《中国学术期刊网络出版总库》、CNKI系列数据库及维普数据库入选期刊

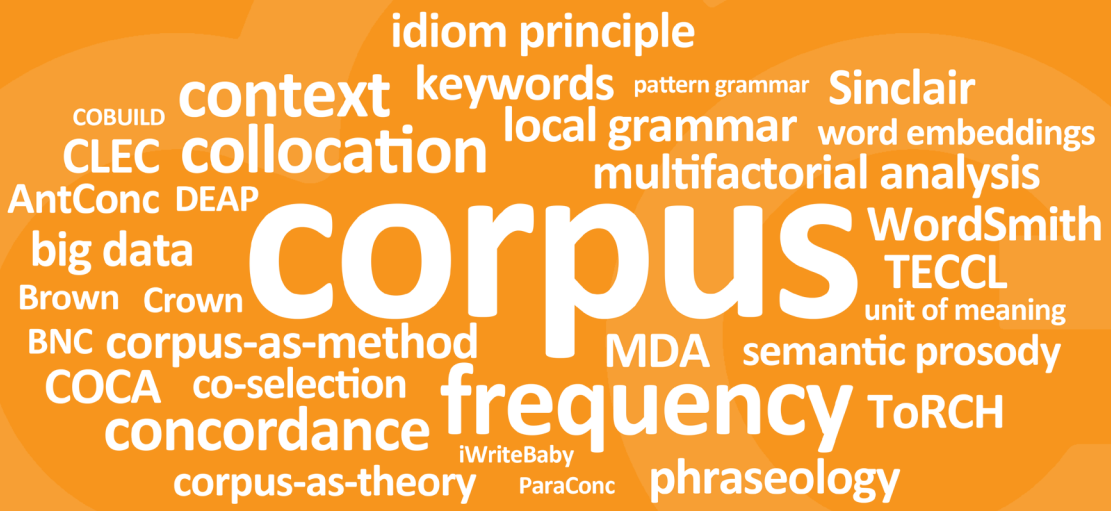
第18辑 2022

语料库语言学

CORPUS LINGUISTICS

第18辑
2022

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



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

外研社

语料库语言学

CORPUS LINGUISTICS

要 目

责任编辑：赵　雪
责任校对：孙凤兰
封面设计：锋尚设计

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

ISBN 978-7-5213-4090-7

定价：35.00元

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

电　　话：（010）88816828
电子邮箱：bfsucrg@sina.com
投稿网址：http://ylyy.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

语料库语言学

CORPUS LINGUISTICS

2022 年　第 18 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》
2022年　第9卷　第2期

目　录

语境共选

研究论文

研制开发

书刊评介

CORPUS LINGUISTICS

Volume 9, Number 2, 2022

Table of Contents

学习者语料库研究的多因素统计方法转向*

华南师范大学 李元科 何安平 黄灵敏

提要：国际上学习者语料库研究的统计方法正在经历变革，由单因素方法转向多因素统计建模的趋势十分显著，但国内学者对此关注不足。本文首先剖析单因素统计方法的短板，阐释多因素统计建模的优势所在。接着汇报基于我国英语高考作文库的两项研究：（1）运用多分类逻辑斯蒂回归建模法研究10项细颗粒句法指标协同影响成绩的机制；（2）运用结构方程模型研究目标语词块在COCA中的频率、搭配强度和准确性协同影响成绩的复杂路径关系。最后指出语料库研究者亟须提升统计素养，这既是研究语言系统复杂性的迫切需求，也是提升语言研究科学性的必要前提。本文对新文科背景下开展语言学的跨学科研究有一定启示作用。

关键词：学习者语料库、多分类逻辑斯蒂回归、结构方程模型

1 引言

起始于20世纪末的学习者语料库研究如今已发展到了新的阶段。新趋势之一是统计方法的革新：从单因素分析转向多因素建模。近年来，一些国际学者通过多因素统计建模法，研究学习者语言的多维度特征如何影响写作成绩，这正成为学习者语料库研究的新热点。然而，国内学者对这些重要转变的关注程度尚显不足。基于此，本文首先梳理国际上学习者语料库研究在统计方法方面的变革趋势，反思以往多用单因素方法有何短板，进而阐释多因素建模的优势，最后展示使用多因素建模法开展国内英语高考作文库的句法和词块研究案例，旨在阐明将多因素统计建模法应用于我国英语学习者语料库研究的可行性和教研价值，以期推动其在国内语料库研究中得到更广泛的应用。

* 本文系国家社科基金一般项目“高中英语学习者词块显隐特征协同影响写作的机理研究”（21BYY187）的阶段性成果。本文通讯作者为李元科。
作者贡献：
李元科：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；
何安平：讨论结论、修改润色；
黄灵敏：数据收集、数据分析。

2 学习者语料库研究的统计方法变革

据Paquot & Plonsky（2017）发表的评估学习者语料库研究统计方法和研究质量的元分析（meta-analysis）报告，在1991—2015年发表于60多种国际期刊和146本编著的共计378项学习者语料库研究中，使用单因素统计方法的研究所占比例高达84%，依次为卡方（23%）、t检验（20%）、相关分析（17%）、方差分析（14%）和对数似然检验（10%）；而使用多因素统计方法检验不同自变量协同影响因变量的研究所占比例却非常低，其中回归分析仅占6%，聚类分析和判别分析各仅占2%（Paquot & Plonsky 2017：79）。报告指出，研究者普遍存在过分倚重单因素统计方法，缺乏查验结果是否满足统计假设，以及缺乏报告效应大小等问题，这些在一定程度上体现了研究者“统计素养的缺失”（lack of statistical literacy）（Paquot & Plonsky 2017：67）。

研究者过分倚重单因素统计方法有可能是因为以往研究大多采用“中介语对比分析”（Granger 1996）的范式，即与英语本族语者相比，学习者对某种语言特征是否存在过多使用、过少使用和错误使用的问题，故研究者较多通过单因素统计方法检验不同群组的使用是否存在统计学上的显著差异。一些学者批评这类研究仅停留在浅层描述层面（Gries 2015b；Paquot & Plonsky 2017；Evert 2018；许家金 2020）。

