指标，运用单因素方差分析判断中外报告两组数据间是否具有显著性差异，并从批评话语分析视角阐释差异背后的动因。选取指标如表 3 所示。

表 3 本研究选用 LIWC-22 指标

分类		词语示例
人称代词	第一人称单数代词	I, me, my, myself
	第一人称复数代词	we, us, our, lets
	第二人称代词	you, your, yourself
	第三人称单数代词	he, she, her, his
	第三人称复数代词	they, their, them
情感过程词	正向情感词	love, excellent, nice
	负向情感词	bad, ugly, horrible
	焦虑词	worried, anxious
	生气词	angry, mad, irritated
	悲伤词	sad, crying, distress
社会历程词	礼貌词	thank, please, thanks, good morning
	交流词	said, say, tell, thank
	家族词	parent, mother, father, baby
	朋友词	friend, boyfriend, girlfriend, dude
	男性词	he, his, him, man
	女性词	she, her, girl, woman

3.3 研究步骤

研究步骤如图 1 所示。首先，从华为公司和 BT 公司官网下载搜集 CSR 报告 pdf 版，通过 OCR 识别转为纯文本格式，使用 Python 清洁去噪，建成语料库。然后，将语料导入 LIWC-22 情感词典获取情感指标数据，再借助 SPSS 24.0 进行单因素方差分析，判断两组数据是否具有显著性差异。最后，从批评话语分析视角阐释差异的动因。

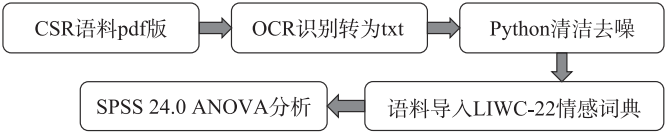


图 1 研究步骤

4 结果与讨论

4.1 人称代词统计分析

话语中的人称代词可以体现出作者对读者的态度 (Carter 2001)，是体现文本与社会情感关系的重要指标。统计结果 (见表4) 显示，华为与BT公司报告在人称代词使用总量上差异显著 ($F=308.5$, $P<0.01$)，效应量 η^2_p 为0.942。比较均值发现，BT公司报告中的人称代词使用数量远多于华为公司，差异主要体现在第一人称复数代词 *we* 和第三人称复数代词 *they* 两项指标上。

表4 华为与BT公司英文CSR报告人称代词使用差异

人称代词	华为公司		BT公司		F 值	P 值	效应量
	均值	标准差	均值	标准差			
人称代词	4.97	0.48	9.53	0.67	308.458	0.000	0.942
第一人称单数	0.05	0.03	0.09	0.06	3.415	0.081	0.113
第一人称复数	2.36	0.44	5.63	0.47	260.406	0.000	0.932
第二人称	0.03	0.02	0.05	0.03	2.687	0.119	0.082
第三人称单数	0.02	0.02	0.03	0.03	2.406	0.138	0.069
第三人称复数	0.53	0.07	1.01	0.19	54.754	0.000	0.739

从表4可以看出，BT公司报告中第一人称复数代词 *we* 的使用频次高于华为公司报告。*we* 具有内包和外排两种用法，既可指代报告撰写者和公司本身，也可指代包含读者在内的所有人 (潘峰、黑黹 2017)。相较于使用 *the company*，第一人称复数代词 *we* 可以增强话语的符号性、亲和性和群体性，降低公司因发布报告而受到指责的风险，而前者则显得更为官方，距离感强 (Aiezza 2015)。华为公司报告中 *we* 的出现频次相对较低，可能由于撰写者为非英语本族语者，权威度及对语言把握的精准度不及英语本族语者高。

使用 Wordsmith 6.0 对两家公司报告中 *we* 的高频搭配动词进行统计发现 (如图2所示)，*we* 的高频搭配动词中两家公司报告共有的词汇有 *work*、*use*、*help*，表达了公司承担社会责任的能动性，后接的宾语也多与企业的战略、政策、管理和绩效相关。

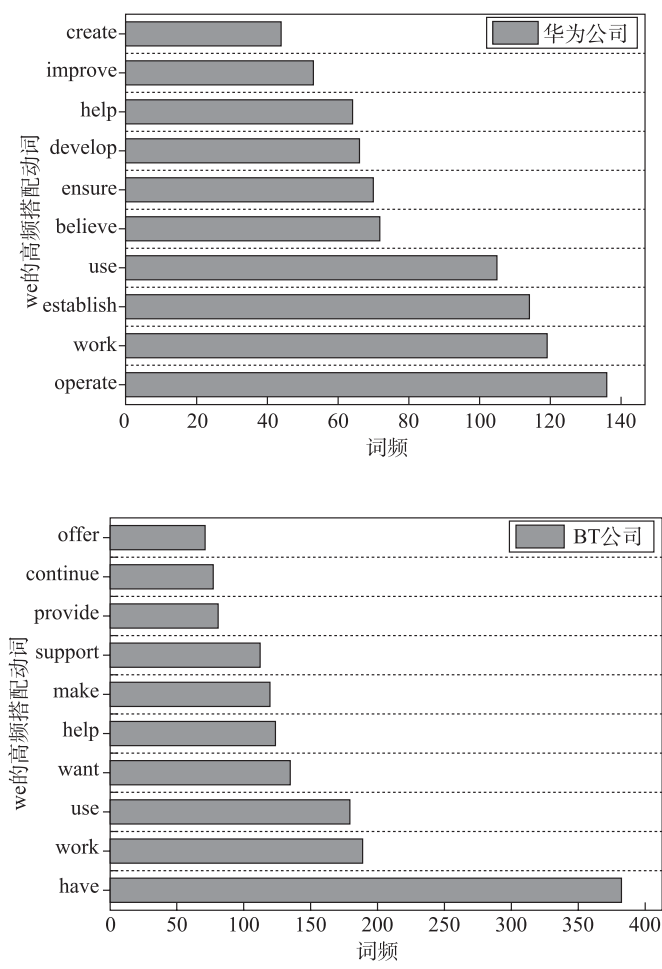


图2 we 的高频搭配动词

华为公司报告中we的高频搭配动词有operate、establish、believe、ensure、develop、improve、create，这些动词语义积极且明确，后接宾语通常为问题的具体解决措施，展现出华为公司执行力强、做事有决心有魄力、勇于直面问题的企业形象，给读者留下深刻的印象。operate的宾语主要包括system、telecommunication、risks、partners，与establish高频搭配的宾语有management、system、network等，believe常与collaboration、suppliers、technology、ICT、success、knowledge等词搭配，develop的宾语主要为products、management、solutions、energy，与improve搭配的宾语有efficiency、sustainability、management、capabilities，create常搭配的宾语为value、business、opportunities、jobs等。从中可以看出，华为公司作为高科技企业注重与供应商、客户等合作伙伴间长期关系的维护，并通过在企业内部建立健全高效的管理体系，提升产业能效，提高产品应用价值，向外部提供就业机会，发挥社会效能，增强与社会的互动。同时，报

告中使用了较多具体实例来支撑观点，描述公司回馈社会的行为，塑造了踏实可靠的企业形象。

BT公司报告中we的高频搭配动词有have、want、make、support、provide、offer等，这些词的意义相对抽象笼统，采用一种较为平和的语气描述公司与社会各方的供需关系。have通常与impact、access、targets、emissions等宾语搭配，support的高频搭配宾语有people、charities、customers、services等，provide常搭配的宾语有training、information、products、guidance等。We在BT报告中的高频使用展现了该公司鲜明的立场，即作为商业主体为外部社会中的客户、雇员、政府、供应商和社区等利益相关者提供价值。需要指出，charity和charities在BT公司报告中高频使用，共计309次，报告中大量描述了公司在慈善方面的所作所为，可能与发达资本主义国家慈善体制起步较早且更为成熟完善有关。此外，BT公司报告中they的使用频次高于华为公司报告，两者差异显著（ $P<0.05$ ）。They通常指代上下文所涉及的公司及读者以外的第三方，可能是某个相关群体，也可能是某类事件，具体要根据上下文语境进行判断。分析索引行发现，两家公司报告中的第三人称复数代词they后面多搭配情态动词。华为公司倾向于搭配need、should、could等中、低量值情态动词，语气委婉，态度平和，情感距离较近，使读者更易接受报告中的观点和主张。通过检索they could我们还发现，华为公司报告中高频使用了so that they could的用法，译为“以便于他们能够”，这一结构的使用连结了华为公司的具体行为及其所产生的社会价值与贡献。例如：

（1）Additionally, we donated 3,000 mobile phones with long-life batteries to victims **so that they could** connect to the outside world.（选自华为英文CSR报告2013）

（2）Huawei also donated 100 mobile phones to local villagers **so that they could** instantly enjoy convenient mobile services.（选自华为英文CSR报告2014）

（3）Huawei also supported online meetings between Chinese and Italian hospitals **so that they could** discuss their respective experience with the disease.（选自华为英文CSR报告2020）

索引行中华为公司为受灾群众和落后地区居民捐献手机产品，提供技术解决方案，心系慈善事业，为中意友好医院搭建互联网络，促进医学的国际交流与合作，华为公司报告的话语与外部社会建立了紧密的情感关系，同时体现出企业求真务实的开拓精神与心系人类的世界人文关怀。BT公司报告中人称代词they与can、will、must等高量值情态动词搭配的情况出现较多，可能与该公司在报告中提出了合作者需要遵守的规则等内容有关。

4.2 情感历程词统计分析

情感历程词可以反映CSR报告中的情感倾向。作者在表达积极的观点态度时使用正向情感词，表达消极的情感态度时使用负向情感词。情感词汇的出现频率及表达方式可以反映公司对相关事件的态度及文本与外部的互动关系（见表5）。

表5 华为与BT公司英文CSR报告情感历程词使用差异

情感历程词	华为公司（n=10）		BT公司（n=10）		F值	P值	效应量
	均值	标准差	均值	标准差			
正向情感词	5.38	0.87	4.71	0.76	48.312	.041	.063
负向情感词	0.79	0.07	1.95	0.29	50.968	.038	.051
焦虑词	0.25	0.06	0.67	0.15	46.321	.045	.068
生气词	0.11	0.03	0.19	0.05	0.025	.888	.001
忧伤词	0.13	0.03	0.15	0.04	0.284	.601	.016

从纵向比较视角来看，华为与BT公司报告中的正向情感词显著多于负向情感词，说明公司均期望以发布CSR报告为契机，用文本中积极的话语情感资源引起潜在投资者共鸣，维护公司与外部良好的社会关系，进而收获更大的企业效益。从横向比较视角来看，华为公司报告中的正向情感词多于BT公司报告，负向情感词少于BT公司报告。GRI《可持续发展报告指南》规定了编制可持续发展报告时的中肯性原则，报告中不仅要传递积极信息，也要客观陈述消极信息和需要改进的方面。数据表明，BT公司在这一原则上做得相对全面，报告中不仅介绍了公司对外部社区所做的贡献，还集中介绍了公司的战略风险、所面临困境以及未来挑战。华为公司报告则更多披露正面情感信息，负面情感信息量远小于正面情感信息量。不过这种情况在2016年之后的报告中有所改变，情感倾向变为相对中立。

两家公司报告中的负向情感词差异主要体现在焦虑词的使用上。面对全球市场的剧烈变化及疫情影响，公司在不同企业文化引领下的应对态度也有所不同。华为公司在面对全球经济放缓及激烈的竞争时表现出了不畏惧、不退缩、直面困难的积极态度，字里行间的焦虑情绪在措辞表达上相对平和。而BT公司近年间面临传统企业的现代转型，公司管理体系采取了一系列重大战略举措，改革发展的阵痛期势必会带来部分焦虑。华为公司报告中正向情感词的高频使用可能与其企业文化相关，公司长期倡导以正向积极的精神文化引领员工进步和行业发展，这种基调也体现在其商业话语中。我们以good、bad、sad等典型情感词为检索项，在语料库中随机提取并分析索引行。例如：

(4) As a provider of the infrastructure for the digital age, Huawei *believes* that as ICT applications become *easier* to use, *more convenient*, and *more affordable*, they will greatly reduce global inequality, bridge the digital divide, and drive the rapid attainment of SDGs. (选自华为英文 CSR 报告 2019)

(5) We want to make sure that we work ethically to fulfil our responsibilities to our stakeholders and *play a positive role* in society. It makes *good* business sense too. (选自 BT 公司 CSR 报告 2019)

(6) Waste *is bad for* the environment and *bad for* business. We want to contribute to an economy that drives down waste and uses resources again and again. (选自 BT 公司 CSR 报告 2020)

分析发现，华为公司对困难的表述方式更委婉，使用了较多比较级，以面向未来的姿态展现出迎难而上、解决困难的决心。而 BT 公司报告对正面与负面信息的用词更为直接，这可能是受到中英不同传统文化价值观影响的结果。中国传统文化强调集体主义，以“和”为贵，追求中庸之道，对情感的表达偏向含蓄深沉，而英国传统文化强调个人主义，情感表达较为热烈直接。

4.3 社会历程词统计分析

社会历程词可以体现话语与社会的情感互动关系。LIWC-22 情感词典中的社会历程词指标能反映出报告在家庭、友谊、性别、礼貌程度等主题方面的比重。对比华为与 BT 公司报告的社会历程词指标发现，两家公司报告中的社会主题元素在整体上具有显著性差异 ($P < 0.05$)，BT 公司报告中的社会历程词出现频次多于华为公司报告，如表 6 所示。

表 6 华为与 BT 公司英文 CSR 报告社会历程词使用差异

社会历程词	华为公司 (n=10)		BT 公司 (n=10)		F 值	P 值	效应量
	均值	标准差	均值	标准差			
社会词	6.52	0.45	12.68	0.99	316.052	.000	.946
家庭词	0.03	0.01	0.13	0.07	18.184	.000	.503
朋友词	0.21	0.04	0.36	0.24	4.248	.045	.241
女性词	0.10	0.03	0.08	0.04	3.604	.074	.167

(待续)

(续表)

社会历程词	华为公司 (n=10)		BT公司 (n=10)		F 值	P 值	效应量
	均值	标准差	均值	标准差			
男性词	0.03	0.02	0.04	0.03	1.515	.234	.078
礼貌词	0.31	0.06	0.46	0.26	1.659	.201	.082
交流词	1.62	0.08	1.51	0.06	3.986	.031	.189

具体而言,差异主要体现在家庭词、朋友词、礼貌词和交流词四项指标方面。纵向比较来看,两家公司报告中的交流词如say、tell、thank均值最大,出现频次高。以thank为检索词分析华为报告语料库,该词后多接宾语Huawei,具体语境为直接引述供应商、员工等相关者的语言来表达对公司为社会所作贡献的感恩。BT公司在家庭主题上更为突出,尤其是在提及BT family的概念上,把企业看作一个团结的大家庭进行人性化管理,重视企业内部团体建设和多元文化建设,增强员工的认同感和自豪感,帮助员工与企业共同成长。在全球三年疫情的时代背景下,华为公司报告更是着墨于描述特殊时期对员工健康的重视,例如为外派员工及随行家人额外增购传染病医疗保险,展现出充满人文关怀的企业形象。如例(7)和例(8)。

(7) With large numbers of people practicing social distancing, self-isolation or confined to their homes during the Covid-19 pandemic, installations become more important than ever to ensure that those with low or no digital skills can keep in touch with **family** and friends, and access vital health services. (选自BT公司CSR报告2020)

(8) We gave medical insurance for COVID-19 to **family members** accompanying our expatriate employees; increased the sum insured for employees affected by work-related or infectious diseases. (选自华为英文CSR报告2020)

BT公司报告中使用的朋友词多于华为公司报告,在BT报告语料库中检索friend/friends/-friendly,出现的搭配有friends and family、user-friendly、child-friendly、environmentally-friendly、family-friendly等。可以看出,BT公司的工作不仅包含常规的慈善和环保等内容,还几乎涵盖了所有的利益相关者,如客户、员工、社区、外部环境等,话语中体现出了公司与社会建立的多维主体情感关系。同时,研究结果也呼应了BT公司在社会责任领域独具特色的关注重点,例如公司十分重视

对少年儿童的帮助，为中学生提供免费的教育资源，并与“儿童热线”机构合作，免费向其电话咨询业务提供技术支撑服务等。

5 结论

本文基于中外跨国公司英文CSR报告语料库，借助情感词典分析了华为公司和BT公司CSR报告中体现的情感因素差异及背后的意识形态动因。研究发现，差异主要体现在以下方面。（1）在人称代词使用上存在显著差异。BT报告中we和they的使用频次显著高于华为报告，体现出BT公司积极承担社会责任的鲜明情感立场及与合作者战略共赢的情感态度，缩短了公司与文本读者间的情感距离。通过分析we和they的相关搭配和索引行发现，华为公司更加注重主体与供应商、客户等合作伙伴间的长期情感维系。（2）在正负向情感历程词使用上存在显著差异。BT报告中的正负向情感词使用较为均衡，华为前期报告中的正向情感词使用较多，后期则趋于减少。此外，两家公司报告中关于阐述应对困难的态度具有差异，华为公司表述较为委婉，BT公司则相对直接，这可能受到中英两国“集体主义”和“个人主义”的不同传统文化价值观影响。（3）相比于华为报告，BT报告内容所涉及的社会元素更加丰富，与外部构建了多维主体情感关系，华为报告中则突出了疫情时代下的人文关怀。

本研究可为我国跨国公司英文CSR报告的撰写提供参考，为商务英语写作教学和企业形象塑造提供借鉴。公司在撰写面向全球发行的英文CSR报告时，可参照国外著名跨国公司同类报告的语言特征，充分考虑目标语读者的行文习惯和文化背景，结合当地实际情况有效传递信息，与外部各方实现恰当的情感构建，增强企业国际竞争力。应当指出，本研究基于LIWC-22情感词典的数据指标，其中情感词的抽取和判别主要是依据词语间的语义联系，较为依赖种子词的数量和质量，个别一词多义也可能会带来噪音（赵妍妍等 2010），未来可以使用R、Python等开源编程语言对基于规则的情感语义网络进行考察，评价单位也可进一步扩展。

参考文献

- AIEZZA C M. Corpus-assisted discourse analysis of modality markers in CSR [J]. *Studies in Communication Sciences*, 2015(1):68-76.
- BOYD R, ASHOKKUMAR A, SERAJ S, PENNEBAKER J. The development and psychometric properties of LIWC-22 [EB/OL]. (2022-04-18) [2023-03-21] <https://www.liwc.app>.
- CARTER R. Working with texts [M]. Second Edition. London: Routledge, 2001.
- FUOLI M. Assessing social responsibility: a quantitative analysis of appraisal in BP's and IKEA's social reports [J]. *Discourse & Communication*, 2012, 6(1): 55-81.

- HAJEK P, OLEJ V, MYSKOVA R. Forecasting corporate financial performance using sentiment in annual reports for stakeholders' decision-making [J]. Technological and Economic Development of Economy, 2014, 20(4): 721-738.
- PENNEBAKER J, BOYD R, BOOTH R, ASHOKKUMAR A, FRANCIS M. Linguistic inquiry and word count: LIWC-22 [CP]. Pennebaker Conglomerates, 2022. <https://www.liwc.app/>.
- 部寒, 王立非. 基于语料库的中美企业财务语篇可读性对比分析[J]. 解放军外国语学院学报, 2021 (1): 71-78.
- 李涵昱, 钱力, 周鹏飞. 面向商品评论文本的情感分析与挖掘[J]. 情报科学, 2017 (1): 51-55.
- 李战子. 评价理论在国际传播语境中的应用与拓展[J]. 外语研究, 2022 (2): 1-6.
- 潘峰, 盛丹丹. 记者招待会汉英口译中的规范及选择——从模糊限制语的翻译谈起[J]. 外语教学理论与实践, 2021 (1): 115-125.
- 潘峰, 黑黝. 新闻发布会汉英口译中的政府形象构建——以人称代词we的搭配词为例[J]. 外语与外语教学, 2017(5): 45-51.
- 邵珊珊, 王立非. 基于语言大数据挖掘的电商英汉评价话语情感分析[J]. 外语电化教学, 2019 (5): 76-84.
- 苏悦, 谢宇雪, 杜晓扬, 等. 基于毛姆小说的20世纪初英国群体心理分析[J]. 心理技术与应用, 2019 (10): 597-604.
- 王立非, 部寒. 中美企业话语情感倾向多维评价测量与对比分析[J]. 外语研究, 2017 (4): 16-21.
- 王琴. 基于语料库的美国媒体中国人口话语建构研究[J]. 语料库语言学, 2022 (2): 109-126.
- 吴育锋, 吴胜涛, 朱廷劭, 等. 小说人物性格的文学智能分析: 以《平凡的世界》为例[J]. 中文信息学报, 2018 (7): 128-136.
- 奚金金, 霍欢, 徐亚. 文本情感分析在网购评论中的应用前景[J]. 信息技术, 2013 (12): 71-74.
- 谢德仁, 林乐. 管理层语调能预示公司未来业绩吗?——基于我国上市公司年度业绩说明会的文本分析[J]. 会计研究, 2015 (2): 20-27.
- 张钊炜. 基于评价系统的评论类文本情感倾向性分析[J]. 语言文字应用, 2018 (2): 138-144.
- 张信勇. LIWC: 一种基于语词计量的文本分析工具[J]. 西南民族大学学报(人文社会科学版), 2015 (4): 101-104.
- 赵妍妍, 秦兵, 刘挺. 文本情感分析[J]. 软件学报, 2010 (8): 1834-1848.

通信地址: 710128 陕西省西安市 西安外国语大学外国语言文学研究院

学术论文中互动元话语及作者身份建构对比研究*

河南师范大学 刘国兵 张君兰

提要：本研究基于Hyland（2005a）的互动元话语理论和孙莉（2015）的身份建构类别，探讨了中国硕士生与国际期刊作者在应用语言学学术论文中互动元话语的使用特征及运用互动元话语建构身份的异同。研究发现，除介入标记外，中国硕士生在学术写作中使用的互动元话语均少于国际期刊作者。从出现频数来看，两类作者运用互动元话语建构的身份类别从多到少依次为研究者、对话者和评价者。中国硕士生建构的小心建议者身份与国际期刊作者不存在差异，但在自发互动者、他发互动者、评我者、评他者、谨慎发起者和自信研究者等身份建构方面与国际期刊作者存在显著性差异。研究结果可为学术英语写作教学提供有益启示。

关键词：互动元话语、作者身份建构、学术论文

1 引言

学术英语写作中，作者运用各种语言资源进行信息传递和自我表征，因此学术文本被视为作者建构身份的重要方式。换言之，作者可以通过语言资源进行学术语篇中的个人身份建构。Hyland（2001）认为，学术文本中作者身份的呈现或隐藏会影响读者对文本的印象，甚至会影响读者对文本的理解。作为学术文本中一种重要的语言资源，元话语泛指关于话语的话语，能够反映出作者对语篇理解的元语用能力（姜晖 2011）。关于元话语的定义和分类，国内外学者持有不同观点。国外学者从社会语言学（Schiffrin 1980）、修辞学（Crismore 1989）、系统功能语言学（Vande Kopple 1985；Hyland 2005a）等方面进行深入探讨；国内学者则从语境构建（李佐文 2001）、元功能（杨信彰 2007）以及多维视角（柳淑芬 2013）等方面阐述元话语的理论基础。元话语的相关讨论中，Hyland（2005a）的人际元话语和互动元话语分类受到学界广泛认可，他提出元话语是与读者共同协

* 本文系河南省高等教育教学改革研究与实践项目“基于OBE理念的英语专业学生写作教学模式创新研究”（2021SJGLX107）的阶段性成果。

刘国兵为本文通讯作者。

作者贡献：

刘国兵：选题构思、研究方法、讨论结论、字数占比（40%）、修改润色；

张君兰：数据收集、数据分析、讨论结论、初稿撰写、字数占比（60%）。

商互动并表现作者自我态度的一种语言形式，用于传递交际意图，实现文本的协商性和互动性。

关于互动元话语与身份建构的关系问题，已有学者讨论了自我指称语对建构作者身份的重要性（Kuo 1999；Tang & John 1999；娄宝翠、王莉 2020），发现学术语篇中自我指称语的恰当使用能够凸显作者身份建构，彰显学术地位。但是其他互动元话语资源对身份建构有何影响，同样值得深入探讨，因此从身份视角对互动元话语进行讨论仍存在很大空间。本研究将以中国硕士生论文和国际期刊论文为研究语料，考察中国硕士生和国际期刊作者的互动元话语使用特征以及两类作者运用互动元话语建构身份的异同，进而从身份视角丰富互动元话语研究。

2 相关文献回顾

2.1 互动元话语

互动元话语指作者评价和读者介入文本的手段，能够用于加强语篇的人际互动。Vande Kopple（1985）根据篇章组织和人际互动功能提出了篇章和人际两类元话语，其中人际类元话语表达作者对文本信息的态度，展示与读者间的互动。在此基础上，Hyland（2005a）提出了引导式和互动式两大类。前者指作者组织构建语篇的手段；后者体现的是作者对命题内容的立场和评价，更能够显示作者与读者之间的互动关系（Hyland 2012），主要包括模糊限制语、加强词、态度标记、介入标记和自我指称语五个子类。

在学术语篇中，作者可以借助互动元话语向读者呈现文本的命题意义，或表达自己的立场，或劝说读者接受自己的观点。就其子类而言，模糊限制语暗示陈述是基于作者的合理推理而不是某种知识，表明作者对观点的谨慎态度。通过模糊语，作者可以展示自己在学术文本中的参与度，更能劝说读者接受作者的命题（Hyland 2005a: 52）。与模糊限制语相反，加强词可以突出作者的认知立场，促进与读者的稳固关系，增加作者对命题的信心和确定程度（Hyland 2005b；Hu & Cao 2015）。态度标记体现作者对文本命题的评价或态度，在加强作者论点的说服力方面发挥关键作用（Hyland 2005a）。介入标记描述作者与读者间的互动协商，提供与读者间对话的开放空间。Hyland（2005a: 151）认为介入标记指的是作者将读者带入文本，与他们保持互动，并预测他们可能提出的各种反对意见。这意味着作者通过预测读者对命题的可能反应以及与文本对话的成员间的学科社区来建构与读者间的互动（Jiang & Ma 2018）。从本质上讲，学术参与是建立在作者意识到读者总是可以反驳主张的基础上的，表明读者在作者如何构建他们的论点方面具有积极的和构成性的作用（Hyland & Jiang 2016）。与介入标记不同，自我指称语则能够帮助作者彰显自己在文本中的地位，建立并实现与读者间的互动。

Hyland & Jiang (2017) 提出作者在学术写作中可以运用自我指称语(第一人称代词)建立个人身份,突出在学术社区中的地位。因此,通过自我指称语建立明确的学术身份有助于彰显作者的贡献,提高作者在自身研究领域的权威性和可靠性(Hyland 2001, 2002)。

关于互动元话语的研究,主要集中在跨语言 and 不同母语背景对比、跨学科对比以及不同学术语篇对比等方面。首先是跨语言和母语背景视角的学术文本中的互动元话语研究。Lee & Deakin (2016) 考察了成功和不成功的中国学生议论文在使用立场和介入资源方面的差异,并对比了二语议论文与一语成功议论文的互动元话语使用差异。郭骅、马磊(2016)以社会学期刊的英文摘要为语料,考察了不同母语背景和不同研究方法对作者使用模糊限制语与增强语的影响。其次是跨学科视角的互动元话语研究。Hu & Cao (2015) 分析学科和范式对互动元话语使用的影响,发现了特定互动元话语资源在跨学科和跨范式中的使用差异。王晶晶、吕中舌(2016)通过对比理工科博士论文和期刊论文,发现博士生在模糊限制语的使用及具体词汇选择上存在差异。Liu & Yang (2021) 对软硬学科研究论文中的元话语进行了历时研究,结果表明两类学科作者为呈现更明确客观的文本,使用的交际资源较多而互动资源较少。最后是不问学术语篇中的互动元话语研究。Wu & Paltridge (2021) 以互动元话语的立场表达为切入点,讨论了中国硕士生和博士生的应用语言学论文在立场资源上的使用差异。徐昉(2015)对比了不同水平英专生的学位论文和期刊论文中的立场标记语,发现学习者总体上显著少用体现作者立场的语言表达。

2.2 身份建构

学术写作中,作者通过投射一种带有个人权威的身份,对自己的主张表现出信心,从而获得可信度。Ivanič (1998) 指出作者在学术社区内利用不同的话语策略协商自我身份,并从三个维度(自传性身份、语篇身份、自我身份)详细探讨了作者身份。自传性身份强调作者以往经历对文本的影响;语篇身份指作者运用一定的话语策略有意识或无意识地向读者传递信息;自我身份指作者在文本中所持的立场、观点或看法。这三种身份相互联系、相互影响,共同构成作者身份。孙莉(2015)从语用学角度出发,将元话语建构的身份分为组织者、互动者和评价者三类,指出三者分别起到语篇组织、人际互动和评价研究的作用。作者在学术论文写作中能意识到语言资源对构建作者身份的重要性,这对传递作者的观点态度和突出作者身份具有推进作用。

学术英语写作中,作者身份建构的相关研究已取得丰硕成果,主要呈现出两个特点。一是集中于对第一人称代词等作者自称语的探讨。例如,Kuo (1999) 探讨了人称代词如何揭示作者在研究中的角色,以及该类词与预期读者和科学学

术团体的关系。他还发现第一人称复数代词的使用频率远高于其他类型的人称代词。Tang & John (1999) 分析了新加坡本科生论文中的作者身份, 认为第一人称代词是作者在文本中最明显的表现形式。Hyland (2002) 考察了中国香港地区本科生二语论文中人称代词的使用, 发现学生未能明显体现作者指称, 而且在提出个人主张时避免使用这些形式。李民、肖雁 (2018) 以学术语篇中的第一人称代词为研究对象, 分析了国内学者与本族语者的使用差异及其建构作者身份的异同。娄宝翠、王莉 (2020) 考察了硕士生与期刊作者对自我指称语的使用情况及身份建构特征, 发现学习者在借助自我指称语呈现作者身份方面还存在不足, 对建构作者身份的其他互动元话语资源的关注不够。Rahimivand & Kuhi (2014) 讨论了应用语言学期刊论文中言据标记、模糊语、加强词、态度标记和自我指称语对作者身份在建构作用。孙莉 (2020) 通过对比硕士生和期刊作者对元话语的使用差异及身份建构特点, 发现中国硕士生运用元话语建构身份的频率明显少于期刊作者。

上述研究表明, 互动元话语的使用受到语言特征和语言文化背景的影响, 中国学习者对互动元话语的使用及其身份建构与本族语者间存在差异。已有研究证明了自我指称语 (主要是第一人称代词) 对作者身份在建构至关重要, 但与自我指称语的身份建构研究相比, 对其他互动元话语资源的身份建构问题的探讨尤显不足。此外, 尽管已有学者对不同水平学习者的互动元话语使用进行实证分析, 但该类研究多聚焦于博硕士学位论文的对比, 缺乏与国际期刊的对比研究。因此, 本研究结合 Hyland (2005a) 的互动元话语分类以及孙莉 (2015) 的身份类别, 探讨中国硕士生与国际期刊作者互动元话语的使用及运用其建构身份的差异, 深入了解中国硕士生的互动元话语使用情况。

3 研究设计

本研究自建两个语料库, 包括中国硕士论文语料库 (Master Theses Corpus, 简称MTC) 和国际期刊论文语料库 (International Journal Articles Corpus, 简称IJAC)。语料数据均选自2016—2020年发表的应用语言学实证性论文, 其中中国硕士论文选自CNKI, 而国际期刊论文选自Web of Science发布的该领域影响因子排名前6的国际期刊 (包括*Applied Linguistics*、*The Modern Language Journal*、*Language Learning*、*Language Teaching Research*、*English for Specific Purposes*、*TESOL Quarterly*)。为保证两个语料库的可比性以及自我指称语判定的准确性, 所选论文均由单一作者撰写, 分别选取20篇硕士论文和40篇期刊论文, 库容分别为316,417和319,782。本研究主要回答两个问题: 1) 中国硕士生与国际期刊作者在英语学术写作中, 互动元话语的使用分别有何特点? 2) 两类作者运用互动元话语建构的身份类型存在何种差异?

研究步骤如下。第一, 根据 Hyland 的分类标准, 判定和标注互动元话语。第二, 依据身份建构类别, 对第一步标出的互动元话语进行身份判定。第三, 将所得数据进行标准化处理, 详细分析中国硕士生与国际期刊作者使用的互动元话语及建构身份的差异。为了保证语料标注的可靠性和准确性, 本研究采取两人交叉验证的方法, 最终结果保持高度一致。需要指出的是, 由于本研究强调的是互动元话语的身份构建, 因此孙莉 (2015) 提出的组织者身份不包含在本文的分析框架中。基于研究目的, 作者提出了研究者身份类别。本研究还借鉴 Martin & White (2005) 对态度系统的分类, 将态度标记细分为情感标记、评判标记和鉴别标记。前人对自我指称语的研究中, 往往将第一人称复数代词 *we* 直接归为自称语, 而本研究将结合具体语境对第一人称复数代词建构的身份类别进行详细判定。

4 结果与讨论

4.1 各类互动元话语的频数特征

表1显示, 国际期刊论文中互动元话语的总频数明显高于中国硕士论文 (Loglikelihood=181.5, $P<0.01$), 这表明国际期刊作者在学术写作中比中国硕士生使用更多的互动元话语, 他们更有意识地使用语言和修辞手段与读者建立积极的关系 (郭骅、马磊 2016)。就互动元话语的五个子类而言, 两组作者使用的模糊限制语存在显著差异 (Loglikelihood=164.1, $P<0.01$)。模糊限制语是使用最多的互动元话语资源, 表明硕士生和期刊作者均倾向于选择模糊语来谨慎地表达自己的观点, 增加命题的可信度, 并为读者提供开放的对话空间。但是, 中国硕士生对模糊语的掌握程度不及国际期刊作者, 可能是因为教师经常给学生一种错误的印象, 即英语学术写作需要直接的表述, 使用模糊语反而会削弱自己的观点 (Wishnoff 2000), 导致学生不敢过多使用该类互动元话语。Loglikelihood 检验还证明了两组作者在使用加强词方面存在显著差异 (Loglikelihood=47.4, $P<0.01$)。这一结果与 Qiu & Ma (2019) 的研究不一致, 他们发现应用语言学领域的硕士生比博士生和专家作者运用更多的模糊语和加强词, 表明他们更尊重自身研究领域的主张, 因此在提出主张时犹豫不决。这些差异可能是因为国际期刊作者对命题的认识态度比中国硕士生更明确, 他们似乎更喜欢用加强词来证明自己的肯定声音, 充分突出其研究成果的创新性。中国硕士生在阅读文献和专著时, 更倾向于参考中文撰写的文献 (Lei & Jiang 2019), 虽然他们能接触到一定的学术话语, 但英语文献输入远远不够, 造成了与国际学术社区在合理用词方面的差距。

表1 IJAC和MTC中互动元话语的频数比较

互动元话语	IJAC		MTC		Loglikelihood	P 值
	原始频数	标准频数	原始频数	标准频数		
模糊限制语	1,584	49.5	932	29.5	164.1	0.00
加强词	935	29.2	653	20.6	47.4	0.00
态度标记	539	16.9	253	8.0	102.7	0.00
评判标记	38	1.2	20	0.6	5.5	0.02
情感标记	141	4.4	74	2.3	20.5	0.00
鉴赏标记	360	11.3	159	5.0	77.8	0.00
介入标记	430	13.5	728	23.0	-80.7	0.00
自我指称语	564	17.6	330	10.4	59.5	0.00
总计	4,053	126.7	2,896	91.5	181.5	0.00

此外，态度资源的分布在两个语料库间也存在显著性差异（Loglikelihood=102.7， $P<0.01$ ），这说明高水平作者在学术论文中更倾向于对命题或观点进行客观评价。在三类态度标记的使用上，研究发现，中国硕士生的整体使用情况与国际期刊作者一致：鉴赏标记最多，情感标记其次，评判标记最少。差异性检验显示，中国硕士生使用的情感标记和鉴赏标记均明显少于国际期刊作者，然而在评判标记的使用上二者不存在显著差异（Loglikelihood=5.5， $p>0.01$ ）。这一结果与孙莉（2020）的发现存在差异，她指出中国硕士生使用的评判标记和鉴别标记显著少于期刊作者。这可能是由语料选取的差异造成的，本研究选取了论文全文，而孙莉的研究仅选取了引言、文献、讨论、结论四部分。这一现象还可以解释为国内学界主张学术语篇应具备客观性，要求学术写作中作者要多评价研究本身，少用含个人情感或表示评判的词汇（孙莉 2020）。

表1还清楚地显示，两个语料库中的介入标记和自我指称语均存在显著差异（ $P<0.01$ ），该结果与孙莉（2020）的研究发现一致。这表明中国硕士生在与读者对话的过程中，倾向于采用引导他人介入的方法。中国硕士论文中介入标记的多用以及自我提及的少用还说明，硕士生受到客观写作风格的影响，可能不愿意以一种明确直接的或者个人的方式与对话者进行互动（Hyland & Jiang 2016）。仔细观察语料发现，中国硕士生常用指代作者与读者的we（inclusive we）引导读者介入文本。指代作者与读者的we表明一个共同的解释和共同的目标（Hyland & Jiang 2016），这可能会影响读者参与作者的论点，说服读者接受作者的观点。介

入标记在两个语料库中的出现频率表明中国硕士生已经意识到通过特定的文本特征与读者建立人际互动，并将读者视为目标受众，从而假定了地位较低的作者角色（Jiang & Ma 2018），因此他们多借助介入标记进行人际互动。此外，研究还发现第一人称代词 I/my 以及指代作者的 we（exclusive we）在两个语料库中高频出现，但硕士生的使用频数少于期刊作者。该结果与 Hyland（2002）和娄宝翠、王莉（2020）的结论一致，他们发现在学术写作中，英语学习者对第一人称代词的使用明显少于本族语者。这表明学习者倾向于向读者呈现自己的研究结果并希望获得接受，所以更容易选择较安全的语言表达（娄宝翠、王莉 2020）。这一现象还说明硕士生作为写作新手，刻意学习国际期刊社区的多位专家合著现象（Kuo 1999）。由于本研究直接将指代作者与读者的第一人称复数归为介入标记，所以该部分统计的自我指称语 we 的使用频率低于国际期刊作者，与娄宝翠、王莉（2020）的结果相悖，她们发现 we 在中国硕士论文和国际期刊中使用频率最高。虽然在学术训练中许多学生被教导要避免使用第一人称，但自我指称语在协商作者论点与话语社区之间的关系中扮演着重要的互动角色，帮助作者创建一个既是学科的仆人又是创造者的身份（Hyland 2001）。

4.2 各类互动元话语建构的身份类别特征

学术作者使用不同的语言策略在学术语篇中与读者互动，建构自己的作者身份（Ivanič 1998）。分析语料发现，硕士生和期刊作者在学术语篇中通过互动元话语来评价语篇，引发互动和体现作者角色。本研究结合孙莉（2015）对身份的研究，将互动元话语构建的身份类型分为评价者、对话者和研究者三类。之后，又对三类身份进行详细划分，通过区分提及自己或他人与读者互动，建构了自发互动者和他发互动者两类对话者身份；根据评价对象（自己或他人）的不同，建构了评我者和评他者两类评价者身份；根据研究者在语篇中的作用和需求，建构了谨慎发起者、小心建议者和自信研究者三类研究者身份。从表2可以看出，中国硕士生和国际期刊作者运用互动元话语建构的三类主身份的频数特征均是研究者最多，互动者次之，评价者最少。下文将对IJAC和MTC中互动元话语所建构的次要身份类别进行详细分析（见表2）。

表2 IJAC和MTC中互动元话语建构的身份类别比较

身份类别		IJAC		MTC		Loglikeli- hood	P 值
主类别	次类别	原始频数	标准频数	原始频数	标准频数		
互动者	自发互动者	473	14.8	314	9.9	30.7	0.00
	他发互动者	413	12.9	667	21.1	-63.0	0.00

（待续）

(续表)

身份类别		IJAC		MTC		Loglikeli- hood	P 值
主类别	次类别	原始频数	标准频数	原始频数	标准频数		
评价者	评我者	446	14.0	209	6.6	85.2	0.00
	评他者	84	2.6	41	1.3	14.7	0.00
	谨慎发起者	1,225	38.3	687	21.7	147.8	0.00
研究者	小心建议者	174	5.4	171	5.4	0.00	0.95
	自信研究者	898	28.1	630	19.9	44.5	0.00

第一,就互动者身份而言,中国硕士生运用互动元话语建构的互动者身份与国际期刊作者间存在差异,即硕士生引发与读者互动的能力存在不足。这可能是由于硕士生为了让读者接受自己的观点,避免体现个体身份,往往通过群体或学术团体来表达自己的观点。相反,国际期刊作者为了强调科研成果的创新之处,多通过自我身份凸显自身科研的突出成就,强调个人贡献。在次类别上,自发互动者和他发互动者在两个语料库中的使用均存在显著性差异(Loglikelihood=30.7, $P<0.01$; Loglikelihood=-63.0, $P<0.01$)。这一结果与孙莉(2020)的发现一致,她发现与国际期刊作者相比,中国硕士生建构的自发互动者身份较少,而他发互动者身份相对较多。正如Tang & John(1999)所解释的,学生作者认为自己处于学术等级的最底层,因此对自己的身份没有安全感。本研究还发现,中国硕士生建构的他发互动者与自发互动者身份间存在显著性差异(Loglikelihood=129.9, $P<0.01$),说明硕士生在建构对话者身份时,多通过他人介入而不是自我提及来与读者进行对话。相反,英语本族语者更倾向于使用自我指称语来向读者展示自身研究的创新性(李民、肖雁 2018)。

第二,就评价者身份而言,中国硕士生建构的评价者身份整体上显著少于国际期刊作者,表明硕士生不擅长运用态度标记建构身份。由于评价者是由鉴赏标记、情感标记及评判标记建构,因此这一结果与上节提到的中国硕士生较少使用态度标记的特征相吻合。在评价者的次类中,中国硕士生建构的评我者和评他者显著少于国际期刊作者(Loglikelihood=85.2, $P<0.01$; Loglikelihood=14.7, $P<0.01$),表明中国硕士生学术写作中较少表达自己的情感和态度,这与学术写作的规范有直接联系。本研究还发现中国硕士生建构的评我者与评他者之间也存在显著差异(Loglikelihood=123.5, $P<0.01$),说明在运用互动元话语表达评价时,中国硕士生倾向于评价自己的研究。这可以解释为,中国硕士生学术写作中较少使用带有个人情感和态度的词汇,而是侧重于通过数据或事实呈现学术观点,

这有助于体现严谨的学术作风。然而,国际期刊作者在与读者建立互动关系时,通过表达自己的情感态度以激起读者的积极回应,促进读者对观点的认同,引导学术声音和成果的推介。

第三,就研究者身份而言,模糊限制语用来谨慎地提出建议,建构了谨慎发起者和小心建议者两类身份。加强词强调作者对命题的肯定程度,帮助作者建立稳固的地位和明显的作者身份,建构了自信研究者身份。如表2所示,中国硕士生建构的研究者身份明显少于国际期刊作者,说明期刊作者比硕士生更善于使用模糊限制语和加强词来建构研究者身份。借助该身份,作者可以谨慎地提出自己的命题,为读者提供对话空间,或者对自己的命题表示肯定,说服读者接受命题。这一结果与Rahimivand & Kuhl (2014)的研究一致,他们发现应用语言学期刊论文中模糊限制语是最优先使用的立场标记,为作者的身份安全提供保障;加强词是第三受欢迎的元话语,能够强化作者身份的建构。本研究还发现两类作者运用模糊限制语建构的谨慎发起者身份存在显著差异(Loglikelihood=147.8, $P<0.01$),但是建构的小心建议者身份在频数上较为接近,不存在差异(Loglikelihood=0.00, $P>0.01$)。这可能是因为国际期刊作者为了强调研究结果的严谨性和合理性,似乎更注重谨慎地对其进行解释,而中国硕士生使用模糊限制语的意识还不是特别充分。模糊限制语是最主要的互动元话语类别,在构建身份时起决定性作用,作者需要在承诺自己的命题和尊重与读者的对话之间取得艰难的平衡(Rahimivand & Kuhl 2014),因此硕士生和期刊作者均较多建构谨慎发起者身份。还需要注意的是,中国硕士生建构的谨慎发起者显著多于小心建议者(Loglikelihood=332.4, $P<0.01$),再次证明了硕士生运用模糊限制语主要建构谨慎发起者身份,解释了模糊限制语能够表达作者对命题的可能性承诺或不确定性(Hyland 2005a),为读者介入文本留出空间。此外,研究还发现加强词所建构的自信研究者身份在两个语料库中也存在显著差异(Loglikelihood=44.5, $P<0.01$)。这与前文讨论的期刊作者使用加强词明显多于硕士生的结果一致,说明加强词在学术社区成员对作者学术能力的论证和评价影响下,能够强化作者身份的构建(Rahimivand & Kuhl 2014)。

5 结论

本研究以Hyland (2005a)的互动元话语模型以及孙莉(2015)的身份建构类别为基础,创建了中国硕士论文和国际期刊论文两个语料库,考察了中国硕士生与国际期刊作者在学术写作中互动元话语的使用以及运用互动元话语建构身份类别的差异。研究发现,中国硕士生在学术写作中使用的模糊限制语、加强词、态度标记和自我提及语均显著少于国际期刊作者,而使用的介入标记显著多于期刊作者,说明中国硕士生在学术写作中使用的互动元话语整体上不如国际期刊作者。此外,两类作者运用互动元话语建构的身份类别在学术论文中的出现频数从多到

少依次为研究者、对话者和评价者。就其次类而言,中国硕士生建构的自发互动者、他发互动者、评我者、评他者、谨慎发起者和自信研究者等身份均与期刊作者间存在显著差异,但建构的小心建议者身份无显著差异。我们认为这些差异与中国学生所接受的写作指导或中国学术界固有的写作模式有关。本研究对大学英语学术写作教学具有一定启发意义。

参考文献

- CRISMORE A. Talking with readers: metadiscourse as rhetorical act [M]. New York: Peter Lang, 1989.
- HU G, CAO F. Disciplinary and paradigmatic influences on interactional metadiscourse in research articles [J]. *English for Specific Purposes*, 2015, 39(3): 12-25.
- HYLAND K. Humble servants of the discipline? Self-mention in research articles [J]. *English for Specific Purposes*, 2001, 20(3): 207-226.
- HYLAND K. Authority and invisibility: authorial identity in academic writing [J]. *Journal of Pragmatics*, 2002, 34(8): 1091-1112.
- HYLAND K. Metadiscourse: exploring interaction in writing [M]. London: Continuum, 2005a.
- HYLAND K. Stance and engagement: a model of interaction in academic discourse [J]. *Discourse Studies*, 2005b, 7(2): 173-192.
- HYLAND K. Disciplinary identities: individuality and community in academic discourse [M]. Cambridge: Cambridge University Press, 2012.
- HYLAND K, JIANG F K. "We must conclude that ...": a diachronic study of academic engagement [J]. *Journal of English for Academic Purposes*, 2016, 24(4): 29-42.
- HYLAND K, JIANG F K. Is academic writing becoming more informal? [J]. *English for Specific Purposes*, 2017, 45(1): 40-51.
- IVANIČ R. Writing and identity: the discoursal construction of identity in academic writing [M]. Amsterdam: John Benjamins, 1998.
- JIANG F K, MA X. "As we can see": reader engagement in PhD candidature confirmation reports [J]. *Journal of English for Academic Purposes*, 2018, 35(5): 1-15.
- KUO C. The use of personal pronouns: role relationships in scientific journal articles [J]. *English for Specific Purposes*, 1999, 18(2): 121-138.
- LEE J J, DEAKIN L. Interactions in L1 and L2 undergraduate student writing interactional metadiscourse in successful and less-successful argumentative essays [J]. *Journal of Second Language Writing*, 2016, 33(3): 21-34.

- LEI J, JIANG T. Chinese university faculty's motivation and language choice for scholarly publishing [J]. *Ibérica*, 2019, 38: 51-74.
- LIU G, YANG Y. A diachronic study of multi-disciplinary metadiscourse in research articles [C] // ICDEL. 2021 the 6th International Conference on Distance Education and Learning. New York: Association for Computing Machinery, 2021: 121-132.
- MARTIN J, WHITE P. The language of evaluation: Appraisal in English [M]. New York: Palgrave Macmillan, 2005.
- QIU X, MA X. Disciplinary enculturation and authorial stance: comparison of stance features among master's dissertations, doctoral theses and research articles [J]. *Ibérica*, 2019, 38: 327-348.
- RAHIMIVAND M, KUHI D. An exploration of discoursal construction of identity in academic writing [J]. *Procedia Social and Behavioral Sciences*, 2014, 98(9): 1492-1501.
- SCHIFFRIN D. Meta-talk: organizational and evaluative brackets in discourse [J]. *Language and Social Interaction*, 1980, 50(3-4): 199-236.
- TANG R, JOHN S. The 'I' in identity: exploring writer identity in student academic writing through the first person pronoun [J]. *English for Specific Purposes*, 1999, 18 (suppl 1): s23-s39.
- VANDE KOPPLE W. Some explanatory discourse on metadiscourse [J]. *College Composition and Communication*, 1985, 36(1): 82-93.
- WISHNOFF J. Hedging your bets: L2 learners' acquisition of pragmatic devices in academic writing and computed-mediated discourse [J]. *Second Language Studies*, 2000, 19(1): 119-148.
- WU B, PALTRIDGE B. Stance expressions in academic writing: a corpus-based comparison of Chinese students' MA dissertations and PhD theses [J]. *Lingua*, 2021, 253(2): 1-18.
- 郭骅, 马磊. 中外社会学期刊论文摘要的人际互动元话语研究[J]. 西安外国语大学学报, 2016 (4): 39-43.
- 姜晖. 元语用视角下的功能性言语探究[J]. 当代外语研究, 2011 (4): 15-19.
- 李民, 肖雁. 英语学术语篇互动性研究——以第一人称代词及其构建的作者身份为例[J]. 西安外国语大学学报, 2018 (2): 18-23.
- 李佐文. 论元话语对语境的构建和体现[J]. 外国语, 2001 (3): 44-50.
- 柳淑芬. 元话语理论基础的多维视角探析[J]. 南京航空航天大学学报(社会科学版), 2013 (1): 75-78.
- 娄宝翠, 王莉. 学习者学术英语写作中自我指称语与作者身份构建[J]. 解放军外国语学院学报, 2020 (1): 93-99.

- 孙莉. 中国硕士学位论文英文摘要的语用身份建构研究[J]. 外语与外语教学, 2015 (5): 15-21.
- 孙莉. 中国硕士学术英语写作中元话语使用及其身份建构特征研究[J]. 西安外国语大学学报, 2020 (4): 28-33.
- 王晶晶, 吕中舌. 中国理工科博士生学术英语写作模糊限制语研究[J]. 外语教学, 2016 (5): 52-56.
- 徐昉. 二语学术语篇中的作者立场标记研究[J]. 外语与外语教学, 2015 (5): 1-7.
- 杨信彰. 元话语与语言功能[J]. 外语与外语教学, 2007 (12): 1-3.

通信地址: 453007 河南省新乡市 河南师范大学外国语学院

实证类汉语学术期刊论文中的自我指称与作者身份构建*

北京大学 王亚敏 宫雪 安卓玛

提要:本研究基于自建语料库,采用对应分析和多重对应分析方法,考察实证类汉语学术期刊论文中的自我指称形式与作者身份建构、论文语类、搭配动词之间的关系。研究发现:自我指称形式,多采用“我们”“本研究”和“本文”,其中抽象主体“本文”和“本研究”功能分工较为明确,前者主要与研究动词、意愿动词、认知动词和言语动词搭配,在摘要、方法语类中,承担“研究过程叙述者”“引导者”“语篇构建者”三种低风险身份;后者则倾向与结果动词共现,出现在讨论、结论语类中,构建“观点持有者”和“新知识创建者”两种高风险身份。相比之下,第一人称复数“我们”具有多功能性,在论文不同语类中搭配不同的动词,其所构建的作者身份涵盖以上五种及独有功能“群体代表”。本研究对汉语学术语篇中自我指称的使用特征进行了细致描写,研究结果可为学术汉语写作教学提供参考。

关键词:汉语学术语篇、自我指称、作者身份、语类、搭配动词

1 研究背景

学术论文是作者高度参与的社会性言语行为(Hyland 2000)。作者身份(authorial identity)是学术写作中不可或缺的修辞手段,在呈现作者自我的同时,体现作者与读者之间的互动关系,从而有效地实现学术语篇的交际功能。传统写作观认为,学术写作应该强调客观性,不应在学术语篇中显现作者身份,更不应使用表明作者身份的主观性语言,因此对学术语篇中作者身份的研究长期未受重视(Arnaudet & Barrett 1984; Lester 1993)。然而,近期研究表明,学术写作与作者身份构建之间存在交互关系,作者身份共存于同读者的关系之中,恰当地构建作者身份,不仅不会影响学术语篇的客观性,还有利于促进作者与读者的有效互动,推动作者在其所属学术领域展现学术成果,因此学术语篇中作者身份构建的研究显得尤为重要(Tang & John 1999; Hyland 2002; Çandarlı *et al.* 2015)。

自我指称(self-mention)是作者身份构建最常见的语言实现方式(Kuo 1999;

* 宫雪为本文通讯作者。

作者贡献:

王亚敏:选题构思、研究方法、数据收集、讨论结论、初稿撰写、字数占比(50%)、修改润色;

宫雪:选题构思、研究方法、数据收集、数据分析、初稿撰写、字数占比(50%)、修改润色;

安卓玛:选题构思、数据收集、修改润色。

Hyland 2001; Karahan 2013等),是作者在学术语篇中建立、维系与读者之间的互动关系并彰显自我身份的有效途径(Zareva 2013),也是其赢得话语权威的有力工具(Hyland & Jiang 2017)。Ivanič (1998)较早提出“作者自我”(authorial self)与文本中自我指称形式的使用密切相关,并根据作者参与程度由高到低,将自我指称的形式划分为第一人称代词(如I、We)、第三人称名词(如the author、the writer)和抽象主体(如the research、the paper)三类:Hyland (2002)也认为后两类通过名词短语的使用,降低了作者在学术语篇中的身份凸显度,第三类则将作者责任转接给非人称的、无生命的评价主体,从而增强研究本身的客观性和可靠性,是作者身份最为隐性的参与。不难看出,不同形式的自我指称形式彰显着作者在场(authorial presence)程度。

近年来,关于英语学术语篇中自我指称的研究成果相当丰富,主要涉及自我指称使用的跨学科差异(Hyland 2002; Harwood 2005; Gao 2017)、跨文化差异(Dueñas 2007; Çandarlı *et al.* 2015; Leedham & Fernández-Parra 2017)、学习者水平差异(Ivanič & Camps 2001; Luzón 2009)、语类差异(Martínez 2005; Dueñas 2007)等方面。不同文化背景的作者在学术语篇中的自我指称使用有着不同特点和功能。目前,围绕汉语学术语篇中自我指称的研究较少,如柳淑芬(2011)以中英文论文摘要中的自我指称为研究对象,研究发现中文摘要中作者参与度明显偏低,而英美作者则表现出较高的参与度,这可能是由作者与读者之间的社会距离、学术传统、文化传统等多种因素造成的。该项研究考察了学术语篇中自我指称形式的使用特点,但尚未涉及其背后构建的作者身份差异。吴格奇(2013)将研究对象扩大为期刊论文全文,考察了中英实证类期刊论文中自我指称的作者身份构建,研究表明汉语论文作者注重构建“研究者”身份,更加凸显自身的专业形象。以往研究表明,学术语篇中自我指称的使用受到多种因素的影响,而现有研究多为基于频次比较的单因素分析,吴文虽考察了实证类期刊论文的全文语料,但缺少基于论文不同语类的细致考察。此外,自我指称的使用及语篇功能的实现与其搭配动词密切相关,王月丽、徐宏亮(2019)就专门考察了学术英语写作中第一人称与搭配动词的使用情况,而汉语学术语篇的研究还未涉及对语类差异和搭配动词这两项因素的探究。因此,本研究基于自建语料库,集中探讨实证类汉语学术期刊论文中自我指称的使用特征,拟采用对应分析和多重对应分析方法,进一步考察自我指称形式与作者身份建构、论文语类、搭配动词之间的关系。

2 研究设计

2.1 语料来源

根据本文的研究目的及研究问题,我们在语料选择过程中主要考虑了以下因素。(1)学科领域。为保证所选期刊论文语料的学科领域一致性,本文所选论文

均来自语言学及应用语言学专业的汉语类核心期刊，主要包括以下5种期刊：《世界汉语教学》《语言教学与研究》《汉语学习》《语言科学》《语言文字应用》。（2）论文类型。本文所选的论文均为实证类论文，且同时具有“引言”“研究方法”“结果”“讨论”“结论”五个部分。（3）时间跨度。本文所选论文的发表时间为2016—2020年，发表时间较新且时间跨度不超过5年。（4）作者数量。本文所选论文均为独作文章，考虑到多位作者合作的论文各部分可能存在撰写风格不一致的情况，因此不在本文的选取范围之内。符合以上4个条件的期刊论文共计51篇，363,830词。

2.2 语料标注

期刊论文中，自我指称具有多种形式，如“我们”“作者”“笔者”“本文”“本研究”“本实验”等。对以上形式进行统计后发现，使用较多的是“我们”“本文”“本研究”，而“作者”“笔者”“本实验”频次较低，且集中分布于个别语篇中。因此，本研究主要对三种典型高频的自我指称形式进行考察，即“我们”“本文”“本研究”，并对其作者身份类型、所处语类、搭配动词进行了系统性标注。标注过程分为三轮：第一轮，由两位语言学及应用语言学专业博士研究生对前100条语料进行“背对背”预标注，结束后计算两者所作标注的内部一致性，若信度较低，则需要标注员对不一致的类型进行充分讨论，并重新标注，最终一致率达到87%；第二轮，两位标注员分别对895条语料进行正式标注，标注完成后对其结果进行信度计算，标注结果一致率达到82.91%；第三轮，由第三位标注员对第二轮中17.09%的不一致类型进行判断，最终再由三位标注员进行充分讨论，从而确定其标注类型。具体标注类型如下。

2.2.1 作者身份类型

本研究参考 Tang & John (1999) 的研究，根据作者身份凸显程度，将自我指称的身份类型由弱到强划分为：群体代表 (representative)、引导者 (guide)、语篇构建者 (architect)、研究过程叙述者 (recounter of the research process)、观点持有者 (opinion-holder)、新知识创建者 (originator)。其中群体代表、引导者、语篇构建者和研究过程叙述者属于低风险身份，而观点持有者和新知识创建者属于高风险身份，分类详情及示例如表1所示。

表1 自我指称所构建的作者身份类型

作者身份类型	功能释义	例句
群体代表	指一个群体，涵盖的范围可以是学术话语社团，也可以是更大的群体，作者把自己看作这个群体的代表。	在四维空间里，时间是非常重要的。无论在哲学讨论中还是物理学研究中，时间都是永恒的主题。在语言使用中，我们也常常需要对状态/事件作时间上的定位。

(待续)

(续表)

作者身份类型	功能释义	例句
引导者	引导读者完成语篇阅读。	在考察“了”和情状类型的关系以前，我们先看句法、韵律和语篇对“了”使用的影响。
语篇构建者	陈述研究目标，说明研究框架。	本文不仅统计归纳了泰国留学生产出普通话陈述句焦点重音音域和时长延长量两方面的特征，更抽丝剥茧地深入探究了其特征的形成机制，以期为相关的教学措施研究起到抛砖引玉的作用。
研究过程叙述者	叙述研究过程。	我们采用访谈法来进一步了解数据背后的原因。
观点持有者	对于既有的某事、某物、某种观点、做法，表达自己的观点、看法。	我们认为，在增加辨析频次的同时，应该注意第二次辨析的时间。
新知识创建者	基于所做的研究工作，提出新观点，创建新理论。	本研究提出了一个三维交互模型——即考虑发音器官、听觉因素和标记性效果这三个因素综合分析汉语鼻音感知过程。

2.2.2 论文语类

关于学术期刊论文，英语学界的研究成果较为丰硕。其中，Swales提出的引言部分CARS（Create a Research Space）模型（Swales 1990）和研究性论文IMRD（Introduction-Method-Results-Discussion）模式（Swales & Feak 1994），对后期学术期刊论文研究具有指导性意义。大部分应用语言学实证类学术期刊论文遵循IMRD模式，具有引言、研究方法、结果、讨论、结论五个子语类。这里所说的语类是指交际事件的类型，具有相同交际目的的交际事件为同一语类（Swales 1990）。学术期刊论文的五个子语类具有不同的交际目的，比如引言的交际目的是创建研究空间，方法部分的交际目的是论证研究方法的合理性等。因此，它们分属于五个不同的语类。本研究拟考察作者在不同语类中，即实现不同的交际目的时，所使用的自我指称形式以及所构建的身份类型有何差异。主要包括以下六种语类：摘要、引言、研究方法、结果、讨论、结论。

2.2.3 搭配动词类型

为细致描述实证类学术期刊论文中自我指称形式的使用情况，我们也对与自我指称形式搭配的动词作了统计和分类。动词分类参考Thomas & Hawes（1994）对报道动词的分类以及王月丽、徐宏亮（2019）对与We/I搭配动词的分类，详情及示例如表2所示。

表2 搭配动词分类

动词类型	定义	主要动词
意愿动词	阐述研究目的。	试图、尝试、旨在
研究动词	阐述研究过程，如实验设计、研究方法、研究内容等。	设计、提取、采用
结果动词	报告研究结果，可分为两类：一是客观重述研究结果；二是在报告结果的同时加入作者的主观态度，如作者明确认同这一结果。	发现、表明、显示、证实
认知动词	报告作者的认知过程、心理过程。	认为、假设、推测
言语动词	报告作者的言语行为。	总结、简称、称为、回答

2.3 分析方法及工具

本文所采用的统计分析方法为对应分析（Correspondence Analysis）和多重对应分析（Multiple Correspondence Analysis，简称MCA）。两者均为探索性的统计方法，用于发现变量间的关联。对应分析可用于展示样本与变量间的对应关系。对应分析的本质是对多维数据的降维处理，与因子分析类似，但因子分析存在一定的不足。因子分析分为R型因子分析和Q型因子分析，R型因子分析用于考察变量（指标）间的相关关系，Q型因子分析用于考察样本间的相关关系。如果我们既关心变量（指标）间、样本间的相关关系，也关心两者之间的对应关系，因子分析便无法满足这样的需求，此时就需要用到对应分析。当变量涉及多个多分类变量时，还需要用到多重对应分析。

本文中的样本指的是语篇中自我指称所构建的身份类型，包括“群体代表”“引导者”“语篇构建者”“研究过程叙述者”“观点持有者”“新知识创建者”6种，变量指的是自我指称形式及其所处的论文语类、与之共现的动词类型，以上3个变量均属于多分类变量。对应分析的工具为R语言平台上{ca}工具包。多重对应分析使用的工具包为{ca}和{FactoMineR}。

3 结果和分析

3.1 自我指称形式与作者身份类型的对应关系

为考察自我指称形式与作者身份类型的关系，我们首先对不同自我指称形式所构建身份类型的频次进行了统计。总的来说，实证类汉语学术期刊论文中的自

我指称在形式上更倾向于使用第一人称复数“我们”(55.86%),其次是抽象主体“本研究”(24.92%)和“本文”(19.22%),所构建的身份类型涵盖六大类,低风险身份(73.52%)远多于高风险身份(26.48%),最为典型的身份为“研究过程叙述者”(45.14%),其次是“新知识创建者”(15.86%)和“语篇建构者”(14.75%),再次是“引导者”(10.73%)和“观点持有者”(10.61%),“群体代表”最少(2.91%),结果如表3所示。

表3 不同自我指称形式所构建身份类型的频次

作者身份类型	我们	本文	本研究	总计
群体代表	26	0	0	26
引导者	50	29	17	96
语篇建构者	71	37	24	132
研究过程叙述者	214	86	104	404
观点持有者	65	12	18	95
新知识创建者	74	8	60	142
总计	500	172	223	895

我们进一步采用对应分析方法,考察自我指称形式与身份类型的关系。结果表明,维度1和维度2的累积解释比例达到100%,因此对应分析的结果可以用二维坐标图来表示。如图1所示,横坐标为维度1,右侧为正向、左侧为负向;纵坐标为维度2,上方为正向,下方为负向。

根据对应分析图,“新知识创建者”和“本研究”位于维度一的负向,说明该身份主要由“本研究”构建。“研究过程叙述者”“语篇建构者”“引导者”身份和“本文”位于维度1的正向,说明以上三种身份主要由“本文”构建。“群体代表”“观点持有者”和“我们”同样位于维度2的正向,说明“我们”主要构建了以上两种身份。值得注意的是,“我们”位于维度1的中点位置,说明“我们”所构建的身份类型比较分散。仅根据此图无法观察出“我们”所构建身份的倾向性,因此,下文将在3.3中对“我们”进行详细分析。