单因素统计方法的局限性与其方法本身的短板不无关系。单因素统计方法的短板之一是无法检验不同自变量如何协同影响因变量。例如，以往较多采用的相关分析虽然可以获得各自变量与因变量之间的相关系数和显著性，或者用方差分析虽然可以检验各自变量在不同水平是否有显著差异，但是所得到的都是每个自变量单独影响因变量的情况，无法得知多个自变量协同影响因变量的总体效应。然而，学习者语言的不同子系统（如词汇、短语、句法、衔接等）之间的协同性会影响语言水平（Norris & Ortega 2009；Verspoor *et al.* 2012），且各子系统内部的不同维度特征之间同样存在协同性或竞争性（郑咏滟、冯予力 2017；郑咏滟 2018）。探究这些复杂关系必须使用多因素统计方法，才能深入剖析学习者语言的多维度特征协同影响语言能力的复杂机制（Gries & Wulff 2013；Gries 2015b）。短板之二是单因素方法和一些多因素方法都无法对自变量是通过何种路径对因变量产生影响开展分析，即它们无法揭示自变量是直接影响因变量，抑或是通过影响某些中介而间接影响到因变量。例如，近年来，一些国际学者使用自然语言处理工具（如CollGram、TAALES），将学生英语写作中的N词组合（n-gram）与英国国家语料库（BNC）或当代美国英语语料库（COCA）中的词块进行比对，从而得知学生使用了哪些英语本族语的词块，并通过这些工具测量出写作者所用英语本族语词块的比率和它们在BNC或COCA中的频率均值和搭配强度均值等，最后探究它们与写作成绩之间有何关联。研究表明：写作者使用COCA词块的比率越高则分数越高（Bestgen & Granger 2014）；词块在COCA中的搭配强度MI均值越高分数也越高

（Bestgen & Granger 2014；Granger & Bestgen 2014；Kim *et al.* 2018；Paquot 2018；Garner *et al.* 2019）；托福独立写作的两词词块在BNC笔语库的频率均值越高则分数也越高（Kyle & Crossley 2016）。但由于这些研究采用的统计方法多为相关分析、方差分析、线性或逻辑斯蒂回归分析，缺乏路径分析，因此仅仅得知这些指标对写作成绩有影响，却无法解释它们是通过怎样的路径对成绩产生了影响。

针对以上短板，国内外不断有学者呼吁学习者语料库研究亟须变革统计方法。例如，Gries（2015a：71）认为，语料库驱动的研究以频次和分布等概率数据为基础，鉴于数据的复杂性，若想在语料库数据分析方面取得重大进展，必须致力于提升统计方法。而多因素统计方法是推动（学习者）语料库研究继续向前发展的“最重要建议”（Paquot & Plonsky 2017：85），因为“多因素统计方法既是语言描写技术，也有很强的理论解释力”（许家金 2020：1）。

纵观近五年国际上学习者语料库研究可发现，使用多因素统计方法渐成趋势。例如，Kim *et al.*（2021）运用结构方程模型研究写作者的认知特性（注意力、工作记忆）以及写作过程中的停顿行为（停顿的平均时长、相邻两次停顿之间的平均词长）与文本特征（作文平均长度、成绩）之间的复杂路径关系。结果显示，虽然写作者的注意力对作文成绩无直接影响，但是注意力通过影响写作过程中“停顿的平均时长”而对作文成绩产生的间接影响却达到了显著水平。换言之，注意力与作文成绩是间接相关，而写作停顿的平均时长是建立它们之间关系的中介。又如，Duan & Shi（2021）运用混合效应模型（mixed-effects modeling）研究学生学术英语写作能力发展与所用词块的搭配强度、结构、功能和语际一致性（congruency）有何关联。通过对31名中国大学英语学习者开展两年半的跟踪，该研究发现学习者所用词块的搭配强度、结构以及两者的交互对推动学术英语写作能力的发展起了重要作用，但词块的功能和语际一致性对写作能力发展的影响并不显著。

除了以上方法，多因素统计方法还包括多元线性回归、逻辑斯蒂回归、聚类分析、随机森林等，它们助力语料库研究在描写深度、预测语言选择机制及揭示多维度特征协同影响语言能力的复杂性等方面都有所突破和创新（许家金 2020）。可见，国际上学习者语料库研究正在转向使用多因素统计建模法，它们正逐渐取代“中介语对比分析法”而成为一种新范式。但目前在国内类似研究还不多见，只有Zhang & Li（2021）、张懂（2019，2020）等。在下一节笔者会展示将多因素统计建模法应用于高考英语作文库的两个研究实例。第一例使用多分类逻辑斯蒂回归（multinomial logistic regression，简称MLR）探究多种句法细颗粒复杂度指标如何协同影响作文成绩；第二例使用结构方程模型（structural equation modeling，简称SEM）研究学生使用目标语词块的频率、搭配强度和准确性协同影响作文成绩的复杂路径关系。

3　运用多因素统计建模法开展高考英语作文库研究的两个实例

下面介绍笔者基于英语高考请求信作文库（简称作文库）开展的句法和词块的两项研究。该作文库（见表1）总库容是120,871词次，它包含2016年广东省900名高三毕业生在英语高考中撰写的求助外教修改个人简历的请求信。他们的成绩介于11—25分，分别属于五个分值等级中的三档（11—15分）、四档（16—20分）和五档（21—25分）。各分值样本数均为60篇。

表1　作文库情况简介

分档	请求信数量	作文长度（形符）			
		最小值	最大值	均值	标准差
五档	300	96	229	149.03	27.539
四档	300	89	217	129.87	25.960
三档	300	72	208	117.92	27.205

3.1　使用MLR研究多种句法特征协同影响作文成绩的机理

MLR是处理因变量为多分类（即至少有三个水平）变量的一种回归分析方法（温忠麟 2016：246），它不但可以检验出对因变量有显著影响的若干个自变量，还可以对它们协同预测因变量模型的效果展开评估。

考察的10项句法特征源自高中英语课程标准（中华人民共和国教育部 2020：181-183）要求掌握的四种从句句型（状语从句、宾语从句、限定性定语从句和非限定性定语从句）和六种名词性短语结构（形容词+名词、所有格+名词、名词+名词、名词+动词分词形式、名词+介词短语、动词名物化）。笔者首先使用自然语言处理技术（TAASSC）（Kyle 2016）将这些句法特征在每篇作文中的比率自动提取出来[1]。然后用EXCEL将它们在900篇作文中的比率（简称指标）汇总。接着在SPSS平台上使用皮尔逊相关分析，对10项指标之间的相关系数展开检验，结果显示任意两项指标之间都不存在共线性问题（$r < 0.7$）（见附录1）。因此继续在SPSS平台上使用MLR，将10项句法指标作为自变量，将作文的分档作为因变量，通过“逐步向后删除法”（stepwise backward elimination），将显著影响作文分档的指标提取出来，结果见表2。

表2　显著影响作文分档的句法指标

指标	似然比检验		
	卡方	自由度	显著性（P）
状语从句比率	12.743	2	0.002
动词名物化比率	26.620	2	0.000
形容词+名词比率	34.952	2	0.000
名词+名词比率	8.285	2	0.016
名词+介词短语比率	9.439	2	0.009

表2的似然比检验结果显示，状语从句、动词名物化、形容词+名词、名词+名词、名词+介词短语这五项结构的比率显著影响作文的分数档次（P值均小于0.05）。而其余五项句法特征（宾语从句、限定性定语从句、非限定性定语从句、所有格+名词、名词+动词分词）的比率则对作文分档未有影响（P值均大于0.05）。为检验表2中五项指标构建的预测作文分档模型（简称预测模型）的效度，笔者进一步用MLR产出四项评估结果，分别为此模型与仅截距模型（intercept-only model）的似然比检验、拟合优度、效应量和预测准确性，结果见表3。

表3　评估预测模型效度的四项结果

似然比检验			
模型	卡方	自由度	显著性
仅截距模型			
预测模型	142.134	10	0.000

拟合优度			
	卡方	自由度	显著性
皮尔逊	528.297	588	0.963
偏差	517.033	588	0.984

效应量（伪R方）	
考克斯–斯奈尔	0.377
内戈尔科	0.425
麦克法登	0.216

预测准确性				
实际分档	预测分档			
	三档	四档	五档	正确百分比
三档	216	57	27	72.0%
四档	102	99	99	33.0%
五档	27	96	177	59.0%

表3显示，预测模型在四项效度评估中都达到标准。首先，表3中的似然比检验结果显示，与仅截距模型（即不包含表2中五项显著指标的模型）相比，预测模型显著提升了准确预测作文分档的效果（P = 0.000 < 0.05）。接着来看表3中的拟合优度，皮尔逊卡方的显著性值（0.963）和偏差卡方的显著性值（0.984）都大于0.05，表示用此模型来预测作文分档的结果与它们的实际分档在总体上并无差异，可得知“模型的总体预测效果理想”（Petrucci 2009：200）。再看表3中的效应量，依据“麦克法登效应量介于0.2和0.4之间是评估模型能够较好解释因变量方差的基准”（Petrucci 2009：200），此效应量（0.216）达标。最后，看表3中的预测准确性，与基准比率（33%）[2]相比，虽然此模型预测四档作文的准确率（33%）一般，但是它准确预测三档作文的比率（72%）和五档作文的比率（59%）都远超过基准。换言之，表2里五项句法特征对预测三档和五档请求信作文有较高的准确率。

在用MLR建立数据模型之后，笔者在三档和五档作文里对表2中五项预测性句法特征开展批量提取和质性分析，发现它们在不同水平作文之间的比率差异与写作者的语用得体性和词汇能力有协同关联。以表2第一行的状语从句为例，鉴于之前TAASSC的分析结果显示，它在五档作文的比率是三档的2.5倍，笔者手动提取了这两个分档作文中的所有状语从句，发现它们的差异主要归因于if-状语从句的使用率：if-状语从句在五档的频次是三档的1.73倍。进一步观察五档作文里的401例if-状语从句，有70%以上（287例）出现在I+will/would+appreciate+it+if-状语从句句型中，如I will appreciate it if you can do me a favor，表达“间接请求”（Economidou-Kogetsidis 2011）。反观三档作文，此类if-状语从句仅有2例；而表达“直接请求”（Petrucci 2009：200）的语例却很多，例如：I hope that you can help me有33例，I hope that you can give me a hand有22例，I want you to help me有19例，I need you to help me有17例。由此可见，五档作文者多用更为礼貌得体的间接请求类句式，他们的语用能力明显优于三档作文者。又例如，表2第二行的动词名物化在五档作文中的比率是三档的4.8倍，两者的差异主要归因于该话题作文关键词apply的名物化形式（application）的使用率。Application在五档作文中的频次（412次）是三档（79次）的5.2倍。对apply进行名物化后形成的正确搭配application forms在五档作文中有24例；在三档作文中却无一例。相反，三档作文含有大量把apply误用为修饰语的错误搭配，如apply book（38例）、apply letter（12例）、apply forms（8例）、apply paper（5例），可见动词名物化及其搭配能够映射写作者运用词汇的水平。

此项研究展示了基于语料库的语言特征分析与MLR建模深度融合的优势。首先是使用TAASSC自动测量出细颗粒句法特征在三个分档作文中的比率。其次是使用MLR检验得知有五项句法特征显著影响了作文成绩，并利用MLR验证了它们预测作文成绩模型的效度达标。最后将这五项预测性句法特征带回到文本里开

展质性分析。结果表明，不同水平学习者使用这些句法结构与他们的语用能力和词汇能力之间具有协同性。

3.2　使用SEM研究目标语词块的频率、搭配强度和准确性协同影响写作成绩的复杂路径关系

SEM综合了因子分析、回归分析和路径分析（许宏晨 2019），它不但能够测量自变量直接影响因变量的效应，还能够测量自变量通过影响某个（些）中介变量间接影响因变量的效应（即检验中介作用）。在本例中笔者使用SEM研究目标语四词词块的频率、搭配强度和准确性对写作成绩的影响。由于本研究具有探索性，故将目标语词块的频率作为自变量，将写作成绩作为因变量，将搭配强度和准确性作为中介变量，做出四条路径假设（见图1）:（1）目标语四词词块的频率直接影响写作成绩的效应是怎样的？（2）目标语四词词块的频率通过影响搭配强度间接影响写作成绩的效应是怎样的？（3）目标语四词词块的频率通过影响准确性间接影响写作成绩的效应是怎样的？（4）目标语四词词块的频率依次通过影响搭配强度和准确性从而影响写作成绩的效应是怎样的？通过SEM测量出不同路径的效应大小和显著性来验证最理想的路径。