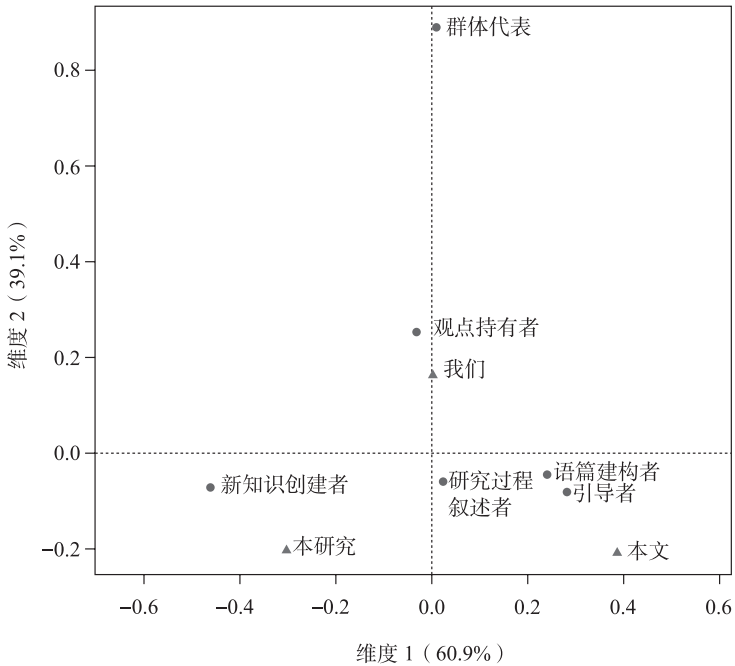


图1 自我指称形式与作者身份类型的对应分析

由此可见，“本研究”“本文”“我们”虽同为自我指称，但各有分工。根据上文的统计结果，主要得出以下两点结论。

一是抽象主体“本文”和“本研究”在使用时分工明确。“本研究”更常用于构建高风险身份，倾向于用来创建新知识，如例（1）：

（1）**本研究**从认知理解的角度提出了汉语动名搭配的新分类，并在此基础上对不同水平二语学习者理解这些搭配的情况进行了考察。

与“本研究”相比，“本文”则更倾向于构建低风险身份，用于叙述研究过程，如例（2）：

（2）**本文**利用国家语委现代汉语语料库（以下简称为“语委语料库”）的语料来计算准入词与“有N”结构、准入词与形容词的搭配强度。

“本文”还可用于描述论文语篇结构和引导读者阅读，如例（3）和例（4）：

(3) 本文以课堂教学及口语考试的实录语料为数据来源,运用微变化分析的方法,对比课堂教学环境下教师输入与韩国学生输出的情况,详细考察韩国学生对不同类型“比”字句的习得认知过程,揭示他们在习得过程中遇到的主要难点。

(4) 最终,本文将语素义和词义的关系分为三类,如表1所示……

二是相较于“本文”和“本研究”,“我们”更倾向于构建群体代表和观点持有者身份,如例(5)和例(6):

(5) 因此,我们有必要从认知习得的角度对汉语动名搭配重新进行分类,围绕这一分类对学习者的相关认知加工过程进行考察。

(6) 我们认为这就是为什么英文中的简单过去时能和所有类型的动词共现的原因。

3.2 抽象主体“本研究”和“本文”的功能分化

本文在进行语料选取时,考虑到实证类学术期刊论文包含摘要、引言、方法等子语类,不同语类具有不同的交际目的,因此作者在论文的写作过程中需要不断变换身份,以实现每个语类的交际目的。身份的变换带来自我指称形式的变化,也预示着将选择不同类型的动词进行搭配。基于以上分析,本小节通过多重对应分析,考察作者身份类型与自我指称形式、论文语类、搭配动词之间的多重对应关系。分析结果如图2所示。

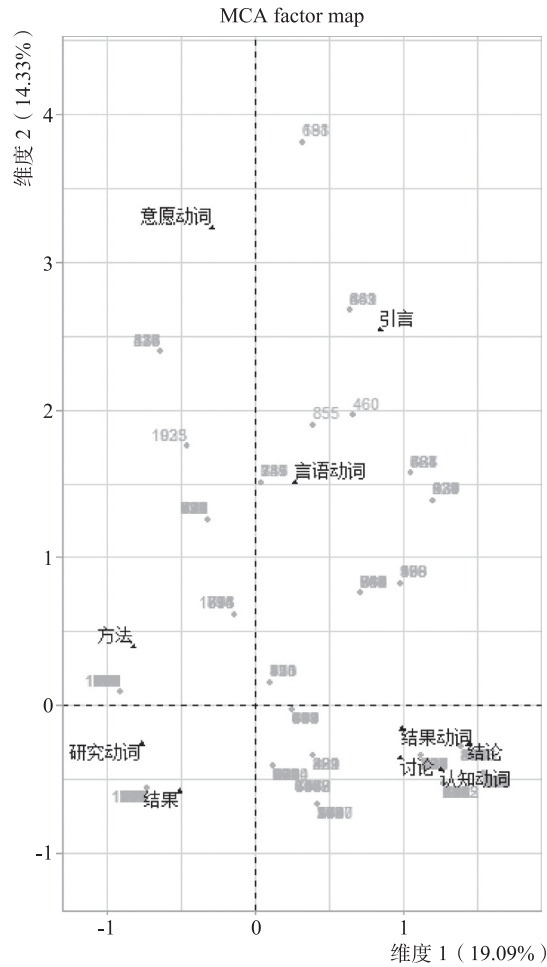


图2 自我指称形式与论文语类、搭配动词的MCA

图中灰色菱形点为各种身份类型的语言使用实例（即2.3中所说的样本），黑色三角形点为变量（包括自我指称形式、论文语类、搭配动词）。维度1、2上具有高贡献值的参项（Estimate值>0.5）。为观察作者身份类型与其他三个变量的对应关系，我们将“作者身份类型”作为补充变量加入图中，如图3所示。

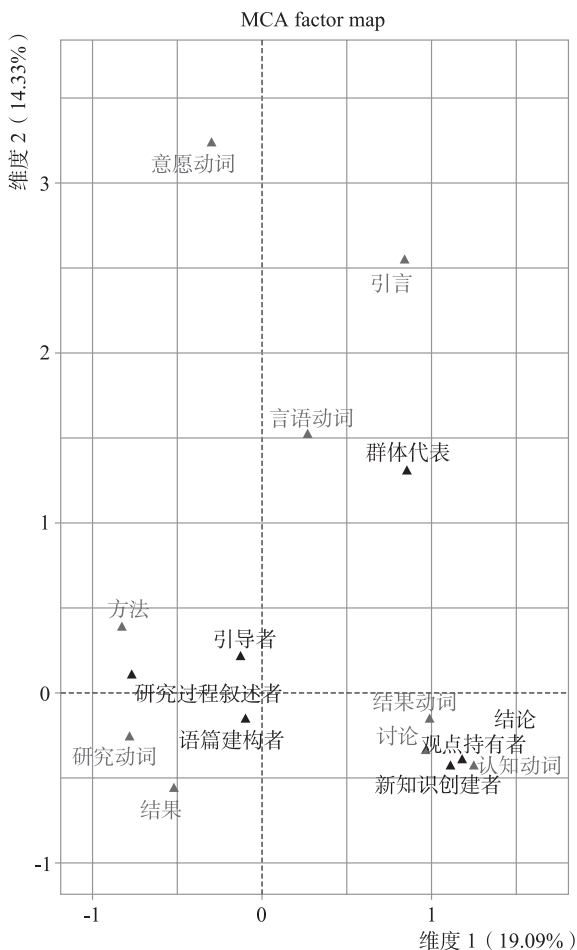


图3 加入补充变量“作者身份类型”的MCA

据图3可知,我们可将6种身份类型归为两类。第一类包括“研究过程叙述者”“引导者”“语篇建构者”三种低风险身份类型,此类身份类型主要由“本文”构建,更倾向于出现在摘要、方法语类中,与之相搭配的动词包括研究动词、意愿动词、认知动词和言语动词,如例(7)和例(8)所示。

(7) 本文通过实证研究考察课堂显性辨析对汉语二语学习者习得近义词差异的影响。

(8) 综合考虑上述因素,本文选择单句中的光杆普通NP作为研究对象,只探讨“们”在单句层面的标记规则及其习得。

第二类包括“观点持有者”“新知识创建者”两种高风险身份类型,主要由“本研究”构建,出现在讨论、结论语类中,与之相搭配的动词主要是结果动词,如例(9)和例(10)所示。

(9) 本研究发现, 语素意识通过两条通路间接作用于阅读理解: 一条通过词汇知识, 另一条通过词汇推理。

(10) 本研究从认知理解的角度提出了汉语动名搭配的新分类, 并在此基础上对不同水平二语学习者理解这些搭配的情况进行了考察。

3.3 第一人称复数“我们”身份构建的多功能性

从上文的统计结果看, “我们”的用例较多, 所构建的身份类型多样, 情况较为复杂。因此, 本小节专门对“我们”所构建的作者身份类型与论文语类、搭配动词的对应关系进行分析。需要指出的是, “我们”在摘要中仅出现了2次, 数据稀疏, 影响MCA图的可视化效果, 故本文在进行多重对应分析时, 去掉了“我们”在摘要中出现的数据。同上文, 我们将作者身份类型作为补充变量加入图中, 结果如图4所示。

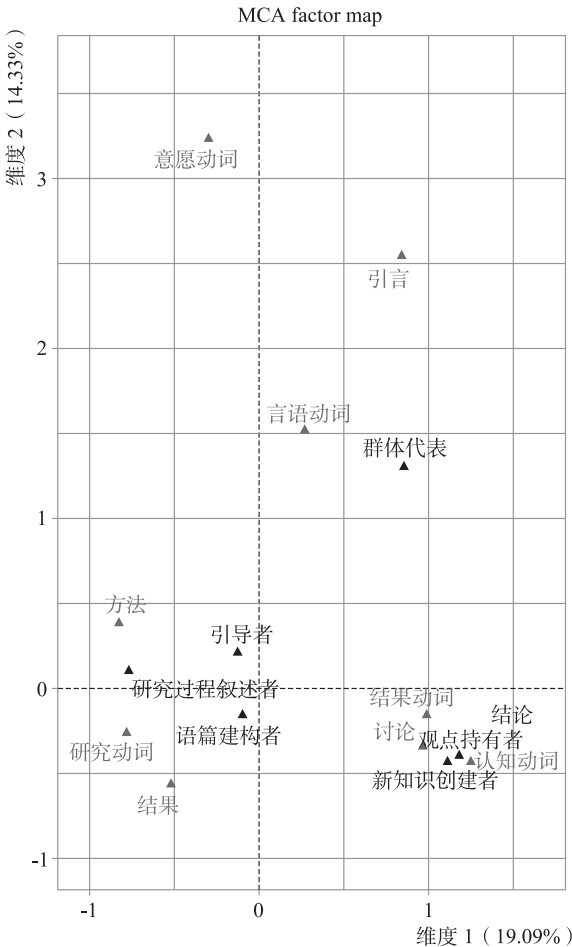


图4 加入补充变量“我们”的MCA

据图4可知,“我们”所构建的身份类型为“群体代表”时,其主要出现在引言语类中,与之搭配的动词是言语动词,如例(11)。

(11) 一般**我们**所说的语块包括习语、高频搭配等。

“我们”所构建的身份为“研究过程叙述者”和“引导者”时,主要出现在方法语类中,与之搭配的动词是意愿动词,如例(12)。

(12) **我们**尝试通过实验研究来做出回答。

其所构建的身份为“语篇构建者”时,主要出现在结果语类中,与之搭配的动词为研究动词,如例(13)。

(13) **我们**首先计算五类差异在前测、后测、延后测的整体平均得分(即将五类差异得分相加除以五),然后再分别考察课堂辨析对每类差异习得的影响。

“我们”所构建的身份为“观点持有者”和“新知识创建者”时,主要出现在讨论、结论语类中,与之搭配的动词是结果动词和认知动词,如例(14)和例(15)。

(14) 我们也证明,不同语言具有不同的非完成体,表达不同的非完成意义,如简单的部分视角,强调进行的过程,提供静态视角等等。

(15) **我们**认为应该是第一学期开设的所有课程都对职前教师的语法信念变化产生了影响。

4 结语

本研究基于自建语料库,采用对应分析和多重对应分析方法,考察了实证类汉语学术期刊论文中自我指称形式与作者身份建构、论文语类、搭配动词之间的多重对应关系。研究发现:(1)在自我指称形式上,汉语学术语篇更倾向于使用第一人称复数“我们”,其次是抽象主体“本研究”和“本文”,较少用第一人称单数(如“我”)或第三人称(如“作者”和“笔者”);(2)在作者身份构建上,

低风险身份远大于高风险身份，按照出现频次高低依次为“研究过程叙述者”“新知识创建者”“语篇构建者”“引导者”“观点持有者”和“群体代表”；（3）不同自我指称形式在作者身份建构上存在差异，与所处论文语类、搭配动词之间存在多重对应关系，总结如表4所示。

表4 汉语学术语篇中自我指称形式与作者身份构建的多重对应

自我指称形式	论文语类	搭配动词	作者身份构建	
抽象主体	本文	研究动词	研究过程叙述者 引导者 语篇构建者	低风险身份
		意愿动词		
		认知动词		
		言语动词		
第一人称 复数	本研究	讨论	观点持有者 新知识创建者	高风险身份
		结论		
	我们	引言	群体代表	低风险身份
		方法	研究过程叙述者 引导者	低风险身份
		结果	语篇构建者	低风险身份
		讨论 结论	观点持有者 新知识创建者	高风险身份

具体来说，抽象主体“本文”和“本研究”功能分工较为明确，前者主要与研究动词、意愿动词、认知动词和言语动词搭配，在摘要、方法语类中，承担“研究过程叙述者”“引导者”“语篇构建者”三种低风险身份；后者则倾向于与结果动词共现，出现在讨论、结论语类中，构建“观点持有者”和“新知识创建者”两种高风险身份。相比之下，第一人称复数“我们”具有多功能性：在引言语类中搭配言语动词构建“群体代表”低风险身份，在方法语类中搭配意愿动词实现“研究过程叙述者”和“引导者”低风险身份，在结果语类中搭配研究动词表达“语篇构建者”低风险身份，在讨论和结论语类中与结果动词和认知动词共现实现“观点持有者”和“新知识创建者”高风险身份。

本研究聚焦于实证类汉语学术期刊论文中自我指称的使用特征，综合考虑了自我指称形式、作者身份类型、论文语类及搭配动词等多种因素，通过细致的描写进一步理清了“本研究”“本文”“我们”三种典型高频自我指称形式的使用差异，研究结果有助于提升二语学习者和母语者的学术论文写作能力，尤其是指导

学生如何在汉语学术语篇中恰当地表达作者身份，并在一定程度上为学术汉语写作教学提供参考。

参考文献

- ARNAUDET M, BARRETT M. Approaches to academic reading and writing [M]. Upper Saddle River: Prentice-Hall, 1984.
- ÇANDARLI D, BAYYURT Y, MARTI L. Authorial presence in L1 and L2 novice academic writing: cross-linguistic and cross-cultural perspectives [J]. *Journal of English for Academic Purposes*, 2015, 20(2): 192-202.
- DUEÑAS P. “I/we focus on...”: a cross-cultural analysis of self-mentions in business management research articles [J]. *Journal of English for Academic Purposes*, 2007, 6(2): 143-162.
- GAO X. A cross-disciplinary corpus-based study on English and Chinese native speakers’ use of first-person pronouns in academic English writing [J]. *Text & Talk*, 2017, 38(1): 93-113.
- HARWOOD N. “Nowhere has anyone attempted ... In this article I aim to do just that”: a corpus-based study of self-promotional I and we in academic writing across four disciplines [J]. *Journal of Pragmatics*, 2005, 37(8): 1207-1231.
- HYLAND K. *Discipline discourse: social interactions in academic writing* [M]. London: Longman, 2000.
- HYLAND K. Humble servants of the discipline? Self-mention in research articles [J]. *English for Specific Purposes*, 2001, 20(3): 207-226.
- HYLAND K. Authority and invisibility: authorial identity in academic writing [J]. *Journal of Pragmatics*, 2002, 34(8): 1091-1112.
- HYLAND K, JIANG F K. Is academic writing becoming more informal? [J]. *English for Specific Purposes*, 2017, 45(1): 40-51.
- IVANIČ R. *Writing and identity: the discoursal construction of identity in academic writing* [M]. Amsterdam: John Benjamins, 1998.
- IVANIČ R, CAMPS D. I am how I sound: voice as self-representation in L2 writing [J]. *Journal of Second Language Writing*, 2001, 10(1-2): 3-33.
- KARAHAN P. Self-mention in scientific articles written by Turkish and non-Turkish authors [J]. *Procedia-Social and Behavioral Sciences*, 2013, 70(1): 305-322.
- KUO C. The use of personal pronouns: role relationships in scientific journal articles [J]. *English for Specific Purposes*, 1999, 18(2): 121-138.
- LEEDHAM M, FERNÁNDEZ-PARRA M. Recounting and reflecting: the use of first person pronouns in Chinese, Greek and British students’ assignments in engineering

- [J]. *Journal of English for Academic Purposes*, 2017, 26(3): 66-77.
- LESTER J. *Writing research papers: a complete guide* [M]. New York: Harper Collins Publishers, 1993.
- LUZÓN M. The use of “we” in a learner corpus of reports written by EFL engineering students [J]. *Journal of English for Academic Purposes*, 2009, 8(3): 192-206.
- MARTÍNEZ I. Native and non-native writers’ use of first person pronouns in the different sections of biology research articles in English [J]. *Journal of Second Language Writing*, 2005, 14(3): 174-190.
- SWALES J. *Genre analysis: English in academic and research settings* [M]. Cambridge: Cambridge University Press, 1990.
- SWALES J, FEAK C. *Academic writing for graduate students (second edition)* [M]. Ann Arbor: University of Michigan Press, 1994.
- TANG R, JOHN S. The “I” in identity: exploring writer identity in student academic writing through the first person pronoun [J]. *English for Specific Purposes*, 1999, 18(1): 23-39.
- THOMAS S, HAWES T. Reporting verbs in medical journal articles [J]. *English for Specific Purposes*, 1994, 13(2): 129-148.
- ZAREVA A. Self-mention and the projection of multiple identity roles in TESOL graduate student presentations: the influence of the written academic genres [J]. *English for Specific Purposes*, 2013, 32(2): 72-83.
- 柳淑芬. 中英文论文摘要中作者的自称语与身份构建[J]. *当代修辞学*, 2011 (4): 85-88.
- 王月丽, 徐宏亮. 中国英语学习者学术写作中第一人称使用发展特征与身份构建研究[J]. *外语教育研究前沿*, 2019 (3): 58-64.
- 吴格奇. 学术论文作者自称与身份构建——一项基于语料库的英汉对比研究[J]. *解放军外国语学院学报*, 2013 (3): 6-11.

通信地址: 100871 北京市 北京大学对外汉语教育学院

语料库方法在汉语第二语言习得研究中的应用研究*

安徽大学 彭家法 孙梦馨

提要：本文对国内中文核心期刊论文中汉语二语习得语料库方法的使用情况进行了梳理，就汉语二语习得语料库方法的发展历程、研究现状、作用和局限等问题作了探讨，评述了当前汉语二语习得研究中语料库方法的发展和不足。分析结果显示：（1）语料库方法的发展历程主要分为起步期、发展期和飞跃期三个阶段，目前是该方法使用的飞跃期；（2）该方法近些年在语料收集、偏误分析研究方面有理论和应用的发展，但在用例分析、习得顺序研究等方面仍没有发挥出应用价值，仍存在着用例分析不全面、习得顺序研究数量不足、方法单一、过程不完整等问题；（3）该方法具有反映中介语全貌等两点优势以及在实际应用中存在着研究内容不平衡等两点不足。针对语料库方法应用于汉语二语习得研究所存在的问题，本文进行了相关的解释分析，提供了解决问题的途径，并展望该研究方法的发展前景。

关键词：语料库方法、汉语二语习得、偏误分析、正确用例、习得顺序

1 引言

语料库方法是二语习得研究中一种常用的定量研究方法，它依托语料库中大量真实的母语或学习者语料，通过对语料样本的统计分析，总结母语或学习者语言的使用特征或情况，归纳语言运用的规律。

近些年，有不少学者对语料库方法作了总结。冯丽萍、孙红娟（2010）指出语料库分析是习得顺序研究常用的一种方法。张宝林（2011）详细说明了语料库对汉语二语习得研究的重要性，提出大规模语料能保证研究结论具有客观性和普遍性。毕晋等（2017）基于4种国内核心期刊，统计概述了汉语二语习得研究中语料库方法的应用领域和使用频率，并对中介语研究语料来源作了分类。曹贤文（2020）总结了语料库方法应用于汉语二语习得研究时主要采用“偏误分析”“频率分析”和“中介语对比分析”3种分析方法。虽然前人对语料库方法的应用进行了很好的总结，但是仍存在着一些不足。他们缺少对该方法在应用中取得的发展和存在的问题的总结。语料库方法是当前汉语二语习得研究中常用的量化研究方

* 孙梦馨为本文通讯作者。

作者贡献：

彭家法：选题构思、研究方法、讨论结论、字数占比（20%）、修改润色；

孙梦馨：数据收集、数据分析、讨论结论、初稿撰写、字数占比（80%）。

法,它能客观反映母语和二语习得者语言使用情况,并且其提供的庞大样本使得研究结论具有较高的可信度和科学性,极大促进了语言特征方面的研究。梳理总结语料库方法近些年取得的发展和现存的不足,有助于研究人员更深入地了解该研究方法的使用条件、应用领域等,从而能在习得研究中选择更合适的研究方法,得出更准确的结论。本文将以张宝林(2011, 2021)等的理论说明为依据,结合相关论文,梳理语料库方法在汉语二语习得研究中的使用情况。第一节为引言部分,主要介绍什么是语料库方法以及本文的研究意义。第二节对该方法的发展历程作了阶段划分,总结不同阶段该研究方法在各研究领域的应用情况。第三节总结其在语料收集方面取得的进展。第四节分析总结该方法在偏误分析领域的应用成果。第五节考察该方法在用例分析和习得顺序研究中的使用情况,发现其在这两个研究方面目前存在用例分析不全面、习得顺序研究数量不足、方法单一、过程不完整等问题,并针对上述问题进行了解释。最后的结语部分总结语料库方法的优势和局限,针对其局限性提供相应建议以供研究人员参考,并对语料库方法的发展前景作一展望。

2 发展历程与现状分析

我们筛选出了1984—2021年《世界汉语教学》《语言文字应用》《语言教学与研究》《汉语学习》中与汉语二语习得研究相关的论文128篇¹,其中运用语料库方法的共计62篇。施家炜(2006)对二语习得研究领域作了细致分类,共分为“学习者语言特征研究”“学习者外部因素研究”“学习者内部习得机制研究”和“学习者研究”四大类,其中每大类下还包含如“偏误分析”“习得顺序”等12小类。他指出,国内汉语二语习得研究大部分只集中于学习者语言特征研究领域,其他领域鲜少涉及。因此,我们以语料库方法应用的研究领域为主要分期依据,结合语料库建设情况,从历时和共时层面对语料库方法的发展历程作一划分。

语料库方法主要可以分为起步期、发展期和飞跃期三个阶段。20世纪90年代中期至2004年是该方法的起步期,这一阶段,国内中介语语料库陆续建成,为语料库方法的使用提供了条件。该时期发文数量并不多,仅有6篇,研究的主题都是学习者语言特征领域。偏误分析相关文献共有3篇,比如陈小荷(1996)根据中介语语料库对留学生使用副词“也”的情况进行了偏误分析;习得顺序研究如王建勤(1997)借助中介语语料库对留学生习得“不”和“没”否定结构的顺序进行了考察。

1 我们选择这四种期刊有两方面原因:其一是这四种期刊一直是CSSCI来源期刊;其二是这些期刊都刊发了较多与汉语二语习得研究、教学相关的论文。我们从1984年开始统计是因为汉语二语习得研究以鲁健骥(1994)引入“偏误”和“中介语”概念为开端;以2021年结尾是因为本文撰稿时期为2022年初。

2005—2012年是该方法的发展期,语料库建设加速发展,CCL语料库、HSK动态作文语料库等相继投入使用。该阶段发文数量有20篇,研究主题开始涉及内部习得机制,比如黄月圆等(2005)依据中介语语料库,通过考察留学生汉语反身代词“自己”的习得情况,证实了普遍语法和母语会影响成人的二语习得。除了偏误分析外,习得顺序也逐渐成为研究热点,共有7篇文献进行了探究,如周小兵、邓小宁(2009)和谢福(2010)等。

2013—2021年是该方法使用的飞跃期,语料库建设取得了长足进步,HSK动态作文语料库不断更新完善,BCC语料库、全球汉语中介语语料库也相继建成并得到充分使用。这一阶段的发文数量达到了36篇,除了学习者语言特征领域外,内部习得机制领域也成了研究热点,共有6篇相关文献,比如蔡淑美、施春宏(2014)和刘旭(2018)等。学习者外部因素领域也出现了相关研究论文,比如程燕、肖奚强(2020)借助语料库考察留学生四字成语的输出和输入情况,证实了课堂教学、目的语环境和输入频率等会影响留学生的习得情况。

根据张宝林(2011, 2021)对应用语料库方法的理论说明以及相关论文的分析,我们归纳出该研究方法在语料收集、偏误分析、用例分析、习得顺序研究等方面应用时应注意的事项。语料收集包括语料收集方法和语料筛选;偏误分析包括偏误类型的归纳和偏误原因分析;用例分析涉及正确用例的分析和用例类型的归纳;习得顺序研究涉及方法选择、原因解释的问题。从62篇论文中语料库方法的使用情况来看,现有文献在语料收集、偏误分析领域做得相对较好;而在用例分析、习得顺序研究方面仍存在不足,比如存在缺少正确用例分析和用例类型归纳、习得顺序研究数量不足、研究方法单一、缺少原因解释等问题。下面我们将具体介绍语料库方法在使用上的发展和存在的问题,并针对不足之处给出相应的原因解释及建议。

3 汉语第二语言习得语料收集

现有文献在语料收集和语料筛选方面做得较好。语料收集方法科学、多元,语料筛选也有相应的标准。

毕晋等(2017)将中介语研究语料来源分成“经验式语料”“跟踪式语料”和“语料库语料”。我们依据前人的理论,结合筛选出的与汉语二语习得研究相关的128篇论文,将汉语二语习得研究语料的收集方法分为教学实践法、内省法和语料库法。从20世纪80年代开始,教学实践法就已经运用于汉语二语习得研究了,比如汪宗虎(1985)收集了教学中留学生词语运用出现的偏误,并展开分析。而随着汉语中介语语料库、现代汉语语料库的建成和普及,汉语二语习得研究在20世纪90年代末陆续出现使用语料库进行语料收集的方法,主要有两种方式,一种是

通过字符串检索的方式直接提取语料，即“语句检索提取”方式；另一种是通过全篇阅读、手动查找的方式提取语料，即“篇章手动提取”方式。篇章手动提取指的是当所需要的语料无法用字符串的形式进行检索时，研究人员会阅读全篇语料，从中手动提取相关的语句。比如李榕、王元鑫（2021）在考察留学生第三人称回指习得情况时，从中介语语料库中筛选出128篇留学生作文，然后通过人工阅读全篇的方式挑选出含有第三人称回指的句子。进入21世纪以来，有些学者从生成语法视角出发研究汉语二语习得，他们多通过内省法获得汉语本体研究的语料，比如袁博平（2017）等，虽然是依靠个人语感造出的句子，但也保证了语料的自然流畅。

检索出结果后，需要对语料进行筛选。在语料的筛选方面现有研究都做得较好。我们根据张宝林（2011，2021）的有关理论和上文选取的期刊论文，归纳了几个筛选语料的方法。首先，要注意检索出的内容是否符合要求。张宝林（2021）指出，由于目前的语料库检索都是形式检索，因此会出现许多“与信息内容相关，但并非是需要查找的内容”，这时候就需要我们仔细甄别。其次，在偏误研究时，还要注意将句子和语段结合，比如张宝林（2010）就曾举“把”字句泛化的例子来说明。最后，我们选取语料作为例句时，一般先从中国文学作品中选取，因为其表达更符合汉语母语者语感。但是并非所有中国文学作品中的语言表达都是自然、通顺的，这时就需要我们注意甄别语句的通顺与否。如果实在找不到符合要求的中国文学作品语料，可以从外国文学作品中寻找，但需要确保其译文水平尽可能接近自然汉语的水平。

4 汉语第二语言习得偏误分析

目前语料库方法在偏误分析研究方面做得较好。偏误类型方面，现阶段大多数文献都能采用定量分析的方法，依据学习者语料对偏误类型进行划分统计；偏误原因分析普遍较为详细，且原因分类更加多样。但是依然存在着一些不足：偏误类型归纳方面仍有少数文献直接套用前人的类型归纳结果；偏误原因分析方面仍有少数文献分析太过笼统。

在对外汉语教学研究中，鲁健骥（1994）归纳出了遗漏（回避）、误加、误代、错序四种偏误类型。但在实际研究中，我们应该根据学习者偏误语料归纳偏误类型，而不是直接沿用上述结论。张宝林（2011）曾指出，既往偏误分析相关研究几乎都沿用了这四种分类，这就使得研究缺乏针对性和个性化，失去了意义。根据对论文的梳理，我们共筛选出了33篇涉及语料库方法和偏误分析的文献，其中有23篇文献都比较符合我们对偏误类型归纳的要求，前两个阶段有10篇，现阶段有13篇。这些文献都采用不同方法或者从不同角度，归纳出新的偏误类型或给出更细致的下位分类，比如陈小荷（1996）按照偏误发生的语境条件把副词“也”

出现的语序偏误分成“‘也’在主语前”等三种详细的下位类别；周小兵、邓小宁（2009）采用“显性偏误”和“隐性偏误”的分类方法对含“得”字的两种补语句偏误语料进行归纳分类；蔡淑美、施春宏（2014）从“语块配位方式”角度将二价名词习得的偏误分为“隔开式”“框架式”“话题式”等类型。不过目前偏误类型归纳研究方面也仍然存在着尚未完全摆脱套用结论的问题。现阶段与偏误类型归纳相关的论文研究共有17篇，我们发现仍然有一小部分论文（4篇）间接或直接沿用鲁健骥（1994）的四种偏误类型。比如牛长伟、李君（2019）在讨论“什么”类代词的偏误类型时，直接依据鲁健骥的分类将“什么”类偏误语料归纳为“副词遗漏”“代词误代”“副词误代”“标点误代”四种类型，且每一类的偏误占比也并没有都统计展现出来。

从偏误类型归纳的方法上来看，现阶段研究较前两个阶段有所发展，普遍更重视运用定量分析。前两个阶段的偏误类型归纳总体上缺少量化统计，大部分文献都只是举例说明某类偏误类型的表现，并未统计该偏误类型在留学生总偏误语料中的占比，无法推知这样的类型是否具有普遍性，比如崔希亮（2005）。张宝林（2011）指出，二语习得相关研究应采用定量和定性相结合的方法，尤其注重定量分析，这样可以保证研究具有最大限度的客观性和较强的说服力。如果只是通过定性的方法主观判断偏误类型，或者依据极少甚至是单个样本划分出偏误类型，我们很难确定该结论是否具有普遍性，也无法从各偏误类型的占比中知道留学生的习得难点是什么。而在现阶段的研究中，研究人员都采用定量分析的方法研究偏误类型。比如蔡淑美、施春宏（2014）用定量分析的方法对留学生习得二价名词出现的偏误类型进行划分，并统计了每一类偏误在总偏误数中的占比，发现“框架式偏误”是留学生习得的难点。

在偏误原因分析方面，应避免过于笼统的分析。张宝林（2011）指出，在分析偏误原因时，我们应该具体深入。我们筛选统计了涉及偏误原因分析的28篇文献，发现现阶段的偏误原因分析较前两个阶段更加详细，且原因分类更加多样。前两个阶段共有14篇文献进行了偏误原因分析，其中有6篇文献对原因的解释相对充分，比如高立群（2001）从认知策略和教学过程两方面解释留学生在不成字部件构成的汉字上错误率高的原因。剩下8篇的偏误原因分析则较为笼统，比如崔希亮（2005）只是简单用“母语影响”一句话解释留学生介词结构出现的位置不当偏误，并未作详细分析。现阶段涉及偏误原因分析的文献也有14篇，其中对原因有详细解释的有11篇，并且多数文献尝试从新的角度进行偏误原因分析。比如李榕、王元鑫（2021）从“语言共性知识干扰”和句子的“完形”心理认知角度对韩语为母语的汉语二语学习者出现的第三人称代词回指“过度使用”偏误进行解释。但目前仍有少数文献存在着原因解释较为笼统的现象。吴继峰（2013）在分析留学生中出现的在重叠形容词前加“很”的冗余偏误时，只是给出了不能

加“很”的原因，却没有对留学生出现这一偏误的原因进行分析；形容词间的误用也只是用留学生没有正确区分近义词词义简单解释，没有作更深一步的探究。

5 汉语第二语言习得用例分析和习得顺序研究

目前的汉语二语习得领域，不少研究人员在使用语料库方法时都忽视了对正确用例进行分析、对用例类型进行归纳。此外，利用语料库进行习得顺序研究的论文数量远少于偏误分析的数量，并且习得顺序研究方法单一，部分研究人员忽视了对习得顺序原因的解释。

5.1 正确用例分析

刘珣（2000）、张宝林（2011）都曾指出，中介语研究不只要进行偏误分析，更要对正确用例进行分析，否则无法看到中介语的全貌及其动态的发展轨迹。从我们收集到的论文来看，虽然近些年有一些研究人员注意到了对正确用例分析的重要性，但总体来说，在运用语料库方法考察留学生偏误语料的同时还对正确的中介语表达做分析的研究依然是少数，在62篇相关文献中只有11篇涉及了正确用例分析，比如杨圳、施春宏（2013）运用语料库分别考察了留学生准价动词的正确和偏误输出用例，通过对正确用例的分析发现留学生可以灵活准确地使用“结婚”等这类协同、离合动词。运用语料库进行正确用例的分析，有助于我们更全面地认识留学生习得汉语某一语法项目的情况。