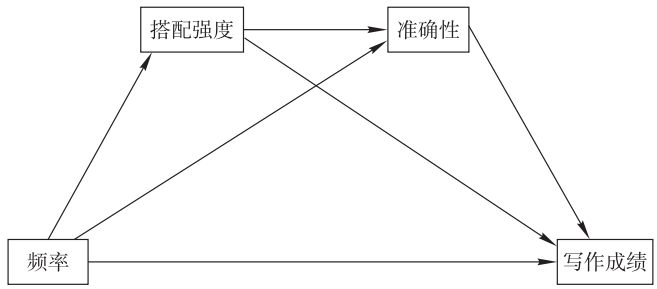


图1　词块的频率、搭配强度和准确性影响作文成绩的路径模型

首先使用CollGram（Bestgen & Granger 2014）从作文库（见表1）的每篇作文中提取出所有连续性的四词组合（如if you can help），然后通过此工具将它们与当代美国英语语料库（COCA）中的四词词块进行比对，从而得知写作者使用了哪些目标语词块。接着使用CollGram测定每篇作文所有四词组合中目标语所占比率，它表征写作者准确使用四词词块的程度（简称准确性）。接下来通过CollGram测量出每篇作文所用目标语四词词块（类符）在COCA总库中的频率均值（简称目标语词块的频率）和MI^3均值（简称搭配强度）。最后，笔者将目标语词块的频率、搭配强度和准确性的结果导入SEM分析工具Process 3.5（Hayes 2018），对上述四条路径的效应开展分析，结果见表4。

表4　目标语词块的频率通过四条路径影响写作成绩的效应

路径	效应	标准误	显著性
（1）目标语词块的频率→写作成绩	–2.682	1.367	P > 0.05
（2）目标语词块的频率→搭配强度→写作成绩	2.397	0.652	P < 0.05
（3）目标语词块的频率→准确性→写作成绩	0.128	0.37	P > 0.05
（4）目标语词块的频率→搭配强度→准确性→写作成绩	5.732	0.767	P < 0.05

如表4所示，路径（1）未达到显著水平（P > 0.05），表明目标语词块的频率对写作成绩无直接影响。路径（3）也未达到显著水平（P > 0.05），表明目标语词块的频率通过影响准确性而间接影响写作成绩也不成立。路径（2）和（4）都达到了显著水平（P < 0.05），而且后者的效应（5.732）是前者（2.397）的两倍多，可见路径（4）与语料库数据的拟合最佳，表明目标语词块的频率需要通过影响搭配强度和准确性从而影响写作成绩。

语言学理论和二语研究结论都能为路径（4）提供“学科理论的解释”（温忠麟、刘红云 2020：95）。首先，基于使用的语言学理论认为，不同语言成分在一起共现的频次是影响它们之间搭配强度的关键因素（Gries & Ellis 2015：231；蔡金亭、陈家宜 2019：5），在词块层面，它体现为若干个单词在一起共现所组成词块的频次会直接影响到它们之间的搭配强度。MI作为一种搭配强度的算法，语料库研究者发现高频词的搭配通常被赋予的MI值较低，而低频词的搭配被赋予的MI值较高（Gries & Ellis 2015：237）。其次，二语认知加工研究表明，词语之间的搭配强度与可预测性有密切关联（Durrant & Doherty 2010），搭配强度越高的词块在通过其内部的前面单词来预测后面单词的准确率就会越高（Jiang & Nekrasova 2007；Ellis *et al.* 2008：388），而且高强度搭配词块还能作为一个整体来吸纳、储存、记忆和输出，从而提升语言产出的正确性（Wray 2002：72）。最后，词汇搭配的准确性正向影响写作成绩也已经被证实（Bestgen & Granger 2014；Crossley *et al.* 2014）。

此项研究展示了将SEM应用于语料库语言特征分析的优势。一是使用它开展路径分析揭示了词块的多维度特征是通过怎样的路径关系影响到写作成绩，它突破了以往研究存在的偏重描写但对深层机理探索不足的局限性。二是研究者可将路径分析的结果与语言学理论和二语研究的结论开展互鉴，这有助于对学生语言的多维度特征协同影响写作成绩的复杂机理作出更加科学的诠释。

4　结语

本文通过文献综述和案例分析，阐述了融合自然语言处理技术和高阶统计深挖句法和词块的多维度特征协同影响写作成绩的理论意义和实用价值，这对新文科背景下开展语言学跨学科研究有两点启示。第一，学习者语料库研究的复杂性不亚于心理学和认知科学等其它学科的研究，语料库研究者同样需要借鉴相关计量技术。这既是语言系统复杂性所决定的，也是提升语言研究科学性所必需的（Gries 2015a：50；Gries 2015b：173）。第二，语料库研究的新趋势是与高级统计相融合，逐步形成一种新的数据驱动路径（Hunston 2017）：有别于传统语料库统计是从某些词语切入，新范式是从数据切入，通过建立数据模型来揭示语料库更深层次的共选和关联本质。诚然，新旧融合的研究范式应该是“计量分析+可视化+人类解释”形成的三位一体，而不是没有任何解释力的数据模型（Evert 2018）。研究者在建立数据模型之后，有必要回归文本作质性分析，或者将模型结果与语言学理论和实证研究结论开展比对，这不但有助于对数据模型作出具有理论支撑的诠释，而且还能对现有理论进行补充、修正或完善，以提升其科学性和在语言教学和理论研究中的实用价值。

注释

1　由于非限定性定语从句的比率和动词名物化的比率未能由TAASSC提供，故通过手动分析获得。前者采用与TAASSC一致的从句比率计算公式，即每种从句数量除以所有小句的总数；后者采用与TAASSC一致的名词短语结构比率计算公式，即每种名词短语结构的数量除以所有名词短语的总数。

2　由于此作文库里三个分档的作文数量均等，因此每个分档用于评估预测准确性的基准比率都是33%（100%/3）。

3　MI是Mutual Information（互信息）的缩写，它是测量搭配强度的一种算法。MI值越高，则搭配强度越强。

参考文献

BESTGEN Y, GRANGER S. Quantifying the development of phraseological competence in L2 English writing: an automated approach [J]. Journal of Second Language Writing, 2014, 26(1): 28-41.

CROSSLEY S, SALSBURY T, MCNAMARA D. Assessing lexical proficiency using analytic ratings: a case for collocation accuracy [J]. Applied Linguistics, 2014, 36(5): 570-590.

DUAN S, SHI Z. A longitudinal study of formulaic sequence use in second language writing: complex dynamic systems perspective [J]. Language Teaching Research,

2021: 1-34.

DURRANT P, DOHERTY A. Are high-frequency collocations psychologically real? Investigating the thesis of collocational priming [J]. Corpus Linguistics and Linguistic Theory, 2010, 6(2): 125-155.