5.2 用例类型归纳

在进行汉语二语习得研究时，我们也应对留学生学习某一语法项目时产生的正确语例进行归纳分类。对用例进行分类不仅有助于汉语的本体研究，也有利于发现留学生学习该语言点的习得顺序，但在62篇相关文献中仅有2篇涉及此类研究。王建勤（1997）通过汉语中介语语料库收集整理了914条与“不”和“没”有关的正确语料，并根据这些语料将“不”归纳成10类结构，比如“不（太）+V”“不（难；好）+V”等；将“没”归纳成4类结构，比如“没+V”“没N/Adv+Adj”等。蔡淑美、施春宏（2014）将留学生二价名词的正确用例归纳成“(……NP……)+V+N”等10种用例类型，并与袁毓林相关文章中所列举的格式进行对比，发现留学生在使用中还出现了袁文没有提到的格式，这一发现无疑会促进汉语本体研究和对外汉语教学研究。

5.3 习得顺序研究

虽然从我们收集到的论文来看，语料库方法主要应用于偏误分析和习得顺序研究，但两个主题间的发文数量还是有明显差异的，运用语料库方法进行习得顺

序研究的论文数量远少于偏误分析。涉及偏误分析的论文有36篇,而涉及习得顺序的只有11篇。

5.3.1 习得顺序研究方法

根据现有的习得顺序研究文献,我们发现在利用语料库方法进行习得顺序研究时,研究人员采取的习得顺序研究方法较为单一。现有习得顺序研究中有7篇都采用了使用频率和(或)正确率排序法,比如彭淑莉(2008)通过计算平均使用频率和平均正确率排名之和的方法研究汉语带宾语的“被”字句的习得顺序;肖奚强、周文华(2009)根据使用频率和正确率对留学生的汉语趋向补语习得顺序进行排序。张利蕊(2018)曾指出,习得顺序常用的研究方法除了有正确率排序法、使用频率排序法,还有正确使用相对频率法、阶段计分法、习得区间法、蕴含量表法等,但从现有文献的使用情况来看,后面这些研究方法相对来说用得较少,甚至是极少。施家炜(1998)、谢福(2010)等采用了正确相对使用频率的方法计算习得顺序;王建勤(1997)、施家炜(1998)采用习得区间法对语法项目的习得等级进行划分;而蕴含量表法、阶段计分法只出现在施家炜(1998)的研究中。因此,在未来的习得顺序研究中,研究人员可以尝试采取其他研究方法进行习得顺序的排序。

5.3.2 习得顺序原因解释

从现有文献看,习得顺序原因解释仍有待改进。目前存在的问题是大部分研究人员都忽略了对习得顺序原因的解释。根据我们收集到的文献,关于习得顺序的研究中仅有4篇文献探究了习得顺序原因或者影响因素。比如施家炜(1998)认为普遍语法“参数重设”、认知难易程度、语言输入的时间、数量、频率、语言结构在二语中的使用频率与广度、语言标记和语言教学等是决定或影响习得顺序的重要因素。彭家法(2009)认为语言间的差异也是影响二语习得顺序非常重要的因素。周小兵、刘瑜(2010)认为特征凸现度、形式意义关系复杂度、语言结构复杂度、语篇诱发因素和语用功能复杂度会影响到语法点的认知难度,进而影响习得顺序。因此,在今后的习得顺序研究中我们也应加强对影响习得顺序因素的考察,从而为汉语第二语言习得提供科学的教学策略。

总的来说,目前汉语第二语言习得在用例分析和习得顺序研究方面仍有不足之处,而这与研究理念、依据的理论方法、语料库的建设都有密切关系。自20世纪80年代以来,偏误分析一直是汉语第二语言习得研究的热门领域,而偏误分析本质上来说是一种归纳法,在这种理念的影响下,研究人员缺乏中介语是动态发展的观念,往往只关注汉语学习者出现偏误的用例,而不会去考察正确的语言现象,更不会归纳留学生正确用例的类型。此外,现有的、免费开放的大型汉语中介语语料库HSK动态作文语料库只收录了中高级阶段留生产出的语料,缺少初级阶段语料,且只做了字、词、句、篇、标点符号5个层面的偏误标注,检索系

统也不够完善，常常检索到不符合要求的语料，这些都会对习得顺序研究和正确用例分析产生极大阻碍。

6 语料库方法的作用和局限

语料库方法的优势很明显。首先，张宝林（2011）指出，该方法能使研究结果具有客观性和普遍性。国内现有的语料库规模都较为庞大，至少收录了几十万字以上的语料。留学生自然、真实产出的语料为研究提供了足够的样本，保证了研究结果的普遍性和客观性。其次，毛文伟（2011）指出，语料库方法能更加全面地反映学习者的习得情况，把握不同阶段学习者产出的特征和发展轨迹。语料库收集了汉语学习者不同阶段正确和偏误的语料，可以反映中介语的全貌和动态发展轨迹，便于研究人员把握学习者的语言特征。

语料库方法在应用过程中也暴露出一些不足之处。首先就该方法本身而言，它应用的研究领域有一定限制，像学习动机、学习策略这些内容是无法进行考察的。此外，它也无法考察学习者有意采取“回避”策略或因“回避”偏误而缺失的语料，在这种情况下可以选择问卷调查的方式代替或补充。其次，目前该方法在汉语二语习得中的应用存在着研究内容不平衡的情况。现有研究都集中于学习者语言特征，对习得机制、学习者外部因素等方面的研究相对较要少得多。而在学习者语言特征上也存在研究内容不平衡的情况，研究偏误分析的文献远多于习得顺序，而像语用特征方面的研究在我们收集到的文献中还尚未看到。此外，上节中我们也提到，目前很多研究在用例分析方面比较欠缺。我们认为语料库方法之所以会出现上述问题，主要与语料库建设及研究理念有关。语料库系统的缺陷会影响该方法的使用，比如语言的语用特征研究成果极少，这与国内中介语语料库目前缺少语用层面的标注有关。而习得顺序、用例分析研究不足，也与语料库语料不完备、语料标注、检索功能不完善有关。当然，研究人员的研究理念也对研究内容有所影响，如前文所述，归纳法的理念对二语习得研究人员产生了极大的影响，使得大家都忽视了其他的研究领域和研究内容。

针对语料库方法存在的问题，我们主张通过完善语料库建设和转变研究理念加以解决。工欲善其事，必先利其器，语料库建设方面可以扩大语料来源，增加语义、语用层面的标注，改进检索系统等，为语料库方法的应用提供良好的基础保障。在研究理念方面，除了使用归纳法，研究人员还应多尝试使用演绎法或者演绎和归纳相结合的方法进行习得研究。虽然语料库性质与归纳法思想一致，但这不妨碍我们在使用演绎法获得语言规律的假设后，通过语料库搜寻例证。在归纳和演绎法理念的影响下，语料库方法能更多地应用于其他研究领域，比如习得机制的研究等。

总的来说,语料库方法在未来汉语二语习得研究中的应用前景非常可观。首先,随着语料库建设的发展,未来语料库将会拥有更庞大的客观、自然的语料,这是其他研究方法无法实现的,运用语料库方法定量分析得出的数据结果将更加科学可信。其次,丰富的语料来源将使得语料库方法既适用于归纳法,也适用于演绎法。在总结语言使用规律,掌握语言或中介语使用全貌上,语料库方法将成为研究人员的首选,其效果是其他研究方法难以替代的。在对演绎出的语言规则进行例证补充时,也可以通过语料库搜寻相关语料。最后,该方法应用的研究领域将会进一步扩展,语义、语用,以及二语习得的个体差异等方面的研究。

7 结语

本文回顾了语料库方法的发展历程,主要经历了三个阶段,即起步期、发展期和现阶段的飞跃期。文章以前人的理论为依据,结合相关文献,梳理了现有汉语二语习得研究中语料库方法的使用情况,总结概述了近年来该方法在语料收集、偏误分析研究方面取得的成果,以及在用例分析、习得顺序研究等方面仍存在的问题,并对语料库方法的优势和局限性作了探究。对于语料库方法在使用中存在的问题,本文提供了相应的解释和解决途径,以期该方法在未来汉语二语习得研究中能得到更好的应用。

参考文献

- 毕晋,肖奚强,程仕仪.新世纪以来汉语作为第二语言习得研究成果分析——基于四份CSCI中国语言学来源期刊文献的统计[J].语言与翻译,2017(4):74-82.
- 蔡淑美,施春宏.基于汉语中介语语料库的二价名词习得研究[J].语言文字应用,2014(2):85-95.
- 曹贤文.二语习得研究“需求侧”视角下的汉语学习者语料库建设[J].华文教学与研究,2020(1):38-46.
- 陈小荷.跟副词“也”有关的偏误分析[J].世界汉语教学,1996(2):54-60.
- 程燕,肖奚强.韩国留学生汉语成语使用状况考察[J].汉语学习,2020(2):95-103.
- 崔希亮.欧美学生汉语介词习得的特点及偏误分析[J].世界汉语教学,2005(3):83-95.
- 冯丽萍,孙红娟.第二语言习得顺序研究方法述评[J].语言教学与研究,2010(1):9-16.
- 高立群.外国留学生规则字偏误分析——基于中介语语料库的研究[J].语言教学与研究,2001(5):55-62.

- 黄月圆, 杨素英, 高立群, 等. 汉语作为第二语言反身代词习得考察[J]. 汉语学习, 2005 (5): 49-60.
- 李榕, 王元鑫. 中高级阶段韩国留学生汉语篇章第三人称回指的习得研究[J]. 世界汉语教学, 2021 (2): 276-288.
- 刘旭. 泰国大学生汉语名词习得机制探析——以名词句法功能习得为中心[J]. 语言文字应用, 2018 (3): 114-123.
- 刘珣. 对外汉语教育学引论[M]. 北京: 北京语言文化大学出版社, 2000.
- 鲁健骥. 外国人学汉语的语法偏误分析[J]. 语言教学与研究, 1994 (1): 49-64.
- 毛文伟. 二语习得量化研究中两种数据采集方法的对比研究[J]. 日语学习与研究, 2011 (1): 12-18.
- 牛长伟, 李君. 汉语中介语“什么”类代词的偏误分析及教学对策[J]. 汉语学习, 2019 (4): 94-102.
- 彭家法. 对外汉语个别实习中的语法教学[J]. 皖西学院学报, 2009 (3): 125-129.
- 彭淑莉. 汉语动词带宾语“被”字句习得研究[J]. 汉语学习, 2008 (2): 91-99.
- 施家伟. 外国留学生22类现代汉语句式的习得顺序研究[J]. 世界汉语教学, 1998 (4): 77-98.
- 施家伟. 国内汉语第二语言习得研究二十年[J]. 语言教学与研究, 2006 (1): 15-26.
- 汪宗虎. 词语教学中的病句分析和批改[J]. 语言教学与研究, 1985 (3): 80-86.
- 王建勤. “不”和“没”否定结构的习得过程[J]. 世界汉语教学, 1997 (3): 92-100.
- 吴继峰. 形容词AABB重叠式的习得研究[J]. 汉语学习, 2013 (3): 91-95.
- 肖奚强, 周文华. 外国学生汉语趋向补语句习得研究[J]. 汉语学习, 2009 (1): 70-81.
- 谢福. 基于语料库的留学生“是……的”句习得研究[J]. 语言教学与研究, 2010 (2): 17-24.
- 杨圳, 施春宏. 汉语准价动词的二语习得表现及其内在机制[J]. 世界汉语教学, 2013 (4): 558-573.
- 袁博平. 计算复杂性与第一语言迁移——以汉语第二语言态度疑问句为例[J]. 世界汉语教学, 2017 (1): 85-104.
- 张宝林. 回避与泛化——基于“HSK 动态作文语料库”的“把”字句习得考察[J]. 世界汉语教学, 2010 (2): 263-278.
- 张宝林. 外国人汉语句式习得研究的方法论思考[J]. 华文教学与研究, 2011 (2): 23-29.
- 张宝林. 汉语中介语语料库检索系统透视[J]. 天津师范大学学报(社会科学版),

2021 (6): 29-37.

张利蕊. 欧美留学生汉语有标复句习得顺序研究[J]. 华文教学与研究, 2018 (3): 78-87.

周小兵, 邓小宁. 两种“得”字补语句的习得考察[J]. 汉语学习, 2009 (2): 65-71.

周小兵, 刘瑜. 汉语语法点学习发展难度[J]. 华文教学与研究, 2010 (1): 24-29.

通信地址: 230000 安徽省合肥市 安徽大学文学院

CQP 语法赋能语言研究及语言学习

湖南工商大学 吴良平

提要：CQP 语法（CQP syntax）是第四代语料库检索平台 CQPweb 所使用的高级检索语法，支持正则表达式和布尔运算，可满足多种复杂查询需求。CQPweb 为全球数十所大学和科研院所采用，CQP 语法为其精华所在。本文对 CQP 语法检索模型和相关概念分解简化，并从词汇、短语和语法等语言学诸层面展示 CQP 语法丰富检索功能，以推动 CQPweb 在教学科研中的进一步深入应用。

关键词：CQPweb、CQP 语法、语料库检索、数据驱动学习

1 引言

得益于功能强大、承载语料丰富和免费开源，第四代语料库检索平台 CQPweb（Hardie 2012）近年来在语言研究和教学应用中使用日益广泛。当前，世界各地 CQPweb 平台已有数十个，新的采用机构包括伯明翰大学、赫尔辛基大学、悉尼大学等。各平台语料资源丰富且大多向公众开放，既有 BNC 等传统大型平衡语料库，也有各教学科研单位的自建语料库，被广泛应用于数字人文（Fischer *et al.* 2020）、批评话语分析（Baker *et al.* 2019）、社会语言学（Sadowsky 2022; Müller *et al.* 2021）等研究领域。相关教学应用则集中在数据驱动学习（杨素香 2015；刘萍等 2016；刘萍、吴良平 2016；罗琴琴、石敏 2020）和教材开发（Curry *et al.* 2022）。

CQPweb 检索功能强大，但潜力没有充分释放。CQPweb 包含简单查询模式和复杂检索模式，前者支持通配符进行简单查询，后者使用 CQP 语法，支持利用正则表达式和布尔运算进行高级查询，是 CQPweb 检索功能核心与精华所在（许家金、吴良平 2014）。囿于 CQP 语法概念模型相对复杂、相关文献不足，且 CQPweb 必须基于 Linux 操作系统进行安装这些难题，CQP 语法的使用并不广泛，当前基于 CQPweb 的研究与应用仍多停留在大家较为熟悉和容易掌握的基于通配符的简单查询。然而，随着计算机虚拟化技术发展，CQPweb 安装技术门槛已大幅降低（葛晓帅、张现荣 2021），普通用户稍加学习或培训后即可在 Windows 平台安装好一个加载自己语料库的全新系统，用时可能不需要 1 小时，今后 CQPweb 无疑将为更多机构和个人所采用。随着 CQPweb 用户增长，功能更为全面、强大的 CQP 语法将在更大范围内得以使用。

为便于研究人员、教师和学生使用CQPweb更好地开展研究和教学应用,本文对CQP语法检索模型和相关概念分解简化,对检索过程中的难点予以剖析,并从词汇、短语和语法等语言学诸层面展示CQP语法的丰富检索功能,弥补相关使用资料短缺的不足,以推动CQPweb这一语料库分析利器在教学科研中的进一步深入应用。

2 CQP语法与CQPweb

在CQPweb词表生成(Frequency lists)、索引分析(Query/Concordance)、词组搭配计算(Collocations)、主题词分析(Keywords)等几大功能中,CQP语法在索引分析中居中心地位。这是由于CQPweb在处理索引分析的简单查询模式(Simple Query)时,自动将检索表达式转换为复杂检索模式所支持的CQP语法后再对语料库进行检索(Hardie 2012: 394)。与此同时,CQP语法对基于索引分析结果的词语搭配分析也有着重要影响。

CQP语法历史悠久,历久弥新,但使用资料尤其是中文资料依然相对匮乏。CQP语法原为德国斯图加特大学Christ为“语料库工作台”(Corpus Workbench, 简称CWB)定制的语料库检索语法(Christ 1994),CQP即为Corpus Query Processor的缩写,意指“语料库查询处理器”。这一语法支持正则表达式和布尔运算,专为语料库检索开发,为众多语料库工具所采用。Hoffmann *et al.* (2008)撰写了《基于BNCweb的语料库语言学实践教程》一书,其中第12章专门论述CQP语法,这可能是现有正式出版物中可查询到的最全面的CQP语法使用指南。

CQPweb所使用的CQP语法基于开源后的语料库工作台,因此也可参考语料库工作台及其后续版本的一些使用手册(Evert & The OCWB Development Team 2022)。出于检索平台安全需要,CQPweb的CQP语法仅支持语料库工作台的查询语法,舍弃了所有命令动词,如set、show、dump、sort、define等(许家金、吴良平 2014: 15),因此使用上不能全盘照搬已有文献。

综上所述,CQP语法使用文献相对匮乏,或源于专著难于获取,或源于手册内容冲突,或源于检索模型本身就比较复杂,致使部分教研实践中遇到了一些难题(刘萍等 2016; 罗琴琴、石敏 2020)。本文尝试对CQP语法检索模型和相关概念分解简化,以进一步降低其学习和应用难度。

CQPweb默认检索模式为简单查询,复杂检索模式下的CQP语法须通过手工选取CQP syntax(见图1)开启。简单查询检索表达式与CQP语法之间的转换可先点击检索界面的Query history(见图1),然后点击后续页面中的Show in CQP syntax和Show as Simple Query切换观察,熟悉简单查询的用户可以充分利用这一便利了解两者之间异同与转换规律。和简单查询相比,CQP语法灵活全面,可

以完成更多检索任务，本文仅讨论 CQP 语法。本文与 CQPweb 相关图片均截取自 BFSU CQPweb（<http://114.251.154.212/cqp/>，用户名 test，密码 test）（许家金、吴良平 2014）。

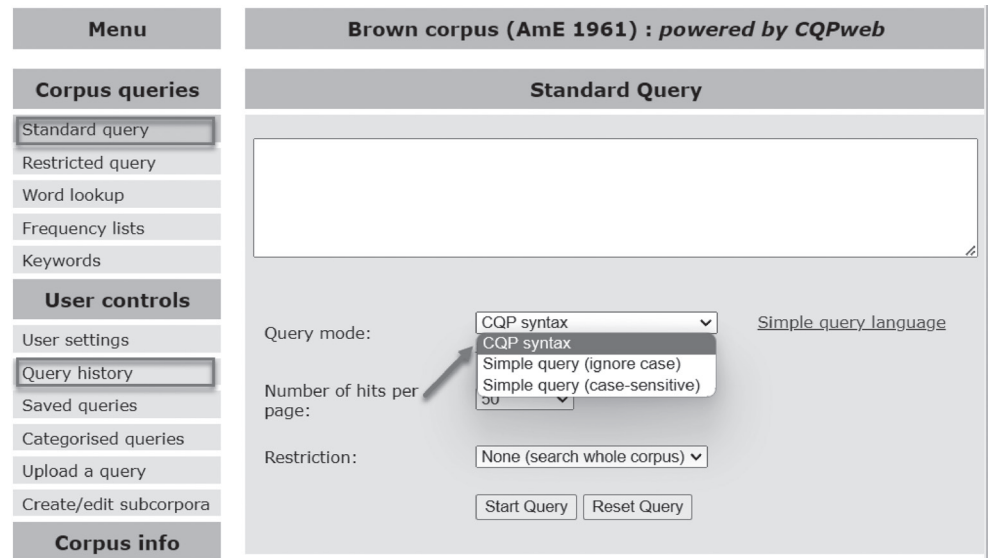


图1 CQP 语法检索界面

3 CQP 语法检索模型与相关概念

理解和使用 CQP 语法通常有两种路径：一种是根据大量样例学习，自下而上逐步归纳提炼，熟能生巧；另一种从检索语法所依赖的数据模型出发，从思想和概念上明晰语法各组成成分及相互关系，自上而下掌握其核心要领，然后逐一掌握具体检索细节。本文认为这两种方式各有利弊，综合起来可能效果最佳。本节对 CQP 语法检索模型和相关概念分解简化，第 4 节举例说明词汇、短语和语法等语言学诸层面的检索。阅读过程中如有疑问，可以在两部分之间来回跳读，以获取最佳阅读效果。

和 AntConc、WordSmith Tools 等众所熟知的语料库检索软件不同，CQP 语法检索文件为竖排格式，有自己独有的一些概念需要提前了解，其数据模型见图 2，与检索语法之间关系如下：（1）<text> 和 </text> 两行一般不参与检索；（2）<s> 和 </s> 相关文献中称之为结构属性（s-attribute），可参与检索，但一般不单独使用；（3）剩余部分称之为位置属性（p-attribute），每列占据一个位置，其中第一列为词形属性，默认名称为 word，第二列为词性属性，实际操作中通常命名为 pos。位置属性是 CQP 语法检索主体。我们接下来从单词检索和多词检索两个相辅相成

的方面讨论可能的最佳检索策略。

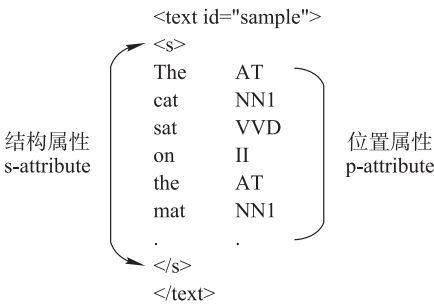


图2 CQP 语法检索文件数据模型图

3.1 单词检索

使用CQP语法对单个单词的位置属性进行检索时，一次可以只检索一个属性，如单独检索其词形属性或词性属性，如[word="cat"]或[pos!="AT"]，这些表达式均可产生输出结果，其中!=表示不等于。针对单个单词的单一属性的检索，我们可用表1中的“属性-值”配对表来说明表达式各构成成分。

表 1 CQP 语法最简表达式

属性	运算符	值
word	=	cat
word	=	mat
word	=	sat
pos	=	VVD
pos	!=	AT

从表1可看出，属性是固定的，但值可以变化。如果要对值进行模糊匹配，就需要引入正则表达式，如[word=".at"]可以一次性检索到cat、mat和sat，其中.号表示任意单个字符。正则表达式是CQP语法功能强大的第一个原因。

然而，上面仅展示了如何检索单个单词的单一属性，如词形或词性，要同时检索单个单词的多个属性，则需要引入CQP语法功能强大的第二个原因，即布尔运算。以[word=".at"& pos="VVD"]为例，检索结果里就只剩sat，而没有了cat和mat，布尔算符&（“和”）起到了预期的过滤效果。CQP语法的三个布尔运算符为

“和”(&), “或”(|), “否”(!), 其功能在4.1节将进一步详细阐述。

3.2 多词检索

了解了单个单词的检索, 多词检索可迎刃而解: 多词检索表达式就是单个单词检索表达式的横向相加。如果将单个单词表示为一个盒子, 那么多词检索就是多个盒子的横向并置, 其中每个盒子内部支持正则表达式和布尔运算, 如图3所示。

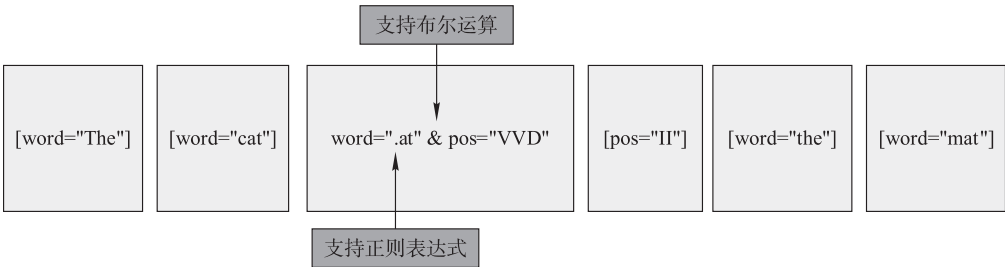


图3 CQP 语法多词检索示意图

4 检索实例

依据上述CQP语法检索模型, 我们可以对不同语言学层面现象进行灵活检索。我们从单词检索(4.1节)、多词检索(4.2节)、语法检索(4.3节)、特殊检索形式(4.4节)以及常见检索问题(4.5节)五个方面加以讨论。由于单个单词的检索是其他语言层面现象检索的基石, 4.1节单词检索内容比较丰富, 将分别展示单个单词按词形检索、按词性检索、按其他标注属性检索、单词单一属性的模糊检索以及单词多个属性的同时联合检索的具体实例, 从实践层面深化对CQP语法检索模型的认识。

4.1 单词检索

(1) 单词按词形检索

CQP语法在CQPweb界面只有一个检索入口, 默认是大小写敏感的, 这与简单检索不同, 后者在检索界面提供了区分大小写(case-sensitive)和不区分大小写(ignore-case)两个不同检索入口。CQP语法中单词词形检索表达式为[word="待检索词"], 如[word="China"]。因为词形属性(即word属性)为单词默认属性, 所以[word="China"]也可简写为"China", 图4是在布朗语料库中的检索结果。需要说明的是, 显性地标记单词各属性有利于人工阅读复杂检索表达式或进行除错, 因此下文按词形检索时, 如不涉及排版美观需要均采用完整表达式而非其简

写形式。

1	<u>brownl_a</u>	weapons , particularly to Communist	<u>China</u>	. The question arose as
2	<u>brownl_a</u>	steamed off into the South	<u>China</u>	Sea , accompanied by a
3	<u>brownl_a</u>	In conferences with Nationalist	<u>China</u>	's dapper , diminutive Vice
4	<u>brownl_a</u>	to the admission of Red	<u>China</u>	to the United Nations .

图4 CQP 语法表达式[word="China"]检索结果

单词检索时如忽略大小写，需要在检索表达式中明确添加“%c”标记，将其置于双引号和右括号之间，如[word="China"%c]，其中%表示否定,c表示case(大小写)，新表达式检索结果如图5所示。不难看出，结果中既有表示瓷器的小写字母开头的 china，又有表示国家名称的大写字母开头的 China。

51	<u>brownl_e</u>	of a bull in a	<u>china</u>	shop . Recently I was
52	<u>brownl_f</u>	<small>小写</small> in Russia , Japan in	<u>China</u>	again . They were always
53	<u>brownl_f</u>	northwest , close to Communist	<u>China</u>	(map , page 250
54	<u>brownl_f</u>	tung (nuts from the	<u>China</u>	wood-oil tree) , perilla

图5 CQP 语法表达式[word="China"%c]检索结果

(2) 单词按词性检索

如果语料库标注有词性，CQP语法也可针对单词的词性进行检索。例如，表达式[pos="JJ"]可检索布朗语料库中的形容词原级，如图6所示。

19	<u>brownl_a</u>	was also recommended by the	<u>outgoing</u>	jury . It urged that
20	<u>brownl_a</u>	enabling funds and re-set the	<u>effective</u>	date so that an orderly
21	<u>brownl_a</u>	effective date so that an	<u>orderly</u>	implementation of the law may
22	<u>brownl_a</u>	be effected " . The	<u>grand</u>	jury took a swipe at

图6 CQP 语法表达式[pos="JJ"]检索结果

单词按词性检索的难点在于需要输入准确的词性赋码，JJ仅可检索出形容词原级，如需检索形容词比较级和最高级，表达式须分别修改为[pos="JJR"]和[pos="JJT"]。那么，为什么要如此修改呢？了解一个语料库的词性赋码集通常有两个途径：（1）根据语料库在线文档进行查找；（2）根据语料库自身进行查找。

BFSU CQPweb中的布朗语料库在线文档齐备，显示其词性标注程序是CLAWS，标注集为CLAWS7，形容词原级、比较级、最高级通过点击界面CLAWS7后显示分别为JJ、JJR和JJT，这就是上面不同形容词检索表达式的书写依据。

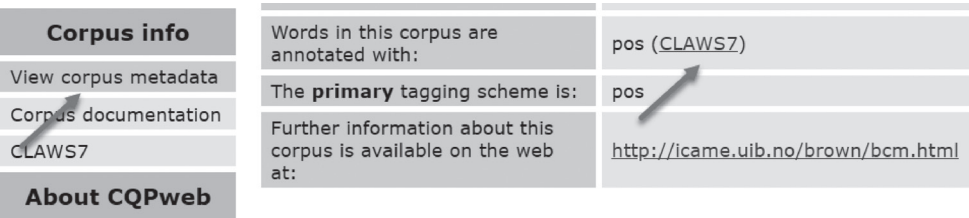


图7 在CQPweb在线查找语料库词性标注集

实践中，如果一个语料库在线文档给出的标注集不是那么明朗，也可根据语料库自身信息进行查找，这里以北外多语种布朗新闻语料库进行说明。如图8所示，北外多语种布朗新闻语料库涵盖的语言非常广泛，包括加泰罗尼亚语（caBrown）、丹麦语（daBrown）、德语（deBrown）、芬兰语（fiBrown）等多国语言，我们选取多数人不熟悉的芬兰语布朗新闻语料库举例说明。



图8 BFSU CQPweb上的北外多语种布朗新闻语料库

BFSU CQPweb界面显示芬兰语布朗新闻语料库（fiBrown Press）有词性赋码（tagged）。一种快速了解其标注集的方法为利用CQPweb的词表生成（Frequency lists）功能，该功能除了能生成词表外，还可生成词性赋码表，操作步骤和结果分别如图9和图10所示。

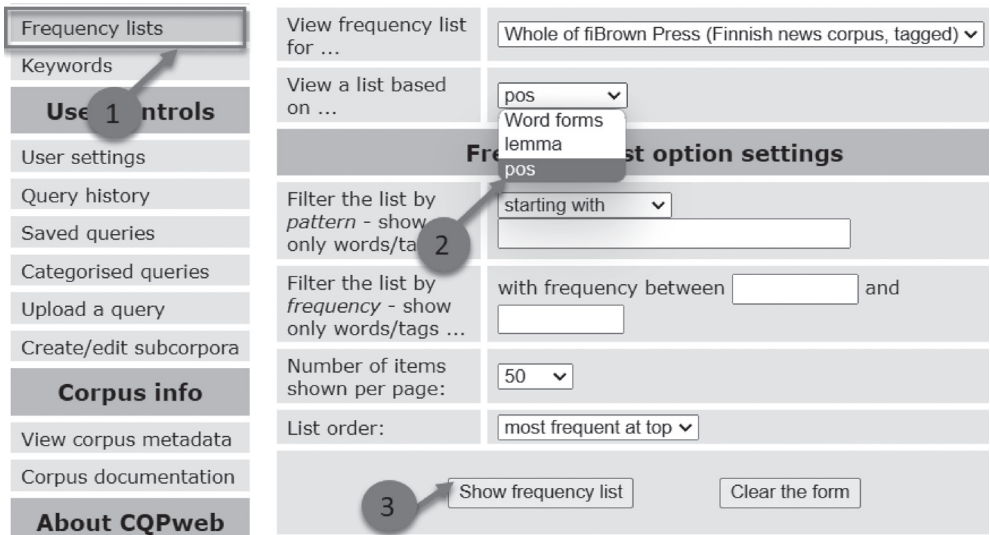


图9 利用 CQPweb 词表生成功能查询词性赋码表

1	<u>Punct</u>	16,582
2	<u>N_Nom_Sg</u>	15,226
3	<u>SENT</u>	14,823
4	<u>Adv</u>	14,067
5	<u>N_Gen_Sg</u>	11,885
6	<u>V_Prs_Act_Sg3</u>	11,091
7	<u>CC</u>	7,273

图10 芬兰语布朗新闻语料库词性赋码表

词性赋码表生成后，可以点击各赋码标签进入语料库查找相应索引行，这样可以通过归纳法了解各标签含义。与此同时，也可以根据在线文档顺藤摸瓜，了解更多赋码细节。以图10中数量排第二位的N_Nom_Sg为例，从命名规则来看极有可能表示单数（Sg= Singular）名词（N=Noun/Nominal），这一点可以根据索引行实例初步证实。针对中间意义不那么透明的标签Nom，在线文档显示标注工具为TreeTagger，进一步查找TreeTagger官网和芬兰语相关的标注集后可知其代表Nominative，为芬兰语名词的15个格之一，至此该标签整体意义基本探明。如果需要按照词性赋码检索自己不熟悉的语料库，可参考上面芬兰语布朗新闻语料库的检索信息查询方案。

（3）单词按其他标注属性检索

如果一个语料库标注有除词性外的其他属性，如原形词（lemma），同样可参照词性赋码的检索方案进行查询。以BFSU CQPweb上加载的AmE Brown Family

Corpora 为例，该语料库标注了每个单词的原形词属性，属性名称为 lemma，表达式 [lemma="have"%c] 可检索出 had、have、has、'd、've、having、's (He's got a lot of friends)、haves (haves and have-nots) 等多种 have 的变体形式。

(4) 单词单一属性的模糊检索

按词形或词性检索单词时，有时并不能精确地描述检索对象。以形态学的能产性研究为例，如要调查 non-、un- 和 in- 三个否定前缀的能产性差异，如果采用语料库研究方法，任何研究者也无法穷尽性地枚举这些前缀开头的所有单词，这时就需要用到正则表达式的模糊匹配功能。表 2 为 CQP 语法中比较常用的正则表达式，为说明方便，pos 的取值范围限定于 CLAWS7 词性标注集。

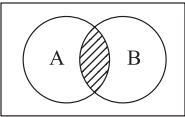
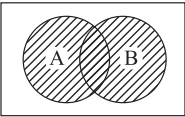
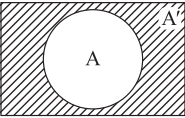
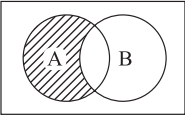
表 2 CQP 语法正则表达式检索简表

符号	符号意义	示例	示例说明
.	任意单个字符	[pos="VBD."]	匹配 VBDR、VBDZ，即 were 和 was
?、*、+、{n}、{m,n}	重复个数，分别表示 0 或 1，0 或更多，1 或更多，n 个、m 到 n 个	[word="colou?r"]	匹配 color 或 colour
		[pos="V.+"]	匹配所有动词
		[word=".*o{2}.*"]	匹配 good, indoor, tools
()	群组符号	[word="(ma)\1.*"]	匹配 mama, mamas 等，其中 \1 表示回指引用第一个群组。
	选择符号	[word="a an"]	匹配 a 或 an
[]	字符集	[word="[ui]n.* non.*"]	匹配 non-、un- 和 in- 开头的所有单词
		[word="[^abc].*"]	匹配非 a, b, c 开头的所有单词

(5) 单词多个属性同时联合检索

理论上 CQP 语法可以对单词的词形、词性、原形词等多个属性同时检索，虽然实践中仅需联合检索词形和词性，这主要依靠布尔运算/集合运算来实现，参见表 3。其中交集运算符为“和 (&)”，并集运算符为“或 (|)”，补集运算符为“否 (!)”，差集运算符为单个运算符的组合。

表3 CQP语法布尔运算简表

运算类型	图示	示例	示例说明
交集		[word="test"& pos="N.*"]	检索语料库中词形为test <u>且</u> 词性为名词的所有单词
并集		[word="test"]][word="testing"]	检索语料库中词形为test <u>或</u> 词形为testing的所有单词
补集		[word!="test"]	检索语料库中词形 <u>不</u> 是test所有单词
差集		[word="test"& pos!="N.*"]	检索语料库中词形为test <u>且</u> 词性 <u>不是</u> 名词的所有单词

4.2 多词检索

CQP语法支持多词序列或短语的检索，可处理连续型短语、非连续型短语以及非连续且位置变异型短语，非常方便。下面试各举例说明。

(1) 连续型短语

连续型短语即计算语言学中的n元组 (n-gram)，语料库语言学也称之为词簇 (lexical bundle) (Biber *et al.* 2004)，CQP语法通过检索连续任意词的占位符[]来实现，换句话说，即检索连续多个中间不带空格的成对中括号[]。例如，要检索三元组或三词词簇，检索表达式可写为[] [] []，要检索四元组或四词词簇，检索表达式相应写为[] [] [] []。这一检索方式显然比使用常规正则表达式更为简洁高效。

(2) 非连续型短语

典型的非连续型短语包括跨词序列 (skipgram) 和搭配框架 (collocational framework) (Renouf & Sinclair 1991)。计算语言学家使用跨词序列来描述非连续单词之间的共现 (Wilks 2005)，这与语料库语言学的搭配框架非常类似。Renouf & Sinclair (1991) 根据伯明翰语料库对典型的搭配框架“a+?+of”“an+?+of”“be+?+to”“too+?+to”“for+?+of”和“many+?+of”进行了详细描写并据此对既有语言理论提出了挑战。要检索搭配框架，以a+?+of为例，CQP语法检索表达式可写为

[word="a"%c] [word=".*"] [word="of"]，或简写为"a"%c [word=".*"] “of”，其在布朗语料库中的检索结果如图 11 所示。

1	brown1_a	The grand jury commented on	<u>a number of</u>	other topics , among them
2	brown1_a	in experienced clerical personnel as	<u>a result of</u>	city personnel policies " .
3	brown1_a	Senate is expected to approve	<u>a study of</u>	the number of legislators allotted
4	brown1_a	would be issued every time	<u>a portion of</u>	the old ones are paid

图 11 CQP 语法检索搭配框架 a+?+of 结果

(3) 非连续、且位置变异型短语

框合结构 (conccgram) 为典型非连续、且位置变异型短语，这类短语由两词或多词构成的所有组合组成，具有构成成分 (AB, ACB) 和位置 (AB, BA) 的双重变异特征 (Cheng *et al.* 2006)。Cheng *et al.* (2006) 发现，由 call/what 构成的双词型框合结构在香港口语英语语料库 (HKCSE) 中仅构成成分变异，两词间变异成分只有 1-3 个插入词，绝大多数出现在非疑问句，主要起重新表述言说内容或基于已说内容引入新话语的功能。使用 CQP 语法在布朗语料库这一书面英语语料库中检索这一框合结构，表达式可写为 ("what"[* "call"])("what"[* "call"]) within s，其中竖线 “|” 表示或者，[* 表示任意个单词，within s 表示将检索范围限定在句子以内，不跨句检索，括号 () 起分组作用，检索结果如图 12 所示。

No	Filename	Solution 1 to 14 Page 1 / 1	
1	brown1_a	campaign on the carcass of	<u>what they call</u> Eisenhower Republicanism , but the
2	brown1_b	when universities and colleges schedule	<u>what they call</u> " Homecoming Day " .
3	brown1_b	1770 's were mercenaries ,	<u>what shall we call</u> the UN troops sent to
4	brown1_d	Newman saw that liberalism (<u>what we now might call</u> secularism) would gradually but
5	brown1_f	may never be encountered in	<u>what we are pleased to call</u> the real life . Dr
6	brown1_g	a man 's flair for	<u>what the professionals in this field call</u> " scenarios " . As
7	brown1_g	used his polarity to illustrate	<u>what I thought had happened to us in that form of liberalism we call</u> Progressivism . It seemed to
8	brown1_g	to represent the viewpoint of	<u>what I would call</u> the unconscious liberal , but
9	brown1_g	tends to give a person	<u>what I shall call</u> depth . I use this
10	brown1_g	or conclusive . Apropos of	<u>what some would call</u> cynicism , I remember an
11	brown1_g	the world ; he displays	<u>what outlanders call</u> the New York mind ,
12	brown1_j	once that these animals are	<u>what we call</u> " queens " , young
13	brown1_p	I tried to believe that	<u>what must have happened was that , restless , disturbed by this telephone call</u> or whatever , she walked
14	brown1_p	I think I was	<u>what you might call</u> a convivial man , and

图 12 CQP 语法检索 call/what 双词框合结构结果

从图 12 可以看出，书面英语中，call/what 构成的框合结构的词序、功能及功能分布与英语口语基本一致，但插入词变异性更大。考虑到除第 7 例和第 13 例外，插入词数量在 1-5 词之间，因此上面检索表达式也可改写为 ("what"[*{1,5}

"call")("what"[]{1,5}"call") within s, 其中 {1,5} 表示插入词数量限定在 1—5 个, 新表达式检索结果如图 13 所示。这表明, CQP 语法中正则表达式的数量限制符也可作用于单词整体, 而不仅局限于刻画单词内部字母的数量特征, 这一点对灵活控制变长型短语的检索非常有利。

1	<u>brownl_a</u>	campaign on the carcass of	<u>what they call</u>	Eisenhower Republicanism , but the
2	<u>brownl_b</u>	when universities and colleges schedule	<u>what they call</u>	" Homecoming Day " .
3	<u>brownl_b</u>	1770 's were mercenaries ,	<u>what shall we call</u>	the UN troops sent to
4	<u>brownl_d</u>	Newman saw that liberalism (<u>what we now might call</u>	secularism) would gradually but
5	<u>brownl_f</u>	may never be encountered in	<u>what we are pleased to call</u>	the real life . Dr
6	<u>brownl_g</u>	a man 's flair for	<u>what the professionals in this field call</u>	" scenarios " . As
7	<u>brownl_g</u>	to represent the viewpoint of	<u>what I would call</u>	the unconscious liberal , but
8	<u>brownl_g</u>	tends to give a person	<u>what I shall call</u>	depth . I use this
9	<u>brownl_g</u>	or conclusive . Apropos of	<u>what some would call</u>	cynicism , I remember an
10	<u>brownl_g</u>	the world ; he displays	<u>what outlanders call</u>	the New York mind ,
11	<u>brownl_j</u>	once that these animals are	<u>what we call</u>	" queens " , young
12	<u>brownl_p</u>	I think I was	<u>what you might call</u>	a convivial man , and

图 13 CQP 语法检索 call/what 双词框合结构新结果

4.3 语法检索

依语料库语言学理论观点, 词汇和语法是一个连续体, 不能截然分开。抛开理论上的分歧不谈, 当前 CQP 语法也支持传统语法概念的检索, 方法是通过检索词性赋码序列部分模拟对语法范畴的检索, 虽然在书写表达式时可能需要发挥敏锐的观察力并反复试错。

以英语宾语从句检索为例, 如果语料库词性标注采用了 CLAWS, 检索表达式可写为 <s> []* [pos="VV.*"] [word="that"& pos="CST"] []*</s>, 其中 <s> 和 </s> 分别为句首和句尾标记, []* 表示任意个单词, [pos="VV.*"] 表示所有实义动词, 不包含情态动词、be 动词、do 动词、have 动词, [word="that"& pos="CST"] 表示词形为 that 的连词。

上述检索表达式也可写为 [pos="VV.*"][word="that"] [:pos!="NN1|J.*:"], 其中 [:pos!="NN1|J.*:"] 表示 that 后不允许出现单数名词或形容词, 以避免 I like that kid 或 I like that lovely kid 这样的句子被误检, 同时 [::] 的特殊标记表示其内部所有成分仅作检索时的限制条件, 不作为返回结果。英语宾语从句第二种检索表达式的返回结果如图 14 所示。

No	Filename	Solution 1 to 30	Page 1 / 1
1	brownl_a	Ala &. The petition	<u>said that</u> the couple has not lived
2	brownl_a	an open date , Hemus	<u>said that</u> Bob Nieman definitely would stay
3	brownl_a	at its annual meeting Friday	<u>noted that</u> state taxing requirements at present
4	brownl_b	for peaceable accommodation . They	<u>realize that</u> by acting in concert ,
5	brownl_b	is a Buddhist , which	<u>means that</u> to him peace and the
6	brownl_e	thing for the engineer to	<u>accept that</u> he can go as far
7	brownl_f	Some hymens are so strongly	<u>developed that</u> they can not be torn

图 14 CQP 语法检索英语宾语从句结果

需要注意的是，无论是检索表达式一还是检索表达二，显然都没有考虑英语宾语从句标记词 *that* 省略的情况，所以实际应用中还可将 CQP 语法检索结果和 Tregex (Levy & Andrew 2006) 等专门的语法检索工具的检索结果进行比对，以进一步优化检索表达式。考虑到学术英语中宾语从句标记词 *that* 通常为显性，如果仅研究学术英语或为相关课堂教学活动提供报道动词的使用样例，那么上述两种检索表达式均有积极参考意义。

4.4 特殊检索形式

CQP 语法的特殊检索形式主要有以下几种：（1）零宽度条件判断符；（2）语言成分边界标记；（3）回指功能标签；（4）自定义宏。

零宽度条件判断符的形式为 `[:]`，其所包含的内容对检索加以条件限制，但限制部分不在检索结果索引行节点词位置显示，可以起到保持包含条件的检索与常规检索输出结果形式一致的作用。检索样例可参见前述对英语宾语从句的第二种检索方法。

语言成分边界标记的一般形式为 `within x`，其中 `x` 可为句边界标记 `s`（`s` 为 CLAWS 软件标注的句边界标记）或其他任意语言成分边界标记，如话轮标记 `u`，这些标记的具体名称在准备和安装语料库时指定，其主要功能是避免跨句或跨话轮等跨越语言成分边界的检索。检索样例可参见 4.2 节对框合结构的检索。

自定义宏可简化检索表达式的书写，实现自定义语义、语法或其他 CQP 语法支持的各类检索。例如，先通过 CQPweb 界面的 User settings->Create a new CQP macro 定义宏名 `emotion` 和宏体 `[word ="happy|angry|sad|excited|anxious"]`，然后回到 CQP 语法检索界面输入 `/emotion[]`，即可一次性检索所有自定义情感形容词。类似地，也可以定义宏名 `conj` 和宏体 `[pos ="CC|CCB|CS|CSA|CSN|CST|CSW"]`，然后回到 CQP 语法检索界面直接输入 `/conj[]`，即可一次性检索 CLAWS 标注过的语料中的所有连词。本文所有检索表达式如果日常使用中需要反复使用，均可自定义为单独的宏以实现快速访问。需要特别提醒的是，如果宏的书写格式不正确或包含

特殊字符，有可能会造成检索系统崩溃，这时删除宏重建即可解决问题。

CQP语法中的标签具有回指功能，例如表达式“a” a:[] “is” “a” [word=a. word]可以检索“a X is a X”这样的同义反复结构（tautology），如But a penalty is a penalty, whether it is minute 94 or minute one。图15为使用该表达式检索The Independent Corpus的部分结果，这类结构在新闻语篇中似乎比较多见。

1	<u>2009_03</u>	an internet station , whereas	<u>a pirate is a pirate</u>	. It 's like if
2	<u>2009_03</u>	but I always thought that	<u>a designer is a designer</u>	and if someone asked me
3	<u>2009_04</u>	the reason why it is	<u>a secret is a secret</u>	, " he concluded .
4	<u>2009_04</u>	's right to say that	<u>a character is a character</u>	throughout the ages , "
5	<u>2009_05</u>	Kingdom Of Fife . But	<u>a winner is a winner</u>	, and this one his
6	<u>2009_06</u>	killer still out there and	<u>a killer is a killer</u>	no matter which way you
7	<u>2009_07</u>	is not . However ,	<u>a hero is a hero</u>	even if he can walk

图15 CQP语法检索同义反复结果

4.5 常见检索错误

在CQPweb系统中使用CQP语法进行检索时，应避免以下几类常见错误：（1）忘记返回CQP语法检索模式；（2）该用半角英文标点符号时用了全角中文标点符号；（3）括号或引号没有成对使用；（4）单词内部使用逻辑符号“|”的时候多个单词或词性赋码误放在多对双引号内。

CQPweb系统中，使用CQP语法完成一次查询开始新查询时，CQPweb会自动返回Simple Query模式，这时须手动切换到CQP syntax模式，否则CQP语法指令无法执行。CQP语法中，所有标点符号，包括双引号、小括号、中括号、大括号，必须全部为英文半角，不能为中文全角。CQP语法中的所有双引号、小括号、中括号、大括号必须成对使用，如果检索出现错误，要根据错误提示确认是否存在单边引号或括号的情形。最后，单词内部使用逻辑符号“|”的时候多个单词或词性赋码必须放在一对双引号内，例如[word="a"|"an"]是错误的，正确形式应为[word="a"|"an"]，同理，[pos="DD"|"DD1"]是错误的，正确形式应为[pos="DD|DD1"]。对于第2、3、4类错误，如果感觉容易出错，可以利用CQPweb界面的检索历史（Query history）复用正确的检索表达式，或自建一个包含常用检索表达式的纯文本文件方便复制粘贴以减少出错几率。

5 结语

CQPweb 支持世界上绝大多数语言的检索, 支持高达 20 亿词规模的单个语料库, 随着 CQPweb 安装难度的大幅降低, CQP 语法也终将从小众走向大众。本文对 CQP 语法检索模型和相关概念进行了分解简化, 指出了一些常见的错误陷阱, 并从词汇、短语和语法等语言学诸层面展示了 CQP 语法的丰富检索功能, 是对 CQP 语法现有使用文献的有力补充。当前国内已有高校根据 CQP 语法的一般指导原则针对自身需求和自建语料库特色创制了校本特色的检索手册辅助教研 (刘萍 2019), 相信未来会有更多简明易懂的类似使用文档问世。

CQP 语法检索模型和理念已融入国际标准 ISO 24623-1 中所描述的 Corpus Query Lingua Franca (ISO 2018), 同时随着底层新语言模型 Ziggurat 的提出和发展 (Evert & Hardie 2015), 未来 CQP 语法将能检索更大规模语料和实现对包括依存语法在内的更多语言层面的检索, 在教学和科研领域的应用前景广阔。

由于 Sketch Engine 和 R 包 RcppCWB 使用了和本文 CQP 语法相似的语料库检索语法, 本文对相关使用者亦有一定参考价值。

参考文献

- BAKER P, BROOKES G, EVANS C. The language of patient feedback: a corpus linguistic study of online health communication [M]. London: Routledge, 2019.
- BIBER D, CONRAD S, CORTES V. If you look at ...: lexical bundles in university teaching and textbooks [J]. *Applied Linguistics*, 2004, 25(3): 371-405.
- BIBER D, JOHANSSON S, LEECH G, et al. Longman grammar of spoken and written English [M]. New York: Pearson, 2000.
- CHENG W, GREAVES C, WARREN M. From n-gram to skipgram to concgram [J]. *International Journal of Corpus Linguistics*, 2006, 11(4): 411-433.
- CHRIST O. A modular and flexible architecture for an integrated corpus query system [C]. *Proceedings of COMPLEX' 94: 3rd conference on computational lexicography and text research*. Budapest: COMPLEX, 1994: 23-32.
- CURRY N, LOVE R, GOODMAN O. Adverbs on the move: investigating publisher application of corpus research on recent language change to ELT coursebook development [J]. *Corpora*, 2022, 17(1): 1-38.
- EVERT S, HARDIE A. Ziggurat: a new data model and indexing format for large annotated text corpora [C] // *Proceedings of the 3rd workshop on challenges in the management of large corpora (CMLC-3)*. Mannheim: Institut für Deutsche Sprache, 2015: 21-27.
- EVERT S, THE OCWB DEVELOPMENT TEAM. CQP query language manual [R/

- OL]. (2022-07-01) [2023-08-01]. https://cwb.sourceforge.io/files/CQP_Manual.pdf.
- FISCHER S, KNAPPEN J, MENZEL K, et al. The Royal Society corpus 6.0 providing 300+ years of scientific writing for humanistic study [C]//Proceedings of the 12th conference on language resources and evaluation (LREC 2020). Marseille: European Language Resources Association, 2020: 794–802.
- HARDIE A. CQPweb - combining power, flexibility and usability in a corpus analysis tool [J]. *International Journal of Corpus Linguistics*, 2012, 17(3): 380-409.
- HOFFMANN S, EVERT S, SMITH N, et al. *Corpus linguistics with BNCweb—a practical guide* [M]. New York: Peter Lang, 2008.
- ISO. ISO 24623-1:2018: Language resource management—corpus query lingua franca [R/OL]. (2018-04-01) [2023-08-01]. <https://www.iso.org/standard/37337.html>.
- LEVY R, ANDREW G. Tregex and Tsurgeon: tools for querying and manipulating tree data structures [C] //Proceedings of the 2006 conference on language resources and evaluation (LREC 2006), Genoa: European Language Resources Association, 2006: 2231-2234.
- MÜLLER M, BARTSCH S, ZINN J. Communicating the unknown: an interdisciplinary annotation study of uncertainty in the coronavirus pandemic [J]. *International Journal of Corpus Linguistics*, 2021, 26(4): 498-531.
- RENOUF A, SINCLAIR J. Collocational frameworks in English [C] //AIJMER K, ALTENBERG B. *English corpus linguistics: studies in honour of Jan Svartvik*. London: Longman, 1991: 128-143.
- SADOWSKY S. The sociolinguistic speech corpus of Chilean Spanish (COSCACH): a socially stratified text, audio and video corpus with multiple speech styles [J]. *International Journal of Corpus Linguistics*, 2022, 27(1): 93-125.
- WILKS Y. 2005. REVEAL: The notion of anomalous texts in a very large corpus [C]. // Tuscan Word Centre international workshop: dial a corpus. Certosa di Pontignano: Tuscany, 2005.
- 葛晓帅, 张现荣. 借助 Docker 容器技术实现 CQPweb 系统的 Windows 部署 [J]. *语料库语言学*, 2021 (2): 148-157.
- 刘萍, 吴良平, 刘丽亚. CQPweb 在 ESP 写作教学中的应用研究 [J]. *外语界*, 2016 (5): 11-19.
- 刘萍, 吴良平. 网络语料库分析系统 CQPweb 的建设及应用—以 HZAU CQPweb 为例 [J]. *中国大学教学*, 2016 (5): 70-75.
- 刘萍. HZAU CQPweb 使用手册 [R/OL]. (2019-12-25) [2023-08-01]. <http://lib.hzau.edu.cn/upload/file/20200319/1584610503206036249.pdf>.
- 罗琴琴, 石敏. 基于 CQPweb 的数据驱动学习在英语写作教学中的应用研究 [J]. 扬

州大学学报(高教研究版), 2020(6): 111-118.

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.

杨素香. CQPweb 在线语料库检索平台及其在外语教学中的应用[J]. 中小学外语教学(中学篇), 2015(7): 1-8.

通信地址: 410205 湖南省长沙市 湖南工商大学外国语学院

汉英人文社会科学文献 平行语料库建设*

福建师范大学 邓劲雷

提要：平行语料库建设可以为翻译研究、翻译实践、语言对比、外语教学等提供数据支持，有着重要的理论和实践价值。国内外虽已建成不少双语平行语料库，但鲜有汉译外的学术文本平行语料库。为服务学术翻译研究和学术外译实践等，我们收集了四十多部汉语学术著作及其英译作品，经过文字识别、校对、句子对齐、赋码等步骤，建成了约一千五百万字词的汉英学术文本平行语料库。此外，我们还为该语料库开发了配套的检索平台。除常见的检索功能外，该平台还支持依存语法检索。

关键词：学术文本、汉英平行语料库、检索平台、依存语法

1 引言

近50年来，语料库凭借语料真实、数据量大、检索便捷等优点深刻地改变了包括翻译在内的语言研究现状。肖忠华（2012：7）认为，语言研究正在经历“语料库‘革命’”。作为语料库语言学的基础工程，语料库建设在这一学科领域中起着举足轻重的作用。目前，研究人员已经建成了诸多规模庞大（百亿字/词）、检索方便的单语语料库，如Davies（2009）建设的美国当代英语语料库（COCA），荀恩东等（2016）建设的BCC语料库。相比单语语料库，双语平行语料库的建设相对滞后，主要原因是“它的构建和加工是很困难的工作”（冯志伟 2010：421）。但是，由于平行语料库有着巨大的理论和实践价值，可以“为语言研究、翻译研究、外语教学、词典编纂和跨语言信息检索等提供最好的平台”（王克非 2012：23），国内外的研究人员投入了大量的时间精力，排除困难，建设了一系列的平行语料库。

然而，目前国内已建好的大型平行语料库（如王克非主持建设的“中国英汉平行语料库”）收集的语料主要以外译汉文本为主，而较少收集汉译外的语料。国内也建成了一些汉译外语料库，但总的来说数量较少，且收集的语料主要以虚构文本为主（如黄立波（2013）主持建设的“中国现当代小说汉英平行语料库”）。目前为止，似乎还没有研究人员构建较大规模的汉译外学术文本平行语料库。

* 本文系教育部人文社会科学研究一般项目“英汉汉英学术文本平行语料库建设研究”（18YJA740010）的阶段性成果。感谢余点、林文婕、范玲丹、万林琳等同学在数据收集和校对工作中的辛勤付出。

学术研究承载着“一个民族或国家智慧与思想的结晶，标志着其文明程度的高下”（陈才俊 2006：130）。21世纪以来，为了让中国学术走向世界，增强中国的国际话语权，我国积极推动学术外译活动，有一大批优秀的汉语学术著作被译成英文。除我国主动推介的汉语学术著作外译外，过去一个多世纪里，许多外国学者也翻译了一大批汉语学术著作。例如，Derk Bodde在1937年翻译出版了冯友兰的《中国哲学史》上册，1953年又翻译出版了下册。然而，目前为止，对汉语学术著作的外译研究总体来说还比较缺乏。为研究汉语学术文献外译的特点，服务学术文献外译，促进中国文化的对外传播，建设汉语学术文献外译平行语料库十分必要。

学术文献平行语料库有着单语学术语料库无法替代的应用和学术价值。目前国内外建设的学术语料库主要有非公开的学术语料库（如Hyland、Biber等主持建设的语料库）和公开的学术语料库（如密歇根学术英语口语语料库（MICASE））。另外，如英国国家语料库（BNC）和COCA等大型通用语料库也收录有学术语体语料。在这些语料库当中，特别值得一提的是北京外国语大学语料库语言学团队主持创建的DEAP学术英语语料库项目。该语料库目前收录了27个学科1亿多词的学术英语语料，是当前公开的覆盖学科领域最广、语料规模最大的学术英语语料库。与单语学术语料库相比，学术文献平行语料库不仅需要收录源语语料（即单语的学术语料）和译语语料，还要对两者进行句子对齐，因此有其独特的应用和学术价值。其应用价值主要体现在：（1）语料库可以转换成计算机辅助翻译软件的记忆库，并生成术语库，帮助提高学术翻译的效率与质量；（2）译员、教师和学生可以利用语料库快速检索术语、短语等的翻译，充分观察相关表达在两种语言当中的对应情况，提高他们的双语敏感性和翻译能力；（3）语料库还可以为机器翻译提供训练语料。其学术价值主要体现在：（1）语料库可以帮助研究人员全面调查词汇、短语、构式等在源语和译入语当中的对应情况，分析两种语言的异同，不同学科、不同译者（如职业译员与特定学科领域的专家）翻译的异同及其成因等；（2）平行语料库当中的译入语还可以单独组库，与单语的学术语料库组成可比语料库，用于分析学术翻译的语言特征等。

2 语料收集

为了比较全面地反映汉语学术文献英译的面貌，我们通过查询美国国会图书馆馆藏图书目录等方法确定了拟收集语料的文献名单。因部分文献收集难度或电子化难度较大，最终收录语料的文献为40多部文史哲文献及其译著（具体著作名单见表1），共约1,500万字词（含标点符号），其中汉语语料为8,367,903字，英语语料为6,730,773词。收录的文献当中，原著出版于民国时期的著作有10部，包括费孝通的《乡土中国》、冯友兰的《中国哲学史》等。这些文献大多由外国学者

翻译出版。原著出版于改革开放后到20世纪末的著作有11部，包括李泽厚的《美的历程》、牟宗三的《中国哲学十九讲》等。这些文献大多数仍由外国学者翻译出版，但是也有一小部分文献是由我国的基金资助翻译出版的，如洪子诚的《中国当代文学史》。原著出版于2000年后的有12部专著，包括罗志田的《裂变中的传承》、陈平原的《触摸历史与进入五四》等，以及9部论文集，包括《中国经济转型30年》等。这些文献当中除改革开放30年系列丛书由福特基金会资助出版外，其他图书主要由我国的基金资助翻译出版。由于目前汉语学术文献英译的数量仍然比较有限，因此本语料库采用全文收录的方式收集语料。从表1可以看出，随着时间的推移，汉语语料形符数与英语语料形符数的比例不断提高。在1949年前，两者的比例大约是1.02：1。而到了20世纪末，这一比例则上升到了1.26：1。进入21世纪以来，这一比例又进一步上升到了1.34：1。

表1 收录著作名单^{*}

时期	形符数（汉/英）	作者	作品
1949年 之前	1,576,248 /1,550,515	费孝通	乡土中国
		冯友兰	中国哲学史（上下两册）、新原道
		李剑农	中国近百年政治史
		梁启超	先秦政治思想史、清代学术概论
		鲁迅	中国小说史略
		毛泽东	寻乌调查
		萧公权	中国政治思想史
1980— 1999年	2,084,382 /1,660,605	冯友兰	三松堂自序
		郭齐家	中国教育思想史
		洪子诚	中国当代文学史
		李泽厚	美的历程、中国古代思想史论、华夏美学
		罗志田	权势转移
		牟宗三	中国哲学十九讲
		乔良、王湘穗	超限战
		王小强、白南风	富饶的贫困
		袁行霈	中国文学概论

（待续）

(续表)

时期	形符数(汉/英)	作者	作品
2000年 至今	4,707,273 /3,519,653	陈来	传统与现代——人文主义的视界
		陈平原	触摸历史与进入五四
		葛本仪	现代汉语词汇学
		葛兆光	宅兹中国
		顾明远	中国教育的文化基础
		陆学艺(主编)	当代中国社会流动
		罗志田	裂变中的传承
		骆玉明	简明中国文学史
		荣新江	敦煌学十八讲
		孙宏开、刘光坤	阿依语研究
		许钧	文学翻译的理论与实践
		朱维铮	重读近代史
		王逸舟(主编)	中国对外关系转型30年
		俞可平(主编)	中国治理变迁30年
		李强(主编)	中国社会变迁30年
		蔡昉(主编)	中国经济转型30年
		蔡定剑、王晨光 (主编)	中国走向法治30年
		郑易生(主编)	中国西部减贫与可持续发展
		卓新平(主编)	当代中国宗教研究精选丛书 基督教卷
		楼宇烈(主编)	当代中国宗教研究精选丛书 佛教卷
		金宜久(主编)	当代中国宗教研究精选丛书 伊斯兰教卷

*因新中国成立前30年出版的汉语学术文献被国外英译出版的数量较低,且该类译著获取难度较大,因此本语料库暂不收录这一时期的文献。

3 语料加工

(1) 语料电子化。在获取到文献后,我们通过扫描、文字识别、人工校对的方式将文献转化为电子文本。为提高人工校对的效率,汉语文献的电子文本先通过百度¹和腾讯²两个云平台的API接口进行文字识别,然后再通过开源工具diff-match-patch³生成两个平台文字识别后文本之间的差异。人工校对只对两个平台识别结果存在差异的文字进行检查。英文的文字识别主要使用开源软件tesseract-ocr⁴完成。

(2) 文本清洁。本语料库主要收录原著和译著的正文部分内容, 因此我们删除了封面、版权页、目录、索引、参考文献等页面内容, 以及原著和译著正文部分中的图片、表格、注释、页眉、页脚等信息。另外, 对于英文当中由于排版需要产生的连字符, 我们通过编写程序检查去掉连字符后的字符串是否为常见的英文单词来判断是否需要删除连字符。

(3) 句子对齐。与单语语料库建设相比, 句子对齐是双语平行语料库建设的一个难点。随着人工智能的发展, 目前计算机具备了跨语言计算句子语义相似度的能力, 也使得句子对齐的准确率有了较大的提高。我们采用开源工具 `vecalign`⁵ (Thompson & Koehn 2019) 对语料进行句子对齐。该开源软件推荐采用脸书的 LASER 模型⁶ 计算句子语义相似度。但经对比发现, 采用谷歌的 LaBSE 模型⁷ 对齐的准确率更高, 大多数情况下语料对齐的准确率在 90% 以上。因此, 我们改用 LaBSE 模型进行语料对齐。另外, 因为 `vecalign` 没有提供分句功能, 所以我们编写了汉语和英语的分句代码。

(4) 语料赋码。语料标注可以为研究人员提供丰富的语料信息, 方便研究人员实施精准检索。目前, 英语语料大多进行单词原形和词性标注。本语料库的英语语料标注工具为斯坦福大学自然语言处理小组发布的基于神经网络架构的开源自然语言处理工具 `Stanza`⁸ (Qi *et al.* 2020)。`Stanza` 可标注的语言多达 60 余种, 而且与该小组早期发布的工具 (如 `corenlp`、`pythonnlp`) 相比, 对词性等的标注准确度有进一步提高。因为目前开源的汉语分词和词性标注工具对本语料库所收集语料的分词和词性标注效果仍然不尽人意, 因此本语料库暂不对汉语进行分词和词性标注。

此外, 我们还对英文语料进行了句法标注。刘鼎甲、王克非 (2018: 280) 指出“经过句法标注的语料库……具有极高的价值和广阔的前景”。目前句法标注的主流方案有两种: 短语结构语法和依存语法。冯志伟 (2017: 295) 指出与短语结构语法相比, 依存句法“在语料库文本的自动标注中, 使用起来比短语结构语法方便”。因此, 我们决定采用依存句法对语料库进行句法标注。与词性和单词原形标注一样, 我们采用上文提到的 `Stanza` 对语料的依存关系进行标注。

为了便于更为精确地检索语料, 我们还对语料的元信息进行了赋码, 赋码信息包括原著作者、原著书名、原著出版社、原著出版时间、译者、译者母语背景, 译著书名、译著出版社和译著出版时间。其中译者母语背景按母语译者、二语 (外语) 译者、母语和二语 (外语) 译者合译、不详四类进行分类。

正如冯志伟 (2010: 421) 所指出的, 平行语料库的构建是很困难的工作。汉英学术文献平行语料库的建设目前仍有不少难点。(1) 文字识别仍存在少量的错误。汉语学术著作较少有电子版格式。因此收录该类语料时, 需要扫描、文字识别。虽然随着深度学习等人工智能技术的发展, 文字识别的准确率有了大幅提高,

但仍存在一些错误，如“曰”有时会被识别成“日”。（2）文本清洁工作量较大。学术文献中有大量的图表、公式、脚注等，不同文献的格式存在差异，自动识别难度较大，需进行人工校对。3）句子自动对齐后仍需进行人工校对。学术翻译有少量省译、添加注释或者把正文内容译为脚注的情况，导致自动对齐后仍有一定的错误率，需进行人工校对。