ECONOMIDOU-KOGETSIDIS M. “Please answer me as soon as possible”: pragmatic failure in non-native speakers’ e-mail requests to faculty [J]. Journal of Pragmatics, 2011, 43(13): 3193-3215.

ELLIS N, SIMPSON-VLACH R, MAYNARD C. Formulaic language in native and second language speakers: psycholinguistics, corpus linguistics and TESOL [J]. TESOL Quarterly, 2008, 42(3): 375-396.

EVERT S. The hermeneutic cyborg: Sinclair Lecture at Birmingham University [EB/OL]. (2018-06-25) [2022-07-15]. https://www.birmingham.ac.uk/research/activity/corpus/news/2018/sinclair-lecture-2018.aspx.

GARNER J, CROSSLEY S, KYLE K. N-gram measures and L2 writing proficiency [J]. System, 2019, 80: 176-187.

GRANGER S. From CA to CIA and back: an integrated approach to computerized bilingual and learner corpora [C]//AIJMER K, ALTENBERG B, JOHANSSON M. Languages in contrast: text-based cross-linguistic studies. Lund: Lund University Press, 1996: 37-51.

GRANGER S, BESTGEN Y. The use of collocations by intermediate vs. advanced non-native writers: a bigram-based study [J]. International Review of Applied Linguistics, 2014, 52(3): 229-252.

GRIES S. Quantitative designs and statistical techniques [C]//BIBER D, REPPEN R. The Cambridge handbook of English corpus linguistics. Cambridge: Cambridge University Press, 2015a: 50-71.

GRIES S. Statistics for learner corpus research [C]//GRANGER S, GILQUIN G, MEUNIER F. The Cambridge handbook of learner corpus research. Cambridge: Cambridge University Press, 2015b: 159-181.

GRIES S, ELLIS N. Statistical measures for usage-based linguistics [J]. Language Learning, 2015, 65(S1): 228-255.

GRIES S, WULFF S. The genitive alternation in Chinese and German ESL learners: towards a multifactorial notion of context in learner corpus research [J]. International Journal of Corpus Linguistics, 2013, 18(3): 327-356.

HAYES A. Introduction to mediation, moderation and conditional process analysis (2nd edition) [M]. New York: The Guilford Process, 2018.

HUNSTON S. Corpus Linguistics in 2017: a personal view [EB/OL]. (2017-07-24) [2022-12-29] http://www.birmingham.ac.uk/cl2017.

JIANG N, NEKRASOVA T. The processing of formulaic sequences by second language speakers [J]. The Modern Language Journal, 2007, 91(3): 433-445.

KIM M, CROSSLEY S, KYLE K. Lexical sophistication as a multidimensional phenomenon: relations to second language lexical proficiency, development, and writing quality [J]. The Modern Language Journal, 2018, 102(1): 120-141.

KIM M, TIAN Y, CROSSLEY S. Exploring the relationships among cognitive and linguistic resources, writing processes, and written products in second language writing [J]. Journal of Second Language Writing, 2021, 53(3): 1-15.

KYLE K. Measuring syntactic development in L2 writing: fine grained indices of syntactic complexity and usage-based indices of syntactic sophistication [D]. Atlanta, GA.: Georgia State University, 2016.

KYLE K, CROSSLEY S. The relationship between lexical sophistication and independent and source-based writing [J]. Journal of Second Language Writing, 2016, 34: 12-24.

NORRIS J, ORTEGA L. Towards an organic approach to investigating CAF in instructed SLA: the case of complexity [J]. Applied Linguistics, 2009, 30(4): 555-578.

PAQUOT M. Phraseological competence: a missing component in university entrance language tests? Insights from a study of EFL learners' use of statistical collocations [J]. Language Assessment Quarterly, 2018, 15(1): 29-43.

PAQUOT M, PLONSKY L. Quantitative research methods and study quality in learner corpus research [J]. International Journal of Learner Corpus Research, 2017, 3(1): 61-94.

PETRUCCI J. A primer for social worker researchers on how to conduct a multinomial logistic regression [J]. Journal of Social Service Research, 2009, 35(2): 193-205.

VERSPOOR M, SCHMID M, XU X. A dynamic usage-based perspective on L2 writing [J]. Journal of Second Language Writing, 2012, 21(3): 239-263.

WRAY A. Formulaic language and the lexicon [M]. Cambridge: Cambridge University Press, 2002.

ZHANG X, LI W. Effects of n-grams on the rated L2 writing quality of expository essays: a conceptual replication and extension [J]. System, 2021, 97(2): 1-14.

蔡金亭，陈家宜. 基于使用取向的英语动词论元构式二语研究[J]. 北京第二外国语学院学报，2019（5）：4-16.

温忠麟. 心理与教育统计：第二版[M]. 广州：广东教育出版社，2016.

温忠麟，刘红云. 中介效应和调节效应方法及应用[M]. 北京：教育科学出版社，2020.

许宏晨. 第二语言研究中的结构方程模型案例分析[M]. 北京：外语教学与研究出

版社，2019.

许家金. 多因素语境共选：语料库语言学新进展[J]. 外语与外语教学，2020（3）：1-10.

张懂. 语料库量化方法在构式语法研究中的应用[J]. 现代外语，2019（1）：134-145.

张懂. 英语与格构式变体的多变量分析及其心理现实性研究[J]. 外语教学与研究，2020（1）：40-52.

郑咏滟. 高水平学习者语言复杂度的多维发展研究[J]. 外语教学与研究，2018（2）：218-229.

郑咏滟，冯予力. 学习者句法与词汇复杂性发展的动态系统研究[J]. 现代外语，2017（1）：57-68.

中华人民共和国教育部. 普通高中英语课程标准（2017年版2020年修订）[M]. 北京：人民教育出版社，2020.