4 语料检索平台建设

4.1 检索功能介绍

许家金、贾云龙（2013）指出良好的语料检索工具的支持是语料库研究有效开展的前提条件之一。为方便检索语料，我们为语料库开发了配套的检索平台，检索主界面如图1所示。因为汉语语料没有分词和词性标注，所以能提供的检索功能相对有限。除普通词语检索外，汉语语料仅支持使用通配符“*”表示一个汉字进行检索。英语语料的检索功能相对比较全面，目前可以支持以下内容的检索。（1）单词或短语。普通检索的检索式最多可包含九个单词，每个单词用空格分开，不区分大小写。（2）通配符“_”和“*”。“_”表示一个字母，“*”表示零至多个字母，通配符可单独使用，也可与英文字母（组合）配合使用，检索含有特定字母（组合）的单词。例如，要检索以“tion”结尾的单词，可输入“*tion”。（3）词性。检索词性的表达式为单词加“.”再加词性。如果要检索某一词性的任意单词，可在“*”通配符后加“.”再加要检索的词性。词性检索采用前方一致的检索方式。例如在词性标注集中，动词词性都以“V”开头，因此检索动词时，只要在检索词后加“.v”即可。如果要检索特定时态的动词，可输入完整的词性形式。如检索动词的进行时形式，可在检索词后加“.vbg”。（4）单词原形。如果要检索某一单词原形的所有单词，可在单词原形前加“-”进行检索。（5）正则表达式。使用通配符后检索式支持使用正则表达式。



图1 检索主界面

检索主界面提供“检索”和“索引”（Concordance）两项功能。“检索”功

能主要对检索表达式检索出的单词、单词组合（检索结果）等在语料库中的频数进行统计。利用该功能还可以实现常用语料库检索工具中的词频表和词簇表提取功能。要输出词频表，只需在检索框中输入“*”，再点击检索即可。要提取词簇表，只需按照词簇中的单词个数输入“*”并用空格隔开即可。“索引”功能主要输出检索结果所在的句子及其译文。例如，在检索框当中输入“仁爱”，并点击“Concordance”，可以得到如图2所示的检索结果。从图2可以看出，在不同的语境中对“仁爱”的翻译不尽相同，有的意译为benevolent、benevolence and love，也有的直接音译为jen and ai。

检索结果 (仅显示前1000条数据)		
1	The benevolent spirit which is characteristic of Du Fu is a clear manifestation of the influence of Confucian thought.	仁爱精神是杜甫的特点，这明显地体现了儒家思想的影响。中国文学概论
2	Mencius, IV/i/1/3; following Legge's translation, Mencius, p. 289.] Therefore he discovered, in the lenient and simple government of the Yin, a principle of benevolence and love, which he combined with the rites and institutes of the Chou, making thereby a system that possessed both ethical foundations and the capacity for practical application. With this, his advocacy of following the Chou now took on profound and farreaching significance; therein, also, was established the ultimate goal, the final stage, toward which Confucius' whole corpus of political thought pointed.	故于殷政宽简之中，发明一仁爱之原则，乃以合之周礼，而成一体用兼具之系统，于是从周之主张始得一深远之意义，而孔子全部政治思想之最后归宿与目的，亦于是成立。中国政治思想史
3	Confucius looked upon providing for the people's nourishment as an essential duty; that is one manifestation of his concept of benevolence and love [jen ai]. Thus he looked upon "the extensive dispensation of succor to the masses" as the achievement of the Sage. [See page 169, and footnote 53.]	孔子以养民为要务，盖亦仁爱思想之一种表现。故博施济众，孔子认为圣人之业。中国政治思想史
4	However, Confucius took benevolence and love as the basis of all government.	然孔子以仁爱为政本。中国政治思想史
5	Briefly, there are four kinds of evidence for this: (1) The two terms, jen and ai, have the same semantic denotation.	约言之，其证有四：(1)仁爱二名之训诂相通。中国政治思想史
6	Since jen and ai have the same definition, it is scarcely possible that they should have conflicting philosophical implications.	仁爱同训，则其义岂相违？中国政治思想史
7	It is greatly to be regretted that this passage does not make clear what it was that Mo Tzu had cited; we have no way of determining whether or not it might have been relevant to benevolence [jen] and love [ai].	尤足见墨之有得于孔，惜乎此文不著所称之言，不能考其是否涉及仁爱耳。中国政治思想史

图2 “仁爱”的索引结果

除对词语、搭配等进行检索外，建设的检索平台还支持依存关系检索，也就是支持对句子中任意一个节点的依存词和管辖词以及它们的依存关系（如主谓、动宾、形名等）进行检索，检索界面如图3所示。依存词和管辖词的检索式与普通检索一样支持通配符、词性、单词原形和正则表达式，但仅允许输入一个单词。Stanza标注的依存关系共有49种，检索时仅需选择想要检索的依存关系即可。例如，要检索动宾关系，仅需将依存关系选为“obj”，点击检索即可出现如图4所示的检索结果界面。从图4可以看出英语译文当中使用频率最高的动宾关系是“pay...attention”，点击该动宾关系后的频率即可出现译文当中含有“pay...attention”的索引结果（如图5所示）。

依存关系检索

依存词 *

管辖词 *

依存关系 obj

检索

Concordance

图3 依存关系检索界面

检索结果（仅显示前1000条数据）

序号	依存词	管辖词	依存关系	频率
1	attention	pay	obj	304
2	us	let	obj	295
3	role	played	obj	291
4	attention	paid	obj	226

图4 动宾关系检索结果

检索结果（仅显示前1000条数据）

序号	dep	cn	en	title
1	obj	他认为,“假使科学救国的论可以成立,我们在中国所以治科学的精神应当如何、如何可以使科学在中国得一根固枝荣的生命”,这才是应该关注的问题。	Lin believed, " To establish the theory of scientific national salvation , we in China must pay close attention to the nature of our spirit of studying science and understand how to make science the root of a prosperous life . "	裂变中的传承 20世纪前期的中国文化与学术
2	obj	(2)重视硬件建设忽视软件建设的倾向 西部地区普遍存在重投资轻技术、重项目轻管理、重设备轻人才的倾向。	5.5.2 . Constructing " Hardware " but Ignoring " Software " A widespread phenomenon in the West is the tendency to pay more attention to investment but ignore technology , pay attention to projects but ignore their management , and pay attention to equipment but ignore trained personnel .	中国西部减贫与可持续发展
3	obj	由此可见,对经济安全问题的理论关注是中国开始对“非传统安全问题”进行理论关注的重要方面。	27 Thus , theoretical concerns for the economic security problem are important aspects indicating that China has begun to pay attention to nontraditional security problems from a theoretical perspective .	中国对外关系转型30年

图5 “pay...attention” 动宾关系的索引结果

4.2 检索功能实现机制

要实现对语料的高效检索，需要对语料进行索引并建立相应的查询语言。目前，大型在线语料库采用的索引及查询系统主要有三种：CWB、Lucene 和 SQL。CWB 是专门为语料库开发的开源索引和查询系统，基于 CWB 开发的在线语料检索平台有 CQPweb（Hardie 2012）。Lucene 是目前使用最为广泛的开源全文检索引擎，基于 Lucence 开发的在线语料检索平台有北大的 CCL（詹卫东等 2019）。SQL 指结构化查询语言，是关系型数据最常用的查询语言，基于 SQL 开发的在线语料检索平台有 COCA（Davies 2009）。由于 SQL 系统检索功能强大、响应速度快、扩展性强、部署方便，因此本语料库选择 SQL 建立语料检索平台。我们将语料库的语料和标注信息存储于 SQL 数据表中，并建立索引，然后通过程序代码将用户输入的检索表达式解析为 SQL 查询语言，再将 SQL 的查询结果转换成合适的格式后展现给用户。

4.3 研究案例

学术语篇的主要作用是转述、交流和讨论他人或自己的研究结果或观点。Thompson & Ye（1991）指出作者转述观点时通常需要对观点进行评价，他们还对具有评价功能的转述动词进行了详细分类。Hyland（1999）发现不同学科对转述动词的使用存在较大差异，原因是不同学科构建知识的方式存在差异。那么，学术翻译中转述动词的翻译是否也会因学科领域的不同而不同？本小节以转述动词

“认为”为例进行调查。

调查的数据来源为语料库当中收录的《当代中国宗教研究精选丛书》中《佛教卷》和《伊斯兰教卷》的原著及英译本。《佛教卷》由台湾辅仁大学的林佩莹博士翻译；《伊斯兰教卷》由香港公开大学的兼职教师 Alex Chan 博士翻译。汉语中动词“认为”与观点之间通常由逗号隔开，因此我们采用检索词“认为”进行检索，发现该检索词在《伊斯兰教卷》和《佛教卷》中的频次分别为54次和51次。检索出例句后，经人工逐句确认“认为”在译文中的对应翻译。具体结果如表2所示。

表2 “认为”在《佛教卷》和《伊斯兰教卷》中的英译情况

佛教卷		伊斯兰教卷	
译入语	频数	译入语	频数
believe	13	省译	15
argue	9	affirm	4
省译	8	argue, assert, think	3
contend, hold	3	indicate, propose, view	2
agree, consider, suggest	2	agree, attribute, concede, confirm,	1
accept, according to, as ... understood, claim, in, in ... opinion, regard, state, think	1	contend, define, elaborate, in the view of, judge, judgment, point, point out, point to, put, regard, remark, say, suggest, make ... more explicit by stating, state	

对比“认为”在两个译本当中的翻译，可以发现《佛教卷》的翻译较多采用believe、hold等立场强度、语气都较弱的词汇，传达出了转述观点或论述可能存在推测成分，为作者和读者就观点的正确性提供了对话、商榷的空间；而《伊斯兰教卷》的翻译则较多采用affirm、assert等立场强度、语气都较强的词汇以及省译的方法，传递出转述的观点或论述是客观事实和正确、权威的信息，也构建了原作者较为权威的身份。造成这种差异可能的原因是《佛教卷》的翻译受中国佛教最主要的宗派——禅宗强调“悟”的影响，倾向于使用believe、hold、suggest等词汇为读者提供思考的空间。

每个学科、每种语言构建知识和作者身份的方式不尽相同。因此，学术翻译要求译者除准确传递信息外，还要考虑语言、文化的差异以及译者所属学科知识、作者身份构建的方式，以尽可能准确地再现原作的立场。

5 汉英学术文本平行语料库的应用

王克非（2012）指出平行语料库的建设可以为外语和翻译的教学与研究、词典编纂、机器翻译等提供语料支持。除上述价值外，本平行语料库还有一些独特的实践和理论价值。

本语料库的实践价值在于可以为传播中国文化和学术外译服务。本语料库收集了较多的中国哲学、历史学、社会学、文学、美学等领域的著作及其译著。这些著作当中有大量承载中国特色文化且英语中没有对等词的词汇，如“仁”“君子”等。翻译这些词汇时，不同译者有不同的译法，甚至同一译者在不同时期、不同语境中的译法也不同。例如，Derk Bodde在翻译《中国哲学史》上册时主要将“仁”翻译作human-heartedness并加注jen，而在翻译下册时则选择将“仁”译作love。本语料库的建立可以帮助译者查询这些文化负载词的翻译情况，并在充分了解翻译现状基础上作出最佳的翻译选择。此外，本平行语料库收录的主要是学术文本，可以帮助提高学术外译的质量。学术文本由于专业性强，与其他语域（如文学文本）的文本在用词和句式上差异较大，只有学术文本平行语料库才能更好地满足学术翻译的需求。

本语料库的理论价值在于可以为学术翻译语言特征的系统描写提供数据支持。学术语体与其他语体有着不同的语言特征，目前已有大量的研究对学术语体的特征进行描写（如Biber 2016）。学术翻译语体与学术语体在语言的使用上也可能存在较大差异。对比本语料库进入21世纪以来的语料与焱炎通用英汉平行语料库（徐秀玲、许家金 2021）的学术题材语料，可以发现本语料库的汉英形符数比（1.34 : 1）比焱炎的形符数比（1.62 : 1）要低得多，这表明汉语学术原著与汉语学术译著的语言特征可能存在较大差异。然而，目前为止，对学术翻译语言特征的调查还比较缺乏。陶源（2018）似乎是目前为止仅有的基于语料库的学术翻译研究专著。该专著主要对俄汉学术翻译进行调查。然而，中国的学术翻译主要是在英汉两种语言之间开展，到目前为止，还没有研究对英汉学术翻译进行系统调查。我们并不清楚，英语学术翻译语体与英语学术语体之间具体有何异同？也不清楚学术翻译语体与其他翻译语体（如文学翻译语体）之间有何异同？本语料库的建设可以为回答上述问题提供数据支持。

6 结语

平行语料库建设具有重要的理论和实践价值，可以为翻译研究、翻译实践、语言对比、外语教学等提供数据支持。目前国内已经建设了不少平行语料库。但总的来说，汉译外的学术语体平行语料库建设还未得到足够的重视。为弥补上述不足，我们收集了40多部跨越过去一个多世纪的汉语学术著作及其译著，通过文

字识别、校对、文本清洁、句子对齐、语料赋码，建成了库容量约为1,500万单词的汉英学术文本平行语料库。为方便检索语料，我们还为该语料库开发了配套的检索平台。该平台除可以实现常见的检索功能（如搭配检索、使用通配符检索）外，还可实现依存语法检索。该语料库的建设可为学术翻译研究和学术翻译实践等提供语料支持。

大型语料库通常采用截取部分片段或部分章节的方式收集著作类语料，以尽可能多收录不同著作的语料，提高语料库的代表性。本语料库没有采用截取部分片段或部分章节的方式收集语料，主要是因为2000年以前汉语学术文献英译出版的数量非常有限。21世纪以来，随着中华学术外译等外译项目的开展，已经有越来越多优秀的汉语学术著作被外译出版。将来扩充21世纪以来的语料时，我们将采用截取片段或章节的方式采集著作语料。

注释

- 1 <https://aip.baidubce.com/rest/2.0/ocr/v1/accurate>。
- 2 <https://ocr.tencentcloudapi.com>。
- 3 <https://github.com/google/diff-match-patch>。
- 4 <https://github.com/tesseract-ocr/tesseract>。
- 5 <https://github.com/thompsonb/vecalign>。
- 6 <https://github.com/facebookresearch/LASER>。
- 7 <https://ai.googleblog.com/2020/08/language-agnostic-bert-sentence.html>。
- 8 <https://github.com/stanfordnlp/stanza>。

参考文献

- BIBER D. Grammatical complexity in academic English: linguistic change in writing [M]. Cambridge: Cambridge University Press, 2016.
- DAVIES M. The 385+ million word Corpus of Contemporary American English (1990-2008+): design, architecture, and linguistic insights [J]. *International Journal of Corpus Linguistics*, 2009, 14(2): 159-190.
- HARDIE A. CQPweb — Combining power, flexibility and usability in a corpus analysis tool [J]. *International Journal of Corpus Linguistics*, 2012, 17(3): 380-409.
- HYLAND K. Academic attribution: citation and the construction of disciplinary knowledge [J]. *Applied Linguistics*, 1999, 20(3): 341-367.
- QI P, ZHANG Y, ZHANG Y, BOLTON J, MANNING C. Stanza: a python natural language processing toolkit for many human languages [R]. Presented at the

- 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations, 2020.
- THOMPSON B, KOEHN P. Vecalign: improved sentence alignment in linear time and space [R]. Presented at the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), Hong Kong, China, 2019.
- THOMPSON G, YE Y. Evaluation in the reporting verbs used in academic papers [J]. *Applied Linguistics*, 1991, 12(4): 365-382.
- 陈才俊. 学术著作翻译原则刍议 [J]. *学术研究*, 2006 (9): 130-134.
- 冯志伟. 双语语料库的建设与用途 [J]. *现代外语*, 2010 (4): 420-421.
- 冯志伟. 自然语言计算机形式分析的理论与方法 [M]. 合肥: 中国科学技术大学出版社, 2017.
- 黄立波. 中国现当代小说汉英平行语料库: 研制与应用 [J]. *外语教学*, 2013 (6): 104-109.
- 刘鼎甲, 王克非. 面向语料库深加工的英汉自动依存句法标注方法 [J]. *现代外语*, 2018 (2): 279-289.
- 陶源. 基于俄汉平行语料库的人文社科类学术文本翻译研究 [M]. 北京: 科学出版社, 2018.
- 王克非. 中国英汉平行语料库的设计与研制 [J]. *中国外语*, 2012 (6): 23-27.
- 肖忠华. 英汉翻译中的汉语译文语料库研究 [M]. 上海: 上海交通大学出版社, 2012.
- 徐秀玲, 许家金. 熒炎英汉平行语料库的创建 [J]. *语料库语言学*, 2021 (1): 149-151.
- 许家金, 贾云龙. 基于 R-gram 的语料库分析软件 PowerConc 的设计与开发 [J]. *外语电化教学*, 2013 (1): 57-62.
- 荀恩东, 饶高琦, 肖晓悦, 臧娇娇. 大数据背景下 BCC 语料库的研制 [J]. *语料库语言学*, 2016 (1): 93-109.
- 詹卫东, 郭锐, 常宝宝, 谌贻荣, 陈龙. 北京大学 CCL 语料库的研制 [J]. *语料库语言学*, 2019 (1): 71-86.

通信地址: 350007 福建省福州市 福建师范大学外国语学院

LawDEAP 法学学术英语语料库的创建*

上海电机学院 王艳伟 干 诚 李俊飞 高利杰 黄张成

提要：LawDEAP 法学学术英语语料库是学术英语语料库DEAP的一个子库。本文介绍LawDEAP的建设思路和实施步骤，对建库目标、语料采集、文件命名与元信息登记、格式转换及文本清理作了具体说明。文章最后探讨了LawDEAP的后期开发及其今后在法学词表研制、词典编纂等法学学术英语教学与研究领域的应用前景。

关键词：学术英语、法学、专用语料库、LawDEAP

1 引言

语料库是一个按照一定采样标准采集而来的、能够代表一种语言或者某语言的一种变体或文类的电子文本集（梁茂成等 2010）。语料库可以按照不同的标准或研究目的进行更精细的分类。按照应用领域的不同，语料库可以分为通用语料库和专用语料库。通用语料库是出于通用的研究目的，广泛采集某语言的口、笔语形式，取样时尽可能考虑口、笔语的主要社会变体、地域变体、行业变体等各种变异及语言使用的各种场合之间的平衡，力求最好地代表一种语言的全貌。而专用语料库通常是出于特定的研究目的，专门收集某个特定领域或主题的语料样本。根据不同的使用目的，专用语料库还可以进一步细分，学术英语语料库即其中一种。1982年，由杨惠中教授领衔启动建设的我国第一代语料库——上海交通大学科技英语语料库（JDEST），对大学英语学术词表的研发居功甚伟。就法律领域而言，目前我国已建有大规模的法律文献数据库，如主要由国家信息中心提供的中国法律法规数据库和北大法宝，也有绍兴文理学院建设的中国法律法规汉英平行语料库（PCCLD）（孙鸿仁、杨坚定 2010）等。这些数据库和语料库虽然功

* 王艳伟为本文通讯作者。

作者贡献：

王艳伟：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（100%）、修改润色；

干诚：数据收集；

李俊飞：数据收集；

高利杰：数据收集；

黄张成：数据收集。

能和定位不同（前者为法律知识和内容提供服务，是法律应用的辅助资源；后者为法律语言研究和教学服务，是学术研究资源），但语料类型都集中在法律法规等立法语言，较大规模的专门采集法学研究论文、综述论文等学术语料的法学学术英语语料库甚为鲜见。

笔者以“法律”“法学”“语料库”为主题词，在中国知网可检索到相关论文16篇，其中9篇讨论法律语言的研究或翻译，7篇探讨法律语料库建设相关问题（如宋北平 2008；王东海、王洁 2008；吴棐弘 2014等）。但若以“学术英语、法学语料库”为主题词，则检索不到相关研究。诚然，已有学者开展了一些法学学术语言相关研究工作，但大都不是基于语料库或基于较大规模的法学专用语料库。如苗青（2014）以法律英语教学为例，多角度考察和探讨了开展学术英语教学可能遭遇的瓶颈及其发展策略。张清、郝瑞丽（2016）介绍了中国政法大学为教改后的学术英语课程编写的《法学英语》教材进行的实践与反思。孙波（2017）从语篇信息理论视角，运用语料库的研究方法，考察了英语本族语者的法学学术语篇中认识情态的使用情况，着重分析了认识情态在语篇信息中的量值和取向分布，并探讨了认识情态对信息功能的调节作用。张磊、卫乃兴（2018）基于Beijing CARE学术文本可比语料库中的中国和西方学者法律子库，对比分析中、西法律学者在使用三个名词评价型式there v-link N of、there v-link N in和n v-link of N时所呈现的高频局部语法型式。郝瑞丽、刘艳（2020）从法律英语切入，尝试从学术英语教学的任务、词汇教学的任务、学术英语词汇教学的任务以及教学词汇的特点等多个角度确定学术英语教学词汇的标准。可见，在现有法学学术英语研究中，法学语料只是相关研究所用语料库中的一个子库，且样本量普遍较小。故设计规模在500万词的LawDEAP法学学术英语语料库的建设对于中国学者了解国际法学研究的话语体系及其语言运用规范，实现与国际法学研究学术共同体的无缝对接，都具有重要的参考价值。

本文报告的LawDEAP法学学术英语语料库，隶属于北京外国语大学中国外语与教育研究中心的语料库共建项目——DEAP学术英语语料库（Database of English for Academic Purposes）。故LawDEAP根据“中国外语教育基金专用英语语料库建设项目”的总体框架进行设计和建设，其建库标准和建库流程与其他领域基本一致，可用于DEAP不同专业领域语料库间的对比研究。下文主要介绍LawDEAP的建设思路、实施步骤及其深度加工和应用前景。

2 建库目标

LawDEAP的建设目标包括以下两个方面。其一，结合国内法学学科的特点，建设一个尽可能覆盖法学11个二级学科、规模较大、时效性强、能体现法学学术英语使用全貌的语料库，并探讨专用语料库规范的建库原则及其语料收集与清理方

式。其二，为我国法学学术英语教学和科研赋能，为法学学术词表的编纂、法学学术语言研究、教材编写等提供较大规模的真实语料。LawDEAP的设计总库容为500万词，建成后的LawDEAP共计5,014,908词次，由697个文本组成。该语料库由研究论文与综述论文这两种主要的学术语类构成，其下设10个子库¹（见表1）。

表1 法学（0301）学术英语语料库结构

研究论文	综述论文
法学理论（030101）	诉讼法学（030106）
法律史（030102）	经济法学（030107）
宪法学与行政法学（030103）	环境与资源保护法学（030108）
刑法学（030104）	国际法学（030109）
民商法学（030105）	知识产权法（030111）

3 语料采集

LawDEAP的语料采集遵循以下4个标准。（1）语料的专门性。因为是法学学术英语专门语料库，LawDEAP的语料取样方案依据我国的学位授予和专业人才培养目录，特别是国务院学位委员会、教育部2011年印发的学科专业目录，其中法学（学科代码0301）一级学科下设法学理论、法律史、宪法学与行政法学、刑法学、民商法学、诉讼法学、经济法学、环境与资源保护法学、国际法学、军事法学、知识产权法共11个二级学科，我们主要根据学科的研究范围和期刊的发文范围确定与这些法学二级学科基本对应的国际期刊。（2）语料的权威性。LawDEAP以“剑桥学术期刊”和ScienceDirect数据库law分类下的期刊文章为数据源，并参考法学专家的意见，最终每个二级学科选取2—6种期刊，共计44种国际期刊（见表2）。（3）语料的代表性。主要收集法学研究中最具代表性的两类文章：研究论文和综述论文（包括题目、作者信息、摘要、关键词和正文）。基本上所有的法学期刊都刊登研究论文，以报道法学研究领域的最新成果。综述论文大多概述法学研究的历史、进展与动态，此类文章只出现在少量的法学期刊上，故在LawDEAP中，研究论文在数量上占绝对优势（87.66%，611/697）。（4）语料的时效性。LawDEAP对上述两类论文全文收录，为确保时间跨度足够小，特别注意收集近年来发表的文章，最终将论文发表的时间控制在2017—2019年，以更好地反映法学学术语言的共时性特征。

具体工作流程如下：

（1）在收集LawDEAP语料时，我们从“剑桥学术期刊”和ScienceDirect数据

库 law 分类下的期刊（共计 44 种）中批量全文下载 2017—2019 年的研究论文和综述论文。

（2）根据国内法学 10 个二级学科的研究范围及上述两个学术期刊数据库中的期刊介绍将二者进行匹配，然后请两位法学博士教师审阅二级学科与国际期刊的匹配结果并提出调整意见，以确定最终的匹配结果（见表 2）。

（3）结合已建成 DEAP 子库负责人分享的经验，按照文章封顶数 800 篇，根据随机抽样原则，从批量下载的 44 种期刊文章中按每刊 20 篇进行等距抽样，排除个别篇幅较短的综述论文，最终共获取 697 篇论文语料，分类存放到各二级学科子文件夹中。

表 2 法学学术英语语料库来源期刊

序号	二级学科	刊物名称
1	法学理论 (Legal Theory)	Asian Journal of Law and Society
2		Cambridge Quarterly of Healthcare Ethics
3		Canadian Journal of Law & Jurisprudence
4		Law & Social Inquiry
5		Legal Studies
6		Legal Theory
7	法律史 (Legal History)	Asian Journal of Comparative Law
8		Law and History Review
9		The Cambridge Law Journal
10	宪法学与行政法学 (Constitutional Law & Administrative Law)	European Constitutional Law Review
11		Global Constitutionalism
12		Government Information Quarterly
13	刑法学 (Criminal Law)	Egyptian Journal of Forensic Sciences
14		Forensic Science International
15		Forensic Science International: Genetics
16		International Annals of Criminology
17		International Journal of Law, Crime and Justice
18		Journal of Criminal Justice

（ 待续 ）

(续表)

序号	二级学科	刊物名称
19	民商法学 (Civil Law & Commercial Law)	Accident Analysis & Prevention
20		Aggression and Violent Behavior
21		Business and Human Rights Journal
22		Children and Youth Services Review
23		Journal of Safety Research
24		Safety Science
25	诉讼法学 (Procedural Law)	Journal of Forensic and Legal Medicine
26		Science & Justice
27	经济法学 (Economic Law)	Health Economics, Policy and Law
28		International Review of Law and Economics
29		Journal of Air Transport Management
30	经济法学 (Economic Law)	Telecommunications Policy
31		Transport Policy
32		World Trade Review
33	环境与资源保护法学 (Environment & Resources Protection Law)	European Journal of Risk Regulation
34		Resources Policy
35		The Electricity Journal
36		Transnational Environmental Law
37	国际法学 (International Law)	Asian Journal of International Law
38		Cambridge Yearbook of European Legal Studies
39		International & Comparative Law Quarterly
40		International Legal Materials
41		Leiden Journal of International Law
42	知识产权法 (Intellectual Property Law)	International Journal of Cultural Property
43		Legal Information Management
44		World Patent Information

4 文本命名与元信息登记

与其他DEAP子库一致，我们对抽样获取的期刊文章按照“一级学科+二级

学科+文献序号+文章类型”的形式进行命名。其中，一级学科law用大写英文字母L表示；二级学科采用两个大写英文字母缩写表示（见表3）。文献序号用三位阿拉伯数字表示。两种文章类型中，研究论文缩写为RA（research articles），综述论文缩写为RV（review articles）。例如，LawDEAP中经济法学二级学科的第85篇文章为研究论文，文件名即为LEL085RA。

表3 法学二级学科名称代码

二级学科	法学理论	法律史	宪法学与行政法学	刑法学	民商法学
代码	LT	LH	CA	CL	CC
二级学科	诉讼法学	经济法学	环境与资源保护法学	国际法学	知识产权法
代码	PL	EL	ER	IL	IP

文本元信息是关于文本信息的信息，用于描述文本信息的命名方式、存储和出处等信息。LawDEAP中的文本元信息包括文件名、子库名称、期刊名称、论文题目、文类、发表年份和DOI号（数字对象唯一标识符）。以上信息经人工汇总登记，最终保存为Excel版元信息文件。

5 格式转换与文本清理

通过网络下载的文本中并非所有语言信息都适合学术英语语料库的建设用途，如需要剔除对学术研究价值不大的文中图表及其说明、公式、参考文献等信息。我们收集的法学期刊论文全部为PDF格式文件，需要转换为方便语料库检索工具读取和提取数据的格式（如DEAP全库统一采用UTF-8文件格式）。为了减少过多的冗余信息对格式转换质量的影响，我们使用Adobe Acrobat Pro软件剪除了每篇文章PDF页面上的页眉和页脚，然后利用文本格式转换工具ABBYY FineReader 14将裁剪后的PDF文件转换为方便校对和清理的Microsoft Word格式，对于部分转换失败的文章，再通过Solid Converter PDF软件重新进行转换。

经过格式转换后的语料不同程度地存在乱码、格式错误等问题，因此必须清理文本，并备份原始文件。文本清理是语料库建设工作中最为耗时费力的环节，除了少量使用PowerGREP，大部分由包括负责人在内的五名工作人员共同完成。在开始文本清理前，负责人根据DEAP总库的建设要求，制定了可操作性较强的文本清理内容和校对要求，随后与所有参与人员进行讨论，确保每人清楚文本清理的具体内容与要求。分工完毕后建立建库微信群，以便项目成员实时保持联系，对文本清理中遇到的各种问题进行讨论，尽最大努力提高文本清理工作的标准化

程度和一致性,以确保后续研究的信度与效度。

具体来说,我们对转换后的Word文件进行了三轮人工清理和校对。首先,打开转换后的Word文件,对照PDF原文,从头至尾逐项开展清理工作,如删除文本中的期刊名称、期刊标识、空行、空格、图表(包括相应标题与注释)、公式以及附录、致谢、项目资助信息、参考文献等,并逐一校正转换过程中出现的拼写错误、标点错误、乱码、文章结构顺序错乱、断句、断行、断段、例子编号错乱等问题,对语料进行第一轮校对。然后,在第一轮文本清理工作完成后,负责人对所有文本逐一进行第二轮校对,确保入库语料质量没有问题后将Word文件另存为纯文本,即Unicode(UTF-8)格式。考虑到Word文件中有一些隐含的文本框,另存为txt文件后信息会缺失,我们对另存后的txt文件又进行了第三轮校对,以确保转换后的文本信息完整且格式统一。最后,针对部分文本存在的格式转换后的脚注嵌入问题,利用PowerGREP软件进行批量清除(Regex: \[w+ (\s w+)*\]),并统计LawDEAP的语料总量(Regex: \b[a-zA-Z\-']+\b),确认无误后入库存放。

6 LawDEAP的后期开发及其应用前景

为了进一步增加语料的代表性,扩展语料库资源在学术英语研究中的应用范围, LawDEAP的后期开发可从以下两方面入手。一是扩大语料库规模。本项目搜集的语料源为“剑桥学术期刊”和ScienceDirect两个数据库,后期开发可以收集其他数据库收录的法学类期刊语料。不仅如此,其他法学研究成果的产出形式,如学术专著、会议论文等,也可以考虑纳入语料采集的范围。此外,还可以考虑增加语料收集的时间跨度,以便于对法学学术语言进行历时性研究。二是增加语料加工的深度。目前LawDEAP中的语料为生语料,研究者在使用时可以根据需要对语料进行各种深度加工,如增加词性标注、特定词类标注、句法结构标注、依存信息标注、修辞结构标注、语步功能标注、评价资源标注、隐喻标注等。

LawDEAP的创建具有广阔的语言研究和教学应用前景。LawDEAP可实现生成索引行、词表,计算搭配词、主题词等多项功能,将发布在语料云网站或CQPweb语料库检索平台,供法学科研人员、语言教师及法学英语学习者使用单词、短语或正则表达式等对不同词形、各种词语组合模式或词簇(n元组合)等进行在线检索,有效服务于法学学术英语教学与研究。

第一,在法学学术英语研究方面, LawDEAP可以为研究者了解和追踪学术热点问题,分析和驾驭法学学术英语的修辞特征等提供数据支持。研究者可以探讨法学学术语言的词汇和句法特征,如名词化、情态系统、and和or连接的并列多项式、多重后置定语结构等,也可以对比不同学科领域在词汇、句法使用方面的差异。

第二，可以基于 LawDEAP 制作法学各个二级学科的常用学术词表或术语表。这些词表可用于法律英语教学大纲、教材的词表编写、法学词典的编纂，也可用于法律英语教学及测试材料的编写等其他用途。

第三，利用法学各二级学科的文本语料，可以对各二级学科的多语言特征进行对比分析。不仅如此，作为参照语料库，基于 LawDEAP，研究者既可以对某个子库中的语言特征进行多重对比分析，以揭示该二级学科过多使用或过少使用的语言项目并探讨其成因，也可以基于 LawDEAP 与其他 DEAP 学科开展多维度比较研究。

第四，研究者还可基于体裁短语学视角（许家金 2017），利用扩展意义单位的分析方法，对词汇或短语进行整句释义，归纳其高频搭配型式，并提供常见例句，进行法学学术英语学习者词典的编纂研究。考虑到学习者的英语水平与认知需求，还可以对例句中特别晦涩难懂的词汇进行适当改写或删减，或配上图片等可视化呈现方式。

7 结语

秉持语料库语言学的理念、技术和方法，作为大型学术英语语料库 DEAP 的一个子库，LawDEAP 的创建过程中恪守 DEAP 总库的规范，在充分听取法学专家意见的基础上收集语料，在语料可及性的范围内精选权威国际期刊和代表性强的法学研究文章，经过文本采集和抽样、元信息登记、格式转换、语料清理和校对入库各个环节，项目成员做到了分工明确、标准统一，确保了语料的代表性、真实性以及文本元信息登记的准确性。该语料库涵盖二级学科面较广，兼顾规模的可行性与语料的时效性，能够从国际学术发表这一侧面体现法学学术英语的典型特征。在法学学术语言研究与教学方面，LawDEAP 可以为法学学术语言的使用特征研究、法律英语教学大纲词表的开发、课堂教学资料的设计和法学学术英语写作平台的构建提供更具针对性的语言素材。在词典编纂方面，LawDEAP 可以作为法学学术术语及其例句抽取的重要数据源。该语料库也具有开放性，研究者可根据需要对 LawDEAP 进行深度加工、扩充或更新。

注释

- 1 鉴于法学二级学科军事法学（030110）的特殊性，在国际学术期刊数据库中未收集到相关语料。

参考文献

- 郝瑞丽, 刘艳. 学术英语教学词汇的教与学——以法律英语为例[J]. 中国ESP研究, 2020 (2): 10-15.
- 梁茂成, 李文中, 许家金. 语料库应用教程[M]. 北京: 外语教学与研究出版社, 2010.
- 苗青. 高校开展专门学术英语教学之瓶颈与对策刍议——以法律英语教学为视角[J]. 外语教学理论与实践, 2014 (3): 65-70.
- 宋北平. 我国第一个“法律语言语料库”的建设及其思考[J]. 修辞学习, 2008 (1): 25-29.
- 孙波. 英语法学学术语篇中信息表达的认识情态研究[J]. 信阳农林学院学报, 2017 (3): 74-78.
- 孙鸿仁, 杨坚定. “中国法律法规汉英平行语料库(PCCLD)”创建的思路、过程与功能[J]. 绍兴文理学院学报(哲学社会科学), 2010 (2): 48-51.
- 王东海, 王洁. “一库三典”的法律语言学研究资源建设[J]. 中国政法大学学报, 2008 (4): 111-116.
- 吴茆弘. 法律翻译教学信息化研究——基于语料库和数据库检索系统的实践探索[J]. 外语电化教学, 2014 (6): 18-24.
- 许家金. 体裁短语学视角下的医学学术英语词典研编[J]. 外语与外语教学, 2017 (6): 52-60.
- 张磊, 卫乃兴. 中、西学者法学论文评价局部语法对比: 对名词型式的探索[J]. 当代外语研究, 2018 (3): 93-99.
- 张清, 郝瑞丽. 专门用途英语教材编写的实践与反思——以《法学英语》教材为例[J]. 中国ESP研究, 2016 (2): 14-24.