通信地址：510631　广东省广州市　华南师范大学外国语言文化学院

附录1　第一实例中10项句法指标的皮尔逊相关分析结果

	adv	ccomp	nonrac	nominalization	amod	poss	nn	vmod	prep	rcmod
adv	1	−.193**	.125*	.244**	.215**	.085	−.081	.021	.159**	−.058
ccomp	−.193**	1	−.038	−.166**	−.314**	−.154**	−.038	−.100	−.280**	−.080
nonrac	.125*	−.038	1	.161**	.084	.061	−.082	.153**	.130*	.069
nominalization	.244**	−.166**	.161**	1	.309**	.241**	−.071	.071	.247**	.072
amod	.215**	−.314**	.084	.309**	1	.275**	.044	.158**	.274**	.110
poss	.085	−.154**	.061	.241**	.275**	1	−.053	.118*	.085	.011
nn	−.081	−.038	−.082	−.071	.044	−.053	1	.033	−.061	−.010
vmod	.021	−.100	.153**	.071	.158**	.118*	.033	1	.011	−.015
prep	.159**	−.280**	.130*	.247**	.274**	.085	−.061	.011	1	−.049
rcmod	−.058	−.080	.069	.072	.110	.011	−.010	−.015	−.049	1

注：** 表示在0.01水平上显著相关；* 表示在0.05水平上显著相关。

adv=状语从句的比率，ccomp=宾语从句的比率，nonrac=非限定性定语从句的比率，nominalization=动词名物化的比率，amod=形容词+名词的比率，poss=所有格+名词的比率，nn=名词+名词的比率，vmod=名词+动词分词形式的比率，prep=名词+介词短语的比率，rcmod=限定性定语从句的比率。

变异语言学视角下英语情态构式多元定量研究*

大连海事大学 李思雨 戴雅宁 孟庆楠

提要：本研究采用基于语料库的变异语言学视角，借助COHA语料库，对影响must、have to、have got to三种构式变体选择的主要因素及其在美式英语中的历时演化路径进行定量研究。结果表明：在1810—2009年，美式英语在表达“必须”这一含义时，出现了must逐渐被have to取代的趋势；影响must、have to、have got to三种变体选择的主要因素按照重要性排序依次为“时态”“体裁”以及“年代”。此外，通过进一步分析三种变体的演化规律，本研究认为英语中存在主要情态动词逐渐被半情态动词[1]取代的趋势。情态动词must的部分含义正逐渐由半情态动词have to表达，而半情态动词have got to的使用频率却呈下降趋势。

关键词：主要情态动词、半情态动词、多元定量研究、构式交替

1 引言

语言是联系社会成员的工具，人们通过语言传达思想、表达情感，由此社会才成为一个联系紧密的整体。然而人的交往活动并不是按照一成不变的规则进行的，沟通方式需要根据环境、人群、主题等进行适当、灵活的调整。因此交际时，为了达到某种目的、传达某种意愿、表达某种情感，说话人常常需要根据情况改变语言方式，从而使沟通更加顺畅有效。情态系统就集中体现了这种语言的变化方式，成为许多语言学家、哲学家及逻辑学家长期研究的课题。

情态是一个语义概念，可以由多种语法类别来实现，并且存在“一义多形”的现象。例如：should与ought to、shall与be destined to、will与be willing to等，每组中的词具有相似的语义及语用功能，可称之为“构式交替”现象。Perek（2015）指出，如果仅关注构式本身，而非探索构式之间的关系，就无法处理构式交替现象。

* 本研究是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构式交替现象研究”（L21CYY004）的阶段性成果。孟庆楠为本文通讯作者。
作者贡献：
李思雨：数据收集、数据分析、讨论结论、初稿撰写、字数占比（40%）；
戴雅宁：数据收集、数据分析、讨论结论、初稿撰写、字数占比（40%）；
孟庆楠：选题构思、研究方法、字数占比（20%）、修改润色。

Gries（2003）提出每个构式都存在基于自身功能而形成的范畴，不同构式范畴有重合的可能，但事实上差异多于共性。结合语言“经济性”原则，笔者认为两个或多个具有相似语义、语用功能的构式并不完全相同。语言使用者对几个相似构式的选择会受到语言内外部因素的限制，例如主语的有生性、时态、小句的取向、体裁、年代、说话人职业、社会阶级和性别等（孟庆楠、罗卫华 2020）。

语言学家对情态的研究取得了丰硕成果，国外学者的研究主要涉及以下三个方面：对情态系统整体进行研究（Depraetere & Reed 2006）；从语义及语用层面对语义相近的情态动词加以区分（Cappelle *et al.* 2019）；分析近代英式英语中表达义务、认识的情态动词及半情态动词的变化规律（Smith 2003）。但这些研究缺乏对影响情态动词选择的句法层面因素的讨论，且缺乏大量可靠数据支持。国内学者则侧重于英语情态语用方面的研究，如与英语教学紧密结合的英语情态研究（高秋萍 2009）、利用情态理论系统分析法庭上的交叉质询（王振华 2004）等，但缺乏对情态系统历时演化路径的深入探讨。

2　研究背景与研究问题

英语情态系统不断变化，主要情态动词的使用频数不断减少、意义趋向单一化；而半情态动词的使用频数日益增加，因此主要情态动词面临被半情态动词取代的风险（Krug 2000；Leech 2003）。关于must与have to的区别，很多研究者从语义、语用层面进行分析，得出must与have to都表示“必须，必要”，但是must通常为“内在驱动”，即说话人本身的主观需要；have to通常为“客观必须”，及客观因素促使“不得不做”某事。同时，Close & Aarts（2010）利用当代英语口语历时语料库（DCPSE），通过语义标注，将情态动词的语义分为认识情态和义务情态，随后进行频数的统计和对比分析，得出情态动词must的使用频数减少与半情态动词have to的使用频数增加存在一定关联。这为研究情态系统历时变化提供了窗口，为后来研究者探究情态动词使用频数减少是否与半情态动词使用频数增加有必然联系提供了参考。但该论文得出的结论是基于当代英语口语语料库的，此结论在书面语中是否依然成立不得而知。金婷茹（2018）对语料进行层次聚类分析，从而区分must和have to，但是侧重于对比中国英语学习者与本族语学习者在学习、运用情态表达时存在哪些不同，同时忽略了对have got to的研究，也缺乏对情态系统历时变化的深入探讨。

基于以上背景，笔者将借助离线版美式英语历时语料库（COHA），重点探讨以下三个问题：（1）must、have to及have got to三种构式变体在美式英语中具体呈现怎样的分布情况及历时演化规律？（2）对must、have to及have got to三种构式变体选择影响显著的因素有哪些，重要程度如何？（3）以must为代表的主要情态动词是否正在逐渐被以have to为代表的半情态动词所取代？