通信地址: 201306 上海市 上海电机学院外国语学院

ShipDEAP 船舶与海洋工程 学术英语语料库的创建*

哈尔滨工程大学 田 苗 滕如玉

提要：随着学术英语的蓬勃发展，近年来国内已有多个学科相继建设了学术英语语料库，但其中鲜有船舶与海洋工程学科相关的学术英语语料库对外公布。ShipDEAP 船舶与海洋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“DEAP 学术英语语料库”（Database of English for Academic Purposes）的子库之一。本文详细说明了ShipDEAP语料库的语料来源、文本构成、语料收集与命名、语料清洁与标注等建设过程，并探讨了该库在船舶与海洋相关学科语言研究与学术英语教学中的实际应用。

关键词：船舶与海洋工程、学术英语、专用语料库、语料库建设

1 引言

学术语篇是知识传播和学术交流的重要媒介，它不仅传递科学信息，而且表达丰富的人际意义（姜峰、Hyland 2020：24）。随着学术交流国际化趋势不断加强，提升我国年轻学者英语研究论文撰写能力、增强我国在国际学术团体中行业话语权的重要性不断显现，学术英语也逐渐成了人才培养的热点之一。在2020年最新颁布的《大学英语教学指南》中，教育部更是明确将专门用途英语纳入大学英语三大类课程之一。学术英语作为专门用途英语的一个重要分支，势必会受到学界的更多关注，其研究规模也将随之进一步扩大。

语料库为学术英语研究带来了实证维度，提供了大量真实的语言数据，研究者通过检索与统计便可以直观地得到学术英语词汇或者语法的分布信息，增强了研究者对学术英语观点论断的可信度（Hyland 2015：292；徐昉 2015：102）。根据姜峰（2022：417）对国内外近40年学术英语研究发展过程的梳理，“语料库”作为方法类的研究主题，在国内学术英语研究中增长幅度最大。Teubert（2005：1）也指出“现如今语料库几乎是所有语言学家的科研必备资源”。可见，语料库业已成为学术英语研究的重要途径。

* 本文系黑龙江省哲学社会科学规划项目“‘船舶与海洋工程’学科学术英语语料库的建设与研究”（22YYE479）的阶段成果。

田苗为本文通讯作者。

田苗：选题构思、研究方法、初稿撰写、字数占比（50%）、修改润色；

滕如玉：数据分析、讨论结论、初稿撰写、字数占比（50%）。

随着学术英语的蓬勃发展,语料库建设也在朝着精细化方向前进,一系列应用于不同领域的学科专业语料库相继出现。与其他类型的专门用途语料库相比,基于某一学科所建设的语料库针对性更强、专业化程度更高,在特定学科的语言研究与实际应用中优势更为明显。国内外已有多个学科相继建设了研究语料库,国际上如数学学科的理论数学领域期刊论文语料库(McGrath & Kuteeva 2012)、应用语言学学科的语言学基础教材语料库(Freddi 2005)、化学学科的化学期刊论文语料库(Valipouri & Nassaji 2013)等,国内如应用语言学学科的《应用语言学》期刊论文语料库(梁茂成、刘霞 2014)、医学学科的医学学术英语语料库(冯欣等 2017)、农业学科的农学学术英语语料库(吕靖、邓飞 2020)等。

然而,就船舶与海洋工程(下文简称船海)学科而言却鲜有相关语料库建设或研究。国内虽有学者探讨过海洋船舶英语论文语料库建设的可行性(郑军 2014),但该语料库后期并未建成或是公开。目前已建成且有一定规模的只有大连海事大学的海事英语语料库(范凤祥 2006),该库总库容 100 余万词次,收录了航海英语、轮机英语和海事条约英语三个领域的英语语料,为航海英语研究提供了便利,但库容较小,且无法充分代表船海学科学术英语语言的全貌,不便于船海学术英语的研究。还有学者建立了一些与船海相关的小型学术英语语料库,如:《船舶与海洋工程学报》英文摘要语料库(李芳萍、李舰君 2015)、轮机英语教学语料库(颜天明、侯慧凡 2018)、海洋工程学术论文引言语料库(张凤婷 2020)等。这些小型自建语料库往往服务于建设者的某一特定研究目的,库容大多不超过 20 万词次,这也决定了其语料选取与应用范围的局限性,无法推广至学界作共享研究使用。此外,上述提到的所有已建成船海语料库在建设之时均未综合考虑与其他学科语料库之间建库标准及语料选取统一性的问题,故而很难实现基于上述语料库的跨学科对比研究。

有鉴于此,我们认为建设一个船海相关学科的学术英语语料库具有较强的现实性与迫切性。本课题在中国外语与教育研究中心“DEAP 学术英语语料库”建设项目的指导下,创建了 ShipDEAP 船舶与海洋工程学术英语语料库,下文将详细介绍 ShipDEAP 的建设过程与应用前景。

2 建库目标

建设 ShipDEAP 语料库旨在实现下述三项主要目标。其一,建成一个库容为 500 万词次且涵盖面广、规模较大、时效性强、能够体现该学科学术英语语言特征的专用语料库。其二,为船海领域的学术英语研究提供大量的语料资源支持。基于本库可探究该领域国际话语社团的通用话语体系,厘清其话语策略、知识构建方式等。其三,服务于船海学科的学术英语教学和词典编撰等工作,为其提供更多的检索与实例资源。

3 语料收集方案

所收集语料的代表性决定了研究结果能否推广至语料所代表语言或语境的整体 (Leech 1991: 27), 因此在确定语料收集方案时要尽可能保证语料的覆盖面与代表性。为实现上述目标, 本课题主要从以下 3 点进行考虑: 学科领域、期刊来源与文本构成。

3.1 学科领域

为全面体现船海学科学术英语的语言特征, 本语料库根据中华人民共和国教育部所颁布的《学位授予和人才培养学科目录 (2018 年 4 月更新) 》, 对标 “工学” 门类中的 “船舶与海洋工程” (学科代码 0824) 一级学科, 确定其下设三个二级学科: 船舶与海洋结构物设计制造、轮机工程和水声工程, 并将其作为 ShipDEAP 语料库的三个子库方向。

3.2 期刊来源

为确保 ShipDEAP 语料库的学科代表性, 本库所选取期刊均来源于船海领域 SCI 国际期刊, 着重考虑期刊的影响因子和分区两个因素。根据上述标准, 每个二级学科初步选定了 50 种 (共计 150 本) 学术期刊。随后, 分别与各二级学科的专家学者商讨, 并参照国内某 211 高校船舶与海洋工程 (双一流建设学科) 相关院系所提供的高质量国际期刊目录, 增加了一些行业公认但影响因子或分区不高的期刊 (如 International Ship Building Progress), 同时剔除掉了一些影响因子高但近五年发文与该领域相关性较低的期刊 (如 ASME Journal of Vibration and Acoustics)。在实际选定过程中又由于期刊资源获取问题, 最终 3 个二级学科各选择 12.7.7 种期刊, 共选取 26 种作为 ShipDEAP 语料库的语料来源, 可参见表 1。

表 1 ShipDEAP 语料库的来源期刊与文本数量

二级学科	来源期刊 (编号)	样本数量
船舶与海洋结构物设计制造 DCNA (Design and Construction of Naval Architecture and Ocean Structure)	Ocean Engineering (1)	25
	Marine Structure (2)	25
	International Journal of Naval Architecture and Ocean Engineering (3)	20
	Cold Regions Science and Technology (4)	15
	IEEE Journal of Ocean Engineering (5)	20

(待续)

(续表)

二级学科	来源期刊 (编号)	样本数量
船舶与海洋结构物设计制造 DCNA (Design and Construction of Naval Architecture and Ocean Structure)	Journal of Navigation (6)	15
	Ship Technology Research (7)	19
	Ships and Offshore Structure (8)	25
	Coastal Engineering (9)	25
	Applied Ocean Research (10)	30
	Journal of Waterway Port Coastal and Ocean Engineering (11)	31
	International Ship Building Progress (12)	10
轮机工程 ME (Marine Engineering)	Energy (1)	40
	Applied Energy (2)	45
	Fuel (3)	40
	Energy Conversion and Management (4)	45
	Applied Thermal Engineering (5)	45
	Control Engineering Practice (6)	20
	Journal of Marine Science and Technology (7)	15
水声工程 UAE (Underwater Acoustic Engineering)	Journal of the Acoustical Society of America (1)	43
	Journal of Sound and Vibration (2)	37
	Journal of Low Frequency Noise Vibration and Active Control (3)	34
	Acta Acustica United with Acustica (4)	37
	Applied Acoustics (5)	38
	Journal of Vibration and Acoustics, Transactions of the ASME (6)	37
	Ultrasonics (7)	34

3.3 文本构成

ShipDEAP语料库的文本构成科学严谨, 主要从文本语类和文本时效性两个方面进行了考虑。为均衡语类, 本库共收录了三种类型的论文, 分别是研究论文

(research articles)、综述论文(review articles)和书评(book reviews)。根据既定期刊文章占比构成,本库所收集文章以研究论文为主。为保证语料库的时效性,本库仅收录2016—2021年发表的文章。

4 语料库建设

4.1 语料收集与命名

在收集文本时,首先通过Web of Science文献库依次检索各二级学科既定期刊名称,将出版年份限定为2016—2021,并按照文献的“被引频次”排序,从高到低选取一定数量的文章。由于各期刊刊文量存在差异,故各期刊所下载文章数量也不尽相同。在语料收集前期,每个二级学科各收集了300篇(共900篇)文章,库容达700余万词次。随后,按照DEAP学术英语语料库预先制定的库容要求,即子库库容为500万词次,我们采取了平均抽样的原则,从每个二级学科随机抽取260篇左右的文章,详见表2。在收集过程中,我们发现船海领域有些综合性SCI期刊可能同时收录船海多个二级学科的学术论文,为此本项目组邀请船海专业的研究生结合其自身专业背景阅读文献标题和摘要,加以区分。

表2 ShipDEAP 语料库学科分布

二级学科	文本数量	库容(纯文本)	库容(头部信息标注版)
船舶与海洋结构物设计制造	260	1,796,064	1,845,136
轮机工程	250	1,843,885	1,898,099
水声工程	260	1,415,874	1,451,249
总计	770	5,055,823	5,194,484

DEAP语料库的其他子库在收录文本时,通常是选择下载PDF格式的论文,再将其直接或间接地转换为TXT格式文本(彭工 2018; 朱晓丽、吴敏 2021等),但船海学科的学术论文中存在大量的公式与图表,若采用上述方法,文本格式转换时会产生大量乱码,将大大增加后期文本清洁的难度。因此本项目选择直接复制网页HTML格式的论文至TXT文档的方式,在此过程中仅选择拷贝语料库研究需要的部分,避免后期再次手动删除语料库研究不需要的内容,最大程度地保证了语料的清洁度。每篇论文都是全文收录,包括标题、摘要、关键词、正文和致谢等,但不包括页眉、页脚、参考文献、附录等语言学信息不强的语篇结构部分。与此同时,在收录过程中详细记录各篇论文的题目、作者、URL等信息,生成各

子学科汇总统计表,方便后期需查阅原文或校对语料时快速定位查找。

ShipDEAP语料库按照“二级学科代码-期刊序号-文本编号及语类编码”的格式对所收集文本统一命名。其中,二级学科代码采用各二级学科英文名称首字母缩写的形式(详见表1);期刊序号由“J+序号”组成;本库所收录的三个语类中,研究论文缩写为RA,综述论文缩写为RV,书评缩写为BR。例如,DCNA-J01-001RA这一文件名代表的是:船舶海洋结构物设计制造二级学科第1本期刊所收录的第1篇研究论文。

4.2 语料清理

文本清洁程度将直接影响到后期基于语料库研究的信度。语料文本如不加以清理会导致词汇分析、搭配统计不准确,以及词性赋码出错或无法进行(梁茂成等2010:32)。ShipDEAP语料库的文本清理工作主要包括以下三个方面。

一是TXT文本格式的转换。由于语料库通用TXT文本格式为UTF-8,所以在收集得到所有TXT文本后,为避免因TXT文件编码而出现的软件识别问题,使用EncodeAnt软件对所有TXT文本进行批量文件编码转换。

二是格式问题的处理。虽然ShipDEAP语料库收集语料所采用的方式最大程度上保证了文本的清洁度,但手工录入语料时由于兼容性等原因不可避免地会出现一些符号、格式不符合规范的情况,如存在全角标点符号、多余的空格和空行等问题。针对此类问题,借助“文本整理器”设置批量整理方案,从而实现这些问题文本的批量清洁。

三是公式、特殊字符与图表的处理。由于船海学科特点,学术论文中存在大量的公式、符号及图表,这部分内容语言学信息不强,无法体现船海学术语言特征,会对检索结果造成一定程度的影响。为最大程度地保留学科专业信息、实现文本可读性,ShipDEAP语料库在保证句法结构完整的前提下,主要采取替换的方式进行处理。这一步骤与文本收集同步进行。

首先,文本中所出现的公式,无论其单独成行还是镶嵌于语句之中,均使用符号@进行替换,详见下例:

(1)

原文: Consider that a node initial located at X moves to r in the current configuration, then the deformation gradient tensor is calculated by:

$$F_{ij} = \frac{\partial x_i}{\partial X_j} = \sum_I \left(\frac{\partial N_I}{\partial X_j} x_{Ii} + \frac{h}{2} \frac{\partial \zeta_I N_I}{\partial X_j} y_{Ii} \right) \quad (28)$$

清理后 Consider that a node initial located at X moves to r in the current
的文本: configuration, then the deformation gradient tensor is calculated by:
@ (28)

其次，船海学术论文中还存在许多纯文本格式下无法显示其原有形态的特殊字符，主要包括上标或下标的符号，ShipDEAP 语料库统一将其替换为形似的可示字符，如将 $W_{compressor}$ 替换为 Wcompressor。

最后，对于文本中的图表，仅保留图表标题及注释，详见下例：

(2)
原文: **Table 1**
Main parameters of the ship model.

Parameters	Symbol	Value of ship model	Value of full scale ship
Waterline length (m)	L_{WL}	6.0	120.0
Waterline breadth (m)	B_{WL}	0.75	15.0
Draught (m)	T	0.25	5.0
Trim angle (deg.)	θ	0	0
Displacement (kg)	Δ	562.5	4,500,000
Wetted surface area (m ²)	S	4.872	1,948.8
Block coefficient	C_B	0.5	0.5

清理后 Table 1. Main parameters of the ship model.
的文本:

4.3 语料标注

经标注后的结构化语料能够帮助研究者深度挖掘语料库信息，实现语料增值。ShipDEAP 语料库的文本标注包括头部文献信息标注与词性赋码。

头部文献信息可以提供文献出版年份、期刊名称等外部属性信息，方便后期为服务不同的研究目的而生成子库。本库的头部文献信息标注使用XML语言，在文本采集时同步进行。预先设置好文献信息标注模板，将其复制粘贴至TXT文档开头的预留区域，并根据所采集文本的具体文献信息填充完整。具体标注信息与释义参见表3。此外，ShipDEAP 语料库使用TreeTagger自动词性赋码器完成词性赋码工作。

表3 头部文献信息标注

头部文献信息标注	释义
<Title></Title>	题目
<Author></Author>	作者
<Affiliation></Affiliation>	附属机构
<Year></Year>	出版年
<Journal></Journal>	期刊名
<Publishing_house></Publishing_house>	出版社
<Text_Collector></Text_Collector>	文本收集者
<DOI></DOI>	数字对象识别码
<URL></URL>	数字文本网址

5 基于 ShipDEAP 语料库的语言研究与教学研究

ShipDEAP 语料库在学术语言本体研究与学术英语教学研究方面均具有广阔的应用前景。学术语言本体研究方面，基于本库探讨船海学术英语在词汇、短语、句式乃至篇章结构等微观、宏观不同层面的语言使用特征，揭示船海领域学术论文的语篇组织策略，并进一步尝试探究该领域学术话语共同体的知识建构方式与人际互动手段，最终反哺学术英语教学。例如，基于 Swales（2004）所提出的语步分析法探究船海学科某一特定体裁文本的语步构建方式，以 Hyland（2005）的互动元话语框架为基础完善船海领域特有的互动元话语表达。此外，ShipDEAP 为学术英语研究提供了新的对比视角，可与已建成的数学（朱晓丽、吴敏 2021）、土木工程（章柏成、杨玲 2020）等其他学科的语料库进行跨学科对比研究，探究学术英语的学科差异性，归纳并描写本学科相较其他学科所特有的话语特征。

学术英语教学方面，ShipDEAP 可以为学术英语写作提供丰富、规范的学术语言使用范例。以吕桂、何安平（2014）和许家金（2017）等为指导，可基于本库编制船海学科专业学术英语词汇表与学术英语写作常用词块表，提高学生对词汇用法和实用搭配构式的掌握能力。此外，本库的建设便于教师开展语料库辅助下的船海学术英语写作教学。教师与学生均可以借助本库总结归纳国际通用的句式表达与语篇架构，辅以词汇表的助力，可有效缩小学生与专家在话语实践方面的差距（姜峰 2021：43），增强学习者的表达规范性和体裁意识，最终提高我国船海学者参与国际学术对话的能力。

ShipDEAP 语料库还可以广泛地应用于船海学科词典编纂、术语提取、机器辅

助翻译、自然语言处理等方面的研究。

6 结论

本文主要介绍了 ShipDEAP 语料库的建设过程与应用前景。作为 DEAP 学术英语语料库的子库之一,从确定建库方案、语料收集与命名再到语料清洁与标注都严格按照预先确定的规范进行,最大程度上确保了语料库的代表性、准确性以及与已建成其他学科子库之间的学科可比性。同时,ShipDEAP 语料库共包含纯文本与词性赋码两个版本,为学者深度检索语料和开展学术研究提供了便利。此外,本库的应用前景十分广阔,不仅能够基于此库进行船海领域的学术语言本体研究,更能将其应用于学术英语教学,为我国船海学者更好地阐述学术观点提供国际通用的语言范式和策略,更好地进行国际学术交流,讲好中国海洋故事。目前,ShipDEAP 为共时单语语料库,为更好地发挥该库的作用,后续可以考虑适当扩容建设。

参考文献

- FREDDI M. Arguing linguistics: corpus investigation of one functional variety of academic discourse [J]. *Journal of English for Academic Purposes*, 2005, 4(1): 5-26.
- HYLAND K. Stance and engagement: a model of interaction in academic discourse [J]. *Discourse Studies*, 2005, 7(2): 173-192.
- HYLAND K. Corpora and written academic English [C]//BIBER D, REPPEN R. *The Cambridge handbook of English corpus linguistics*. Cambridge: Cambridge University Press, 2015: 292-293.
- LEECH G. The state of the art in corpus linguistics [C]//AIJMER K, ALTENBERG B. *English corpus linguistics*. London: Longman, 1991: 8-29.
- MCGRATH L, KUTEEVA M. Stance and engagement in pure mathematics research articles: linking discourse features to disciplinary practices [J]. *English for Specific Purposes*, 2012, 31(3): 161-173.
- SWALES J. *Research genres: explorations and applications* [M]. Cambridge: Cambridge University Press, 2004.
- TEUBERT W. My version of corpus linguistics [J]. *International Journal of Corpus Linguistics*, 2005, 10(1): 1-13.
- VALIPOURI L, NASSAJI H. A corpus-based study of academic vocabulary in chemistry research articles [J]. *Journal of English for Academic Purposes*, 2013, 12(4): 248-263.
- 范凤祥. 轮机英语词汇的量化特征 [J]. *大连海事大学学报 (社会科学版)*, 2006

- (2): 128-132.
- 冯欣, 吴菁菁, 齐晖, 许家金. MedAca医学学术英语语料库的创建[J]. 语料库语言学, 2017(2): 107-113.
- 姜峰. 中国学生学术话语能力发展与教学有效性研究[J]. 外语与外语教学, 2021(6): 34-44.
- 姜峰. 近四十年国内外学术英语研究: 主题与进展[J]. 外语教学与研究, 2022(3): 413-424.
- 姜峰, HYLAND K. 互动元话语: 学术语境变迁中的论辩与修辞[J]. 外语教学, 2020(2): 23-28.
- 李芳萍, 李舰君. 科研论文英文摘要中的衔接——以船舶与海洋工程为例[J]. 哈尔滨职业技术学院学报, 2015(6): 110-111.
- 梁茂成, 李文中, 许家金. 语料库应用教程[M]. 北京: 外语教学与研究出版社, 2010.
- 梁茂成, 刘霞. 语篇内部的短语学特征分布模式探索——以学术论文为例[J]. 解放军外国语学院学报, 2014(4): 1-11.
- 吕桂, 何安平. 专门用途英语词汇语义共选特色探究[J]. 山东外语教学, 2014(1): 36-42.
- 吕靖, 邓飞. AgriDEAP农学学术英语语料库的创建[J]. 语料库语言学, 2020(2): 89-99.
- 彭工. BioDEAP生命科学学术英语语料库的创建[J]. 语料库语言学, 2018(2): 69-77.
- 徐昉. 学术英语写作研究述评[J]. 外语教学与研究, 2015(1): 94-105.
- 许家金. 体裁短语学视角下的医学学术英语词典研编[J]. 外语与外语教学, 2017(6): 52-60.
- 颜天明, 侯慧凡. 基于语料库的轮机英语词汇教学实践探讨[J]. 航海教育研究, 2018(2): 46-49.
- 张凤婷. 中外作者海洋工程学术论文引言中转述动词对比研究[D]. 大连: 大连海事大学硕士学位论文, 2020.
- 章柏成, 杨玲. CivDEAP土木工程学术英语语料库的创建[J]. 语料库语言学, 2020(1): 78-87.
- 郑军. 海洋船舶英语论文语料库创建的可行性研究[J]. 学园, 2014(33): 10-11.
- 朱晓丽, 吴敏. MathDEAP数学学术英语语料库的创建[J]. 语料库语言学, 2021(2): 127-135.

通信地址: 150001 黑龙江省哈尔滨市 哈尔滨工程大学国际合作教育学院

《学习者语料库研究遇见二语习得》述评

北京外国语大学 李维静

Bert Le Bruyn & Magali Paquot (eds.). 2021. *Learner Corpus Research Meets Second Language Acquisition*. Cambridge: Cambridge University Press. xiii+275pp.

1 引言

近年来,基于学习者语料库的二语习得研究持续受到关注,已成为二语习得研究的重要分支。这一分支的研究课题汲取了学习者语料库研究和二语习得研究的优势,同时缩小了这两类研究间的距离,促进了这两个领域的发展。《学习者语料库研究遇见二语习得》一书聚焦两个领域间的合作,收录了9篇具有代表性的实证研究文章。这些研究均来自学习者语料库和二语习得领域的核心学者,旨在介绍学习者语料库研究的最新进展,以及如何利用这些先进的方法探究二语习得的核心问题。

2 内容简介

本书共包含十二篇文章,可分为五个部分。第一部分即第一篇文章,是编者对全书内容的介绍。第二部分包括第二至四篇文章,主要从母语迁移角度探讨中介语的“普遍性”(universal tendencies)和“变异性”(cross-linguistic influences)。第三部分包括第五至八篇文章,关注学习者语言水平的历时变化。第四部分包括第九和第十篇文章。这两篇文章以解决二语习得问题为导向,为学习者语料库的数据分析和语料库建设提供了方法论支持。第五部分包括第十一和十二篇文章,是关于全书内容的总结述评。

本书的第一部分是两位编者 Le Bruyn 和 Paquot 对全书内容的介绍。她们首先从学习者语料库与二语习得研究的差异切入,阐明本书的目的。从源头上看,学习者语料库的孕育具有明显的应用导向,即扩大现有的语料来源,为语言学习和外语教学服务。这与理论驱动的二语习得研究有着本质的差异,因而两个领域在学习者语料库建成初期(20世纪90年代)并无太多的交集。在21世纪初,学习者语料库 CHILDES 已被系统地应用于一语习得问题的探讨,但在二语习得研究课题

中使用学习者语料库的仍是少数。近十年来,两个领域才开始有一些合作,成功的案例包括“接口假说”(Interface Hypothesis; Rankin 2009)和“可加工性理论”(Processability Theory; Bonulla 2015)。编者援引Myles(2015)的观点,重申学习者语料库在二语习得研究中的潜力,点明本书目的,即促进两个领域的系统合作。接着,编者介绍了该书文章的收录标准和编排原则。她们按照研究主题将所收录的九篇实证研究分为三部分:(1)中介语的“普遍性”和“变异性”;(2)二语学习者语言水平的历时变化;(3)语料库分析和发展。编者对各部分的文章予以总结提炼,以便读者快速定位相应的研究。笔者对全书的划分遵从编者的编排原则,下面将逐一对上述部分进行介绍。

本书的第二部分由三篇文章构成,围绕中介语的“普遍性”和“变异性”展开。

第一篇研究聚焦学习者的冠词使用,作者为Ionin和Diez-Bedmar,旨在探讨学习者语言水平、母语中是否有冠词系统、母语是否存在特指性(specificity)这三个因素对英语冠词正确使用的影响。研究从“剑桥学习者语料库”(Cambridge Learner Corpus)中抽取了语言水平在B1-B2级别,母语为俄语和西语的英语学习者作文(体裁为描述性文本和创造性自传),共抽取200篇(各水平50篇)。基于所抽取的学习者语料,研究采用计算机辅助错误标注的方法,对作文中定冠词、不定冠词和零冠词的使用进行了五个层级的手工标注,包括冠词是否正确使用、所使用冠词的类别(例如限定冠词)、所使用的具体冠词(例如零冠词)、所修饰名词的类型(例如专有名词)和单复数情况以及名词词组的语法语境(例如做表语、出现在存在句中)。在这一研究中,作者将冠词正确使用定义为每篇作文中每百词正确/错误冠词数量的(标准化)中位数,以便进行统计分析。通过比较不同水平、不同母语背景学习者的数值,研究发现:当母语(本研究中为西语)有冠词系统时,学习者语言能力与冠词正确使用呈正相关关系;对于母语(本研究中为俄语)没有冠词系统的学习者,其语言能力与冠词正确使用无显著关联。这一结论与以往二语习得实验结论一致,反映了中介语的“变异性”。对特指性的统计分析并未发现这一语义特征在中介语中的“普遍性”。

Werner、Fuchs和Götz的研究旨在诊断学习者过去时间表达的问题,探究中介语中影响过去时间表达的因素。英语中,现在完成时和一般过去时均可用于表达过去时间意义。而在具体语境中,这两种形式的选择或替换往往给学习者带来困扰。作者发现,在关于过去时间“形式-意义”配对的二语习得研究中,母语迁移和普遍学习机制均可对现在完成时和一般过去时替换产生影响。作者采用学习者语料库方法对上述两种因素进行检验。研究选取了母语为德语(母语中有现在完成时和一般过去时替换)和汉语粤方言(母语中无现在完成时和一般过去时替换)的学习者以及英语本族语者的口笔语语料(其中学习者口笔语语料分别来自LINDSEI和ICLE,本族语者的口笔语语料分别来自LOCNEC和LOCNESS),并

对其中4,000条现在完成时用例和3,000条一般过去时用例进行人工标注。标注内容包括动词语义、动词情状体、时间状语、句式（主动或被动）、动词类型（规则或不规则）、是否有否定、时态持久性（例如前文中是否出现过SP）。研究采用MuPDAR（Multifactorial Prediction and Deviation Analysis with Regressions）多因素回归建模的方法，对比学习者和英语本族语者的现在完成时和一般过去时的交替使用情况，定位了母语为德语及汉语粤方言的学习者偏离目标语用法的语境。除此之外，研究结果显示：能够预测学习者和本族语者现在完成时和一般过去时替换差异的显著因素为认知性因素（例如动态持久性）而非学习者母语差异，因此研究认为学习者中介语的现在完成时和一般过去时替换具有“普遍性”特征。

第二部分最后一篇文章来自Meriläinen，文章通过两个具体案例探讨了二语习得过程中跨语言影响（cross-linguistic influence）和普遍习得过程（universal acquisitional process）在中介语中的交互作用。跨语言影响指学习者中介语的变异性，即学习者的二语习得受到母语的影响，不同母语背景学习者的二语表现具有差异性。而普遍习得过程认为，二语习得和母语习得过程均由普遍认知机制引导，因而学习者的二语表现具有普遍性，与学习者的母语无关。此研究中，两个案例分别讨论了间接引语中主谓顺序的倒置现象和介词省略现象。这两个现象在以往的二语习得实验中反映出学习者中介语的普遍性，指示学习者的二语习得过程有过度概括（overgeneralization）和简化（simplification）的认知倾向。为检验这一结论的真实性，作者从学习者语料库ICLE和MEC中抽取了母语为瑞典语、芬兰语、德语、汉语、日语的学习者笔语语料，从LOCNESS中抽取了英语本族语者笔语语料。通过对比学习者和本族语者语料中这两个现象的频数，研究发现：虽然学习者的中介语表现具有普遍性（两个现象在学习者语料中的频数显著高于本族语者的频数），但是母语与英语越相似的学习者，与英语本族语者的频数差异越小。也就是说，对于中介语中的主谓顺序倒置和介词省略现象，跨语言影响的作用比普遍习得过程更为显著。

下面的四篇文章构成了本书的第三部分，核心话题是二语学习者语言水平的测量及其历时变化。

第一篇文章的作者为Polio和Jo Yoon，旨在探索学习者作文中的二元/三元词能否作为测量学习者作文准确度的指标，以及这个指标与经典CAF指标（即复杂度、准确度、流利度）之间的关系。这一研究采用基于用法的研究路径，认为二语学习是学习者不断总结语言输入规律的过程。语言输入中蕴藏的构式、频率、形义配对、典型用例、词语组合等信息对二语习得起重要作用。作者聚焦学习者作文中的二元和三元词语组合，以COCA作为参照语料库类比学习者可能接触到的语言输入，以自建的学习者笔语语料库、类型语料库和论文库代表学习者的二语写作表现。在研究第一步，作者提取了学习者作文中的二元及三元词，随后利

用这些词在COCA中的频数信息计算出这些词的平均互信息值(MI score)及缺失比率(指在COCA语料库中未出现,但是出现在学习者语料库中的二元及三元词比例)。第二步,作者手工标注出这些二元及三元词的使用正误。为探究这类词语组合对学习者的作文准确度的预测能力,作者在统计分析时加入了经典的测量指标,包括单位长度(unit length)、分句复杂度(clause-level complexity)、短语复杂度(phrase-level complexity)、词汇复杂度(lexical sophistication)、准确度、作文得分在内的6个大类,共计11个具体指标。本研究共得出了3个结论:(1)通过错误统计,作者发现参照语料库中缺失的大部分二元(87%)及三元词(77%)属于错误使用;(2)利用因子分析,作者发现二元词缺失比率、二元/三元词平均互信息值仅与经典准确度指标聚合在一起,说明二元/三元词具有成为测量学习者作文准确度指标的潜力;(3)进一步的回归分析显示,二元/三元词指标(47.6%)虽然可以较好地预测学习者的作文得分,但预测能力不及经典的准确度指标(48.5%)。

第二篇文章来自Paquot、Naets和Gries,目的是探究句法层级的短语使用能否作为复杂度指标追踪学习者的历时变化。与上一篇研究相比,本文作者从形式上的词语共现深入到句法层级,聚焦直接宾语构式中动词与宾语之间的共现。另外,虽然两篇文章的作者都选择互信息值作为检验搭配强度的标准,但是本文作者强调,互信息值适用于测算低频共现词之间的共现强度,因而具有指示短语复杂程度的作用,是一个复杂度指标。为实现研究目标,本研究选用的学习者语料库为历时语料库LONGDALE,参考语料库为网络语料库ENCOW14 AX。研究首先从LONGDALE语料库中抽取所有的直接宾语构式。然后参考各构式在ENCOW14 AX中的频数信息,研究得到每一种直接宾语构式的互信息值。最后,计算出每一篇学习者作文中所有直接宾语构式的平均互信息值,并以此数值代表该篇作文的短语复杂度。在统计分析阶段,作者使用了语料库语言研究中较为先进、复杂的建模方法——混合效应逻辑斯蒂回归模型,模拟学习者的学习时长、语言水平及写作主题(固定变量)对学习者的作文中短语复杂度(平均互信息值;反应变量)的影响。学习者的个体差异则作为随机变量投入模型。模型的结果显示,随着学习者语言水平的提高,学习者作文中的短语复杂度得到显著提高;同时,学习者的学习时长(例如所在年级)对短语复杂度并没有显著的预测力。在这两个结论中,前者说明短语复杂度可以作为测量中高级学习者语言能力的指标;后者说明在短语复杂程度上,学习者的作文并未展示出历时变化。

接下来两篇文章的特点是使用学习者历时语料库探究学习者语言水平的历时变化。

Tracy-Ventura、Huensch和Mitchell的研究聚焦学习者的词汇系统,侧重探讨学校语言学习结束后学习者词汇多样性的变化。以往的研究发现,在学校学习结

束后,学习者的词汇变化表现出明显的个体差异性,与学习者的年龄、语言接触时间的长短、态度、动机、语言水平和语言学习结束后的语言接触/使用有关。这些因素中,能够为学习者词汇变化提供较强解释力的因素仅有最后两项。因此,为探究这两种因素如何影响学习者的词汇多样性,本研究选择了一个特殊的语料库——LONGSNAP (Language and Social Networks Abroad Project)。该语料库设计严密,共包含56位高级学习者(母语为法语和西语)6年时间里(共7个时间段:学习前,学习中1、2、3和学习后1、2、3)3类任务(访谈、讲故事、议论文写作)的口笔语语料。该语料库的设计符合二语习得研究方法中的“前测-中测-后测”原则,能够较好地反映出学习者词汇系统的变化。具体操作中,研究将词汇多样性定义为统计数值D得分和动态平均类形比(MATTR),用学习最后阶段(学习中3)的D得分和MATTR数值代表学习者的语言水平。为获取学习者在学习结束后的语言接触/使用情况,本研究补充了相关的问卷数据。结合这两类数据,研究发现语言任务变量与时间变量存在明显的交互:当学校学习结束后,学习者的词汇复杂度在两类口语任务中有显著的提高,但在写作任务中没有显著变化。通过对语言任务、时间和语言水平的建模,研究还发现(学习结束时)语言水平越高的学习者,词汇多样性的变化数值越小,指示较高的语言水平对学习者的词汇多样性具有保护作用。

Verspoor、Lowie和Wieling的历时研究是对Verspoor *et al.* (2012) 横向(cross-sectional)研究的拓展,旨在探索学习者作文中词汇、句法系统的变化情况以及二者之间的习得顺序。本文作者使用的语料库是自建的数据库,收录了22名荷兰高中生在23周时间内(即一个学年)的每周作文。为探究这些学生的作文在23周内的变化情况,作者提前对每篇作文的整体质量进行考核并给出相应分数(打分区间为1—5分,其中1分为最低分,5分为最高分)。操作过程中,研究以23篇作文中头两篇的平均分代表学习者这段时间内的初始语言表现,以末尾两篇的平均分代表学习者的当前语言表现。对比两个时间段的得分,研究发现学习者作文的整体得分有明显的提高,指示着学习者语言能力的提升。为进一步分析学习者词汇、句法系统的变化,研究将统计数值Guiraud和T单位平均长度(mean length of T-unit)作为词汇和句法复杂度的测量指标。研究结果显示,虽然每个学习者的词汇、句法变化情况都有差异,但这两个指标整体呈现出先升后降的趋势,且词汇发展先于句法发展。

本书的第四部分包含两篇文章,为学习者语料库的分析和建设提供了方法论支持。

Wulff和Gries的研究展示了语料库方法MuPDAR(F)在处理二语习得问题中的优势。学习者的二语习得是一个复杂的过程,受到多种语言内部因素(例如语言输入的数量和质量)和语言外部因素(例如语言学习发生的环境)的影响。

同时,二语习得具有个体差异性,每个学习者的发展速度和习得路径均有不同。MuPDAR(F)方法的优势在于能够同时考虑这两个方面的影响,对学习者的中介语进行细颗粒度的分析。MuPDAR(F)在具体操作中有三个步骤:(1)对目标语言使用建模(R1);(2)将学习者语言使用投入模型R1,预测相同语境下目标语言的表现;(3)对目标语言与学习者语言的差异建模(R2),分析影响二者差异的因素。通过上述步骤可以看出,MuPDAR(F)方法突破了传统频数比较的局限性,能够对“学习者中介语-目标语”的差异作综合的概率性诊断。Wulff和Gries的研究聚焦学习者的属格交替现象('s属格和of属格),详细介绍了MuPDAR(F)的操作方法,包括数据处理、模型选择、参数解读等细节。研究发现:(1)对于学习者整体来说,在投入模型的10个因素中,所属与被所属名词间的距离、所属名词的单复数以及音节交替差异这3个因素会导致学习者做出与目标语者不同的选择;(2)对于不同母语的学习者来说,母语为汉语的学习者比母语为德语的学习者更容易受到所属名词复杂度的影响;(3)话题会影响学习者的个体表现。

Bell、Collins和Marsden的文章将学习者语料库、二语习得研究和语言教学联系起来,为建设能够反映二语发展情况、辅助语言教学的“课程”语料库提供了思路。文中详细介绍了语料库设计阶段和语料库建设阶段的注意事项。在语料库设计阶段,研究者需要对学校课程(如课堂时间、周课时量等)、学习者概况(如年龄、年级、兴趣、语言水平、母语背景等)、语言任务(如任务模式、任务类型等)及学习者中介语的测量指标(如词汇多样性指标、句法复杂度指标)有一个全面的了解,以便做出最符合当前研究目标的决策。在语料库建设阶段,作者强调了对小型样本数据进行预实验的重要性。这一步可以检验研究者决策的合理性,同时也保障了后续大型语料库的有效性。另外,文章还指出了当前语料库建设过程中的实际问题,如转写问题(如转写的内容、编码的系统)和错误标注问题(如错误标注的框架、标注者的选择),并给出两点建议:(1)信息公开,即建设者需要明确地说明问题是如何解决的;(2)标准化,即使用统一的编码或标注体系,以提高语料库之间的可比性。

本书的第五部分是两位学者对全书内容的总结和评述。

第一篇评述的作者是Granger。作者采用语料库的视角,首先对学习者语料库研究和二语习得研究的长处与不足进行了介绍。Granger认为,二者的合作恰好可以取长补短,相互促进。其次,Granger对本书收录的九篇实证研究进行了总结,提炼出当前两个领域合作的方向,分别是语料库的建设和使用、母语迁移研究和二语发展研究。最后,Granger明确了学习者语料库研究和二语习得研究合作的前景。她表示虽然目前两个领域还没有互相接纳,但是只有加强合作,才能够促进两个领域之间的互哺。

第二篇评述来自Myles,她从二语习得角度出发,阐明了学习者语料库研究和

二语习得研究相互合作的意义：现有二语习得理论（例如动态系统理论）的发展和验证需要多样、密集的语言数据支撑，需要学习者语料库的资源；同时，学习者语料库的建设和分析需要参考二语习得理论的框架，尤其要关注收集和利用元信息（例如学习者概况、任务信息等），进而更有针对性地解决二语习得问题。从本书中收录的九篇代表性文章来看，Myles认为两个领域间的距离正在逐步缩小，主要表现在两个方面：（1）研究均有明确的习得理论指导；（2）研究的数据量更大、任务类型更多、学习者范围更广。但是，Myles也指出，学习者语料库研究和二语习得研究的系统合作还没有实现。她呼吁两个领域间资源、方法论和经验共享，例如使用相同的诱发任务、统一的转写标准和编码软件等，以推进两个领域间的深度交流。

3 简评

本书汇集了九篇代表性的文章，体现了当前学习者语料库领域和二语习得领域合作下的新进展，向读者展示了两个领域相互学习、相互促进的积极作用。

一方面，相比于传统的二语习得研究，基于学习者语料库的二语习得研究借助语料库语言学的统计方法，能够协同考虑影响语言使用的多个因素及各因素间的交互（如语言水平、任务类型、话题、母语迁移等），进而帮助我们理解二语习得的复杂系统。其次，学习者语料库内大量的真实语料为二语习得研究提供了丰富的学习者语言使用资源。这些大规模的真实语料突破了传统二语习得内省和诱导数据的局限性，能够增加研究结论的概括性和普遍性（王立非、孙晓坤 2005：19）。同时，语料库中的语言资源使得多样化的语言特征研究，如搭配、类联结、词块、语块、语义韵研究等成为可能，是对二语习得研究范畴的拓展。

另一方面，二语习得研究的理论框架和方法论也能为学习者语料库研究所用。学习者语料库研究有必要参考二语习得理论来解读研究结果，提高研究的理论价值（Granger 2009）。另外，二语习得研究中的特定分析方法，如探究母语迁移的模型（Jarvis 2010）也可作为计算机辅助错误分析和中介语对比分析方法的有益补充。除此之外，二语习得研究对元信息的严格把控既是学习者语料库研究的挑战，又为学习者语料库的设计和完善提供了新的思路。

总的来说，本书设计严密，编排合理。但是本书也存在一些不足之处。

第一，书中介绍的学习者语料库 ICLE 和 LINDSEI 需购买后才能使用，这极大地增加了研究的成本。实际上，国内外有不少免费的语料库可供使用。例如 TLC（Trinity Lancaster Corpus；Gablasova *et al.* 2019）和 VOICE（version 3.0 2021）聚焦学习者的口语表现，可作为 LINDSEI 语料库的替代。这两个语料库涵盖不同语言、文化背景的英语学习者，涉及不同类型的口语任务，研究者可以直接在配套

网页上进行检索。对于中国的研究者来说,本土学生的二语习得问题是重中之重。聚焦中国学习者二语写作的语料库CLEC(桂诗春、杨惠中 2003)、WECCL 2.0(文秋芳等 2009)、TECCL(许家金 2016)、iWriteBaby(许家金 2019)可以为研究者提供丰富的作文语料。上述学习者作文语料库均可以通过北京外国语大学的CQPweb平台检索和使用。

第二,本书对目标读者有一定的要求。从标题来看,本书的目标读者是对学习者语料库研究和二语习得研究感兴趣的学者。但实际上,本书有较高的阅读门槛,要求读者有相关的语料库研究经验,懂得基础的统计知识。比如研究者需要了解如何提取目标语料、如何使用R语言软件(例如数据的导入和分析、建立模型)、如何将研究结果图示化、如何解读数据结果等。这些技术细节对于统计经验较少的学者来说十分具有挑战性,使得本书中的研究难以复制。有意参考书中研究方法的学者可以先参阅Levshina(2015)以及Brezina(2018),以更好地把握研究中的技术细节。

参考文献

- BONULLA C. From number agreement to the subjunctive: evidence for Processability Theory in L2 Spanish [J]. *Second Language Research*, 2015(1): 53-74.
- BREZINA V. *Statistics in corpus linguistics* [M]. Cambridge: Cambridge University Press, 2018.
- GABLASOVA D, BREZINA V, MCENERY T. The Trinity Lancaster Corpus: development, description and application [J]. *International Journal of Learner Corpus Research*, 2019(2): 126-158.
- GRANGER S. Corpus research applications in second language teaching [J]. *Annual Review of Applied Linguistics*, 2009(31): 205-225.
- JARVIS S. Methodological rigor in the study of transfer: Identifying L1 influence in the interlanguage lexicon [J]. *Language Learning*, 2010, (2): 245-309.
- LEVSHINA N. *How to do Linguistics with R: data exploration and statistical analysis* [M]. Amsterdam: John Benjamins, 2015.
- MYLES F. Second language acquisition theory and learner corpus research [C]// GRANGER S, GILQUIN G, MEUNIER F. *The Cambridge handbook of learner corpus research*. Cambridge: Cambridge University Press, 2015: 309-332.
- RANKIN T. Verb second in advanced L2 English: a learner corpus study [C]/*Proceedings of the 10th Generative Approach to Second Language Acquisition Conference*. Somerville, MA: Cascadia Proceedings Project, 2009: 46-59.
- VERSPOOR M, Schmid M, Xu X. A dynamic usage based perspective on L2 writing [J].

Journal of Second Language Writing, 2012(3): 239-263.

VOICE. The Vienna-Oxford International Corpus of English (version VOICE 3.0 Online) [DB/OL]. Austrian Centre for Digital Humanities and Cultural Heritage, Austrian Academy of Sciences, 2021.

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.

王立非, 孙晓坤. 国内外英语学习者语料库的发展: 现状与方法[J]. 外语电化教学, 2005 (5): 19-24.

文秋芳, 梁茂成, 晏小琴. 中国学生英语口语语料库2.0[M]. 北京: 外语教学与研究出版社, 2009.

许家金. “中国学生万篇英语作文语料库” 介绍[J]. 语料库语言学, 2016 (2): 108-112.

许家金. iWriteBaby 中国学习者英语语料库的创建[J]. 语料库语言学, 2019 (1): 105-109.

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/国家语言能力发展研究中心

《情态与历时构式语法》述评*

北京航空航天大学 刘娜 李福印

Martin Hilpert, Bert Cappelle & Ilse Depraetere (eds.). 2021. *Modality and Diachronic Construction Grammar*. Amsterdam: John Benjamins. v+251pp.

1 引言

过去20年间,构式语法理论蓬勃发展。该理论在历史语言学领域的广泛应用促使历时构式语法(Diachronic Construction Grammar,简称DxCG)作为独立的语言分析范式形成。DxCG从构式语法的节点和网络观出发,提倡对语言变化进行“网络变化”描写(Hilpert 2021: 61)。情态作为语法研究的核心话题,近年来开始被构式语法领域的学者关注。文集《情态与历时构式语法》便是在这一背景下出版的。该文集共八章,2021年由约翰·本杰明出版公司出版,三位编者Martin Hilpert、Bert Cappelle和Ilse Depraetere均是构式语法领域的杰出学者。文集结合定性和定量分析法,刻画了不同语言中多个情态范畴的构式化和构式变化路径,极大地拓宽了构式语法的研究领域,深化了对情态范畴的研究,值得学界关注。本文首先介绍各章内容,之后简要评价。

2 内容简介

第一章“DxCG中的情态:旧题新解”出自三位编者。结合DxCG和情态研究现状,作者先介绍了文集的三个研究主题:情态构式网络的组织、情态构式形—义历时发展的路径,以及构式化与构式变化的区分;随后简述各章核心内容和贡献;最后总结了本文集的贡献和启示。

第二章“缩略词、构式和构式变化”出自Robert Daus。作者关注了一个边缘性语言现象:英语情态缩略词can't、won't和'd。既往课本、语法书或研究多将缩略式界定为完整式(cannot、will not和would)的口语变体,认为它们对等。作

* 本文系国家社科基金项目“汉语动补结构的宏事件历时语言类型学研究”(21BYY045)的阶段性成果。

李娜为本文通讯作者。

作者贡献:

刘娜:选题构思、讨论结论、初稿撰写、字数占比(90%);

李福印:字数占比(10%)、修改润色。

者则主张缩略式是独立的形义结合体（即构式），应该被作为单独的类别整合到情态系统中。理由是：形式上，缩略式与完整式差异明显（音节压缩、韵律及句法受限等），表明前者在构式网络中已成为独立节点。使用频率上，在过去200年间缩略式的用例已超过完整式。区别性搭配分析（distinctive collexeme analysis）结果显示缩略式均偏好搭配认知/情绪类单音节动词，而完整式偏好搭配言语交际动词。语义上，三个缩略式与完整式偏好表达的情态类型互不相同。作者通过提取显著搭配、标注变量以及建立普通线性混合效应模型，确定了显著影响语言使用者选择缩略式或完整式的因素：二者的情态功能差异。总体上，形式差异、使用频率的增加、搭配偏好的变化以及使用功能的改变已将缩略式与完整式完全区分开，表明前者已成为独立构式，存贮于语言使用者的构式网络中。

第三章“德语 wissen/verstehen-情态构式语法化程度的实证研究”出自 Volodymyr Dekalo。研究亮点是将语法化程度量化。既往研究通常基于某构式同构项类型的扩展判定其语法化程度增强，而作者对此非常审慎。[wissen/verstehen V]（能够做……）是现代德语的近义情态构式，由动词 wissen/verstehen 语法化而来。数据显示两个构式类符频率差异不显著，因此无法根据同构项比较二者的语法化程度。作者对影响二者使用的语言因素进行变量标注，建立混合效应二元逻辑回归模型，确定二者在动词补语位置、是否与其他情态/时态助动词连用以及主语有生性上存在显著差异，证实 [wissen V] 的语法化程度高于 [verstehen V]。作者将这一结果与计算二者同构项类型的结果进行对比，发现后者不仅得出相反结论，而且无法反映两个构式语法功能的变化。作者呼吁应从构式使用特征出发测量语法化，不应假定语法化程度的增强总是反映在同构项类型的扩展上。

第四章“verdienen-构式在词汇和语法环境中的句法变异”出自 Gabriele Diewald、Volodymyr Dekalo 和 Dániel Czicza。本文通过共时变异构拟了历时演变。德语动词 verdienen 有“赚取”和“值得”义，可接名词和动词，形成 [verdienen+N]_{赚取}、[verdienen+N]_{值得}、[verdienen+V]_{应该做} 和 [verdienen+be V-ed]_{应该被做} 这四类构式。作者认为当下共存的四个构式实际对应 verdienen-构式语法化的不同阶段。通过语料观察，发现 [verdienen+N]_{赚取} 需使用有生主语和具体宾语，而 [verdienen+N]_{值得} 多使用抽象宾语，说明前者的词汇义更强、是语法化的源头。[verdienen+be V-ed]_{应该被做} 的使用表明 verdienen-构式中宾语类别从名词短语扩展至动补结构，说明该构式的同构项类型发生扩展。[verdienen+V]_{应该做} 主语为抽象实体，具有道义情态义，是该构式当前语法化的（暂时）最终状态。对四个构式进行简单搭配分析（simple collexeme analysis）并按语义归类，发现 [verdienen+N]_{赚取} 主要搭配“拥有/数量”义名词，语义差异小；[verdienen+N]_{值得} 主要搭配“认知、感觉”义等六类抽象名词，语义差异增加；[verdienen+be V-ed]_{应该被做} 显著搭配11个动词范畴，语义差别进一步增大；[verdienen+V]_{应该做} 则偏好搭配静态动词。本

研究证实“共时分布可成为探究历时变化的窗口”(Kuteva 2001: 9)。

第五章“加利西亚语(Galician)副词certamente和seguramente的语法构式化”出自Vitor Míguez。拉丁语名词mens(头脑,情绪)演变为现代罗曼语副词后缀-mente(表工具或方式)是语法化的经典案例。本文以“确定”义副词certamente和seguramente为例进一步探究了该后缀从表工具/方式到表认识情态,再到表交互主观性的语法构式化历程。既往对交互主观化的研究曾提出单向性演变假说,认为演变须遵循“非主观性→主观性→交互主观性”顺序(Traugott & Dasher 2001: 281)。但对比历时语料后作者发现certamente和seguramente的交互主观义(表加强论断)都早于主观义(表认识)出现,因此对该假说形成挑战。此外,seguramente在发展过程中丧失了交互主观义,例示了“去交互主观化”过程。作者呼吁学界关注认识/示证标记何时以及为何会出现不符合常规发展路径的情况。

第六章“构式展现:非预期义补语句中的后情态助动词”出自Rea Peltola。作者考察了现代法语和芬兰语中“惊叹”义小句构式[[心理动词/名词/形容词]+pouvoir]和[[态度动词]/名词/形容词]+pitää]的形成和演变。经对比历时语料,作者发现:两个构式都表征超越了话语既定认识或评价界限的事件,即“出乎意料”。但说话者识解两个构式的基础不同:pouvoir-构式源自表可能性的情态助动词,在真值层面发挥作用,侧显(profile)“p的存在和非p的不存在”(即最终p而非非p);pitää-构式源自表必然性的情态助动词,与英语动词should(应该)的识解类似,侧显“存在一系列主语所指偏好的q(q1、q2、q3等),被另一个潜在的可选项p否决”(即最终p而非q)。本研究表明语法化过程中构式的词汇和语法成分存在相互作用。

第七章“纵聚合项内部和纵聚合项之间的水平联接:德语指示语的构式网络”出自Elena Smirnova。作者构拟了德语指示性和陈述性言语行为(Peter asks Anna to go home. / Peter says that he is going home)的构式网络。通过对比两个构式及其变体在新高地德语时期的频数分布和变化趋势,本文一方面构建了指示语构式家族成员间的水平联接,确定了构式原型;另一方面通过对比该构式家族与陈述语构式家族的用例及分析对应关系,构建了两个家族间的水平联接。在概念层面上,本文提出这两类纵聚合联接构成了语态构式网络中不同的水平联接。在实证层面上,本文展示了如何借助语料库数据提取和解析这些水平联接关系。

第八章“日语祈使标记koto的构式化”出自Etsuyo Yuasa。既往研究将koto-构式的祈使义归结于构式中的命令类动词,本文则提出不同观点:命令动词多用于法院判决文本,koto-构式常用于口语中,二者使用语境不同。作者主张koto-构式是独立的构式,具有独特的句法和语义特征。语义上,koto受所搭配命令动词的影响,通过语用强化产生“命令”情态义。语义更新促使koto的句法表现向典

型情态构式的句法结构靠拢,后者允准前者使用其结构。至此,新意义与新形式相匹配,构式化过程完成,koto-构式诞生并成为构式网络中的新节点。本文还探讨了文化对言语行为模式的影响,比如日本文化使日语母语者偏好间接交流,因此表命令时,说话者思而不全言,让听话者推断隐含义,这种实践促进了悬停小句(即省略主句)的形成。作者最后指出koto由名词补语标记演变为情态标记不是孤例,其构式化过程可为其他日语情态构式的形成提供参照。

3 简评

情态是人类语言中非常重要的语法范畴,表现为在基础话语命题(表事实/陈述)之上补充或叠加说话者的主观意识,比如命令、要求、直觉、假设、可能、义务、怀疑、劝告、感叹等(Bybee & Fleischman 1995: 2)。情态标记的产生和发展反映了人类认知和语言使用的变化,因此情态研究历来是研究热点。DxCG涵盖了所有的语言演变现象,其对构式化和构式变化的界定和区分极大地推动了传统语法化研究的发展。本文集融合了语法化和构式语法理论,对跨语言和跨语系中的典型/非典型情态构式的形成(构式化)和发展(构式变化)进行详细刻画,并借助语料库和统计方法对结果进行可视化,不仅拓宽了情态和DxCG研究的范围,还推动研究向纵深发展。文集所体现的语法化和构式语法理论的兼容性以及对情态形成和变化的解释力使DxCG的分析模式进一步确立,有助于读者将其应用至汉语研究中。

文集的第一个特色在于突出了情态和DxCG结合研究的三个关键主题。(1)贯彻构式语法的网络节点观。第二、四、七和第八章对不同情态构式的节点、变体、联接关系等进行了细粒度分析,构建了完整的构式网络。(2)探索情态构式形-义发展的路径。既往类型学和语法化研究形成了大量可供验证的假设。第三、五和第六章展开了实证分析,发现相关假设不成立,提醒研究者不可一味重视共性、忽略特性。(3)区分构式化与构式变化。第四、五和第八章对构式化和构式变化进行了探究和解读,既有应用,也有批评。文集整体基于构式视角对情态的发展和变化进行了细致、深入的分析,为一些争论提供了解决思路,也对历史语言学领域的情态研究作出了贡献。

文集的第二个特色在于除了对典型情态构式的关注外,还讨论了多类非典型情态构式,比如:英语缩略式(第二章)、补语小句构式及后情态标记(第六章)、指示/陈述语构式(第七章)等,学界对这些构式的研究不足、了解不深入。文集搭建了情态研究的全局性网络,加深了研究者对情态范畴及其成员关系的理解,也将对其他领域(如语态、句型等)的研究有所启发。

文集的第三个特色在理论应用上,文集基于研究事实对既往研究进行了反思,

并力图突破研究瓶颈、为未来研究树立参照。比如,第三章证实同构项类型扩展不等于语法化程度增加,作者通过建模将语法化程度量化,值得借鉴。第五章中加利西亚语副词的演变历程与既定的单向性演变假说相冲突,例示了研究较少的“去交互主观化”现象,作者呼吁学界重新审视交互主观化的发展路径。第八章对日语名词补语标记演变为情态标记的历程进行了追溯,结合对同类语料和社会文化的分析,作者否定了此前的简单归类,并将研究结论的适用性扩展至其他日语构式,体现了作者的创新意识。

文集的第四个特色是重视真实语料的运用以及定性和定量分析的结合。DxCG关注构式的图式性、能产性和组构性。文集普遍采用了基于平衡语料库的统计建模和可视化分析,考察了不同情态构式的形态、句法、语义和语用特征及其之间的相关性,可视化地呈现了情态构式特性间的复杂关系,提高了研究的可复制性。

当然,本文集也存在不足。第四章作者使用了在第三章中解释力不足的“同构项类型扩展”分析。此外,第五和第六章中历时语料整体形符数较少。另外,文集没有收录对汉语情态范畴的研究,实为缺憾。但总体说来,瑕不掩瑜。目前DxCG和情态构式的结合研究尚处于起步阶段,随着DxCG理论的分析范式不断完善,对情态及相关范畴的研究也将进一步深化。

参考文献

- BYBEE J, FLEISCHMAN S. Modality in grammar and discourse: an introductory essay [C]//BYBEE J, FLEISCHMAN S. Modality in grammar and discourse. Amsterdam: John Benjamins, 1995: 1-14.
- HILPERT M. Ten lectures on diachronic construction grammar [M]. Leiden: Brill, 2021.
- KUTEVA T. Auxiliation: an enquiry into the nature of grammaticalization [M]. Oxford: Oxford University Press, 2001.
- TRAUGOTT E, DASHER R. Regularity in semantic change [M]. Cambridge: Cambridge University Press, 2001.

通信地址: 100191 北京市 北京航空航天大学外国语学院

English abstracts of major papers

Discourse meanings of Beijing Winter Olympic Games based on the LDA model

.....ZHANG Yu & WEI Naixing (1)

The LDA model is a popular unsupervised topical modeling method and can be applied to topic analysis in critical discourse analysis. Many previous studies have used keywords to discover topics, while few have adopted the topic modeling method. Therefore, the present study used the LDA model to examine the topic of Beijing Winter Olympics Corpus consisting of English news reports of Beijing Winter Olympic Games. In addition, this study adopted the corpus-drive approach to investigate the recurrent collocates and semantic prosodies of key words to reveal the discourse construction of Beijing Winter Games and the attitudinal meanings of oversea media. It was found that the media focused on three topics: the bid of the Winter Olympics, preparation of the Games, and government measures. The media positively evaluated the fact that Beijing had become the first “dual Olympic” city and that the Chinese government promoted winter sports. Meanwhile, oversea media criticized the lack of winter sports culture and natural snow in Beijing. This study proves the feasibility of combining LDA, the groups linguistic approach, and critical discourse analysis. It also provides a new analytical framework for future studies.

A study on the co-selection mechanism of the object and complement in English caused-resultative construction

.....LIU Congying, LI Xiaochen & CAO Duxin (15)

This paper investigates the co-selection mechanism of the object and adjective complement in English caused-resultative construction using the collostructional analysis method by considering the make+object+adjective complement construction as an example. It is found that the words attracted by the object and complement of this construction have distinct syntactic and semantic characteristics in the British National Corpus (BNC). Meanwhile, strong collocations between the object and the complement are also identified, proving that there are co-selection patterns between these two positions. These patterns result from the interactions of the construction’s grammatical and semantic features and its pragmatic functions. These findings indicate that the lexical

preferences of the construction's slots are hierarchical and probabilistic, which is the consequence of the interaction of the internal and external grammatical, semantic, and pragmatic factors.

A multivariate quantitative study on word order preference of English binomials from a variationist linguistics perspective

.....*LIN Yibing & MENG Qingnan (27)*

From a corpus-based variationist linguistics perspective, this study explores the impact of 10 constraints on word order preference of noun binomials and adjective binomials and the importance ranking by means of BNC corpus data and R software. It is found that the word order preferences of noun and adjective binomials are affected by different constraints. Number of syllables is the most powerful constraint of noun binomials, whereas final sonority is the most powerful constraint of adjective binomials.

A comparative sentiment analysis of CSR reports between multinational enterprises home and abroad

.....*SONG Tianyi & HUANG Libo (48)*

The study establishes a corpus of English corporate social responsibility reports (CSR Reports Corpus) of Huawei and BT Corporation by making a comparative sentiment analysis between the texts of two companies using LIWC-22 dictionaries. Critical discourse analysis will be used to explain the motivations behind the differences. The results of the study show that differences of emotional construction between the two companies are mainly reflected in the following aspects: 1) Huawei's CSR reports tend to shorten the emotional distance between the readers and the company by describing specific examples of events and to focus on the maintenance of long-term relationships with partners. BT's CSR reports show an obvious win-win attitude toward the external communities. 2) The sentiment tendency in BT's CSR reports is more neutral while Huawei's reports indicate a trend from more positive to a balance of positive and negative sentiment over the years. 3) BT's CSR reports describe events with the company's own unique characteristics and build a more multi-dimensional and comprehensive emotional relationship with the society. Huawei's reports reflect extensive emotional care to employees and clients during COVID-19 pandemic era. The paper attempts to provide some implications for the teaching of business English writing and provide suggestions to Chinese multinational companies regarding the construction of their corporate image overseas.

A comparative study of interactional metadiscourse and authorial identity construction in academic theses

.....*LIU Guobing & ZHANG Junlan (60)*

Based on Hyland's (2005a) interactional metadiscourse theory and Sun's (2015) identity categories, this study investigates the features of interactional metadiscourse and the similarities and differences on identity construction by employing interactional metadiscourse in academic theses in applied linguistics between Chinese masters and international journal authors. The research results are that, interactional metadiscourses employed by Chinese masters in academic writing are fewer than by international journal authors, except for engagement markers. In terms of frequency of occurrence, the identities constructed by the two author groups employing interactional metadiscourse are in the descending order of researcher, interactor, and evaluator. No difference is found in the identity of careful advisor constructed by Chinese masters and international journal authors. However, Chinese masters differ significantly from international journal authors in the constructed identities of self-initiated interactor, other-initiated interactor, self-evaluator, other-evaluator, cautious originator, and confident researcher. The research findings could provide beneficial enlightenments for academic English writing teaching.

Self-mention and authorial identity construction in empirical Chinese academic journal papers

.....*WANG Yamin, GONG Xue & AN Zhuoma (72)*

Based on the self-built corpus, this study investigates the relationship between self-mention forms and authorial identity construction, genre, and collocational verbs in empirical Chinese academic journal papers using correspondence analysis and multiple correspondence analysis. The study found that the forms of self-reference were mostly *we* (我们), *this study* (本研究), and *this paper* (本文), among which the abstract subjects *this paper* (本文) and *this study* (本研究) were mostly used. The former mainly assumes three low-risk roles (recounter of the research process, guide, architect) paired with research verbs, intentional verbs, cognitive verbs, and verbal verbs in the abstract and method sections. The latter tends to construct two high-risk roles (opinion-holder, originator) co-occurring with resulting verbs in discussion and conclusion sections. In contrast, the first-person plural *we* (我们) is versatile with different verbs in different sections of the papers. The authorial identity it constructs covers the above five types and the unique role of representative. This study provides a detailed description of the characteristics of the use of self-mention in academic Chinese discourse, and the results can be used as a reference for teaching academic Chinese writing.

语料库语言学

CORPUS LINGUISTICS

要 目

- | | |
|---------------------------|-------------|
| 基于 LDA 主题建模技术的北京冬奥会话语意义研究 | 张 毓 卫乃兴 |
| 英语“致使-动结”构式宾补共选机制研究 | 刘聪颖 李潇辰 曹笃鑫 |
| 变异语言学视角下英语并列二项式词序多元定量研究 | 林奕冰 孟庆楠 |
| 计算社会科学、文化组学与语言学 | 邵 斌 李雨飞 |
| 中外跨国公司英文社会责任报告情感分析对比 | 宋天祯 黄立波 |
| 学术论文中互动元话语及作者身份建构对比研究 | 刘国兵 张君兰 |
| CQP 语法赋能语言研究及语言学习 | 吴良平 |

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



记载人类文明
沟通世界文化
www.fltrp.com



北外学术期刊



iResearch 微信公众号

责任编辑：赵 雪
责任校对：夏洁媛
封面设计：锋尚设计



定价：35.00元