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研究范式。该研究范式的基本假设为：语言使用者所具有的内在语法知识是动态的、概率化的，语言的变异现象也具有概率属性（Bod *et al.* 2003）。由“概率语法观”（Bresnan 2007）可知，语言使用者在选择不同构式变体的时候，会被诸多语言学内、外部因素所影响。本族语使用者，或许可以在不了解must、have to及have got to细微差别的情况下，在特定的语境中正确选择、使用这三种构式变体，但却无法道出这三种构式变体之间的细微差别。大型历时语料库为探索制约三种构式变体使用的因素提供了帮助。

基于上述理论框架，本研究采用了多因素分析的研究方法。与传统的语言使用频数的描述性方法不同的是，这种新式研究方法会随机抽取适量语料库数据，随后推测出影响构式交替的因素，从而进行人工标注。利用统计学软件，基于回归分析和分类模型等探索性方法，从多个变量中筛选出对特定构式变体选择影响最为显著的因素，而非把统计各种语言变体形符频数作为最关键的研究因素。最后，结合相关语言学理论，解读其蕴含的功能根据。多因素分析比单因素分析、本族语使用者内省式分析等研究方法更加客观、科学，不仅能全面细致地刻画语言现象，还能考察语言特征，并探索这些语言特征与语境变量是如何相互影响的（许家金 2020）。

4 语料来源与标注

为了探究19—20世纪美式英语中have to、have got to、must之间的细微差别，并快速提取相关语料，本研究选用离线版本的COHA语料库。COHA语料库是一个包含约4亿词的大型历时语料库，涵盖了小说、杂志、报纸和非虚构四种体裁，涵盖题材广且具有代表性，能够较为全面地反映美式英语在19—20世纪的变化特征。该语料库中的语料已进行了词形还原及词性标注，便于检索、提取、分析含有关键词及其搭配词的语料。本研究所涉及的反应变量为表达“必须”“不得不”的义务情态动词及半情态动词，即：must、have to及have got to。研究仅关注表达义务的情态动词must和半情态动词 have to、have got to，对于must作名词、must表示推测（如 must have done、must be+名词等）及have to、have got to中的have是实义动词等情况均不讨论。

首先，笔者通过 Perl 编程，对语料进行了初步筛选，发现符合本研究要求的含有must、have to、have got to的语料分别有400,000余条、190,000余条、1,800余条。出于可操作性和含有must和have to的符合条件的语料比例大致为2∶1的情况，笔者按照2∶1的比例分别抽取含有must和have to的语料。抽取0.4%的包含must的语料（包含陈述句、疑问句），得到包含1,600余条语料的原始数据样本；按0.2%

抽取含有have to的语料（包含陈述句、疑问句），得到380条原始语料。考虑到含有have got to的语料总数较少，为了对其进行详细的研究，笔者按12%抽取，得到216条原始语料。手动标注部分变量时，笔者删除了表示认识情态的must以及have是实义动词的have to、have got to等不符合研究要求的语料。此外，笔者还删除了离线版本语料库中的乱码语料。经过严格筛选后最终得到符合本研究所有要求的1,791条语料，其中包含情态动词must的语料有1,065条，包含半情态动词have to的语料有517条，包含半情态动词have got to的语料有209条。同时，笔者核对了语料库中的原文，结合关键词的上下文语境，对少数较难标注的语料逐一进行了标注。在选取和标注影响三种构式变体的变量时，笔者参照了其他学者对must、have to、have got to 的研究，同时学习了Szmrecsanyi *et al.*（2016）对相关变量的分类及标注方式，最后选择了9个可能影响该组情态构式变体的预测变量。为便于后续分析，笔者将变量的因素水平大多设置为三个，如表1所示。

表1　预测变量的名称、因素水平及标注依据

变量名称	因素水平	标注依据
时态	现在时	时态为一般现在时、现在进行体、现在完成体
	过去时	时态为一般过去时、过去进行体、过去完成体（包含must、have to、have got to前有would等）
	其他	must、have to、have got to前有shall、will等表示将来时间的词
体裁	报纸 小说 非虚构 杂志	根据语料库中各条语料的体裁类别进行标注
主语的有生性	无生	主语为组织、机构、国家或形式主语 it
	有生	主语为有生命的人或高等动物
	其他	缺少主语
主语的代词性	代词	主语为人称代词、指示代词、疑问代词、关系代词或不定代词
	名词	主语为（并列的）名词短语或名词性从句
	其他	主语空缺或由表示存现的there充当虚位主语

（待续）

（续表）

变量名称	因素水平	标注依据
动词的语法及物性	不及物	must、have to、have got to之后的动词为不及物动词
	及物	must、have to、have got to 之后的动词为单及物动词、双及物动词、复杂及物动词、及物的短语动词或介词动词
	其他	must、have to、have got to 之后为系动词或是用于标记时体特征的助动词 be/have
小句类型	主句	must、have to、have got to 所在小句为简单句、并列句或主从复合句的主句
	从句	must、have to、have got to 所在小句为定语从句、状语从句或名词性从句
	其他	must、have to、have got to 所在小句为省略句，或是作为独立的插入语成分
小句的取向	否定	must、have to、have got to 所在小句表示否定的语用含义（包括肯定的修辞问句）
	肯定	must、have to、have got to 所在小句表示肯定的语用含义（包括否定的修辞问句）
	中立	must、have to、have got to 所在小句为特殊疑问句或是无明显语义倾向的一般疑问句
动词的动态性	动态	must、have to、have got to 之后为表示动作、事件、状态变化或言语行为类动词
	静态	must、have to、have got to 之后为表示状态、感知、思想、情感类动词
年份	1810—2009	根据各条语料所对应的具体年份进行标注，随后通过 R 语言中的 as. numeric 函数将字符型变量转化为数值型变量

5 数据分析与讨论

在对所有变量都进行标注后，笔者运用R软件，对数据进行条件推断决策树和随机森林分析，探究影响must、have to和have got to三种构式变体选择的主要因素。

5.1 影响三种构式变体选择的多因素分析

条件推断决策树模型的算法原理是：首先进行卡方检验，根据结果决定选取分类特征，如果该分类特征对反应变量的影响显著，则可选取该分类特征进入决策树模型，从而构建最佳决策树模型，将进入模型的数据进行分类和预测。与

反应变量关联最紧密的变量被选取作为第一个分类变量，其他变量根据与反应变量的相关性程度，依次参与模型构建。本研究采用的方法是，在R统计软件中加载 {party} 程序包，使用ctree 函数构建条件推断决策树模型，对全部数据进行统计。笔者将上节所标记的9个变量作为模型的预测变量，反应变量为构式变体类型。由于该变量为分类型变量，因此在条件推断决策模型中，呈现结果时，在最底端使用条形图的形式。图1中的椭圆节点代表预测变量，顶部是节点序号。每个椭圆节点下会有两个分支，分支上标记的是预测变量分裂的条件。模型最底端的条形图即为叶节点，代表符合每种分裂条件组合的数据集，条形图顶部有节点序号及数据总量，底端条形图浅灰色部分、深灰色部分、黑色部分分别代表have got to、have to以及must所占比例，笔者将这三种构式变体简记为g、h、m。预测变量在树形图中所处的层次越高，意味着其对反应变量的影响越大。通过将显著性水平设置为0.05，将末枝最少数据总量设置为30，笔者构建了四层条件推断决策树模型，以便于最后呈现的可视化结果清晰直观、方便解读。数据统计后的结果如图1所示。

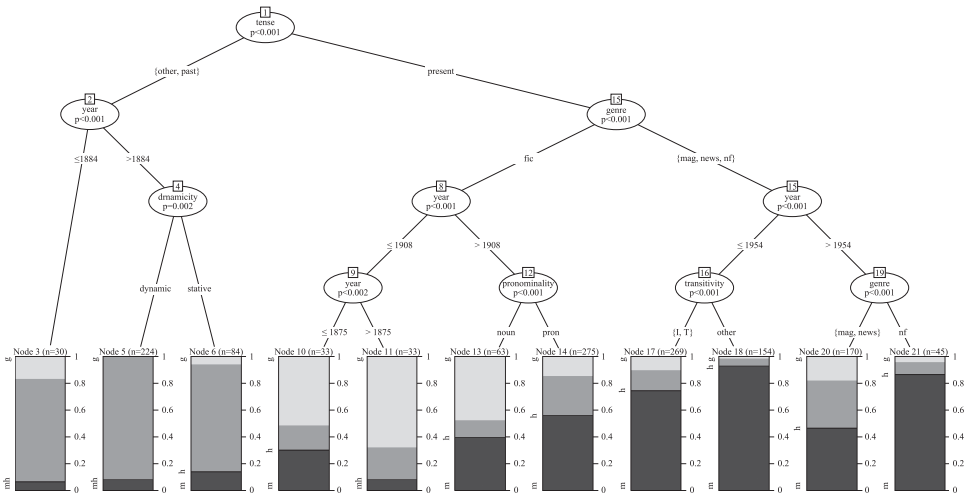


图1　影响must、have to、have got to三种构式变体选择因素的条件推断决策树模型

由条件推断决策树统计结果可知，决策树模型整体具有高度显著性（$P < 0.001$）。笔者利用混淆矩阵（confusion matrix），得出模型分类准确率为71.1%，远高于基准分类准确率59.5%（计算方式：1,065 / 1,065+517+209）。即该决策树模型根据变量特征正确分类的构式变体数量，占整个数据集中构式变体数量的71.1%。

由图1可知，最终共有六个预测变量被选取进入了条件推断决策树模型，分别为：体裁、年代、时态、动词的语法及物性、主语的代词性和动词的动态性。通过

观察决策树的顶部节点可知，影响三种构式变体选择最显著的因素是“时态”。在构式变体所在语句为“现在时”的情况下，多倾向于使用must来表达“必须”的含义，而在其他情况下，则更倾向于使用have to这一能够进行屈折变化的表达。从条件推断决策树的第二层来看，影响构式变体选择的次要因素是“年代”和“体裁”。由节点2的统计数据可知，在表达“过去”和“将来”含义的语料中，1884年以前，have to的使用频数明显高于另外两种构式，而have got to的使用频数也高于must；而在1884年以后，have to所占比例进一步升高，must的使用频数也呈上升趋势，have got to却几乎消失。这一历时演化路径在图形中部的节点8处依然能够得到体现，虽然受体裁影响，have got to所占比例较高，但是总体上have got to的使用频数在历时演变中呈下降趋势，而must和have to的使用频数则呈上升趋势。在节点15中，1954年前，must所占比例最高；1954年后，have got to的使用频数变化并不明显，而must所占比例却呈明显下降趋势，have to的使用频数依然呈上升趋势。由此可以推断，在美式英语历时演化的进程中，时态并不会对三种构式变体的选择产生决定性影响，不管三种构式变体是哪种时态，在表达“必须”这一含义时，must可能正逐渐被have to所取代，这一现象也与Close & Aarts（2010）的观点一致。

除了“年代”这一关键预测变量，“体裁”也在构式变体的选择中起到了关键作用。由节点7的数据可知，在小说这一非正式语体当中，have got to的使用频数高于另外两种构式变体，而在报纸、杂志、非虚构等正式语体当中，have got to的使用频数大幅下降，must所占比例显著上升。这一变化趋势在节点19的数据当中也得到了显著体现，在杂志和报纸中，must的使用频数已经略高于其他两种语体，在非虚构语体中must的使用频数则更高，占据绝对优势。由此可见，在正式语体当中，当表达“必须”这一含义时，倾向于使用must。同时have to也有微弱的上升趋势，并且根据上文的分析，have to在历时演变中更占优势，而have got to虽然在历时演变中不占优势，甚至有逐渐被取代的趋势，但是在非正式语体当中，使用频数依然较高。

从图1也可以看出，动词的动态性、主语的代词性、动词的语法及物性这三个预测变量也进入了条件推断决策模型。由节点4的数据可知，当三种构式后的动词为动态动词时，多倾向于使用have to，当后接静态动词时，have to依然占据主导地位，因此笔者推测动词的动态性并不起到决定性作用。而另外两个变量却对构式变体的选择起到一定作用。由节点12的数据可知，在小说体裁中，若主语为代词，must和have to所占比例上升，must的使用频数更是占据较大优势，而have got to的使用频数则显著下降。由此可见，当主语为代词时，倾向于使用更正式的表达，即must和have to。由节点16的数据可知，在1954年之前的报纸、杂志、非虚构中，关于动词的语法及物性，若构式变体后接系动词be或是用于标记时体特征的助动词be/have，那么must的使用频数占绝对优势。

笔者标注的所有预测变量没有全部进入条件推断决策树模型当中，为了能够全面、直观地反映各个预测变量的相对重要性，笔者通过在R软件中加载{party}程序包，使用cforest 函数构建随机森林模型，对全部数据进行了随机森林分析，并将各个预测变量按照重要性进行排序，结果如图2所示。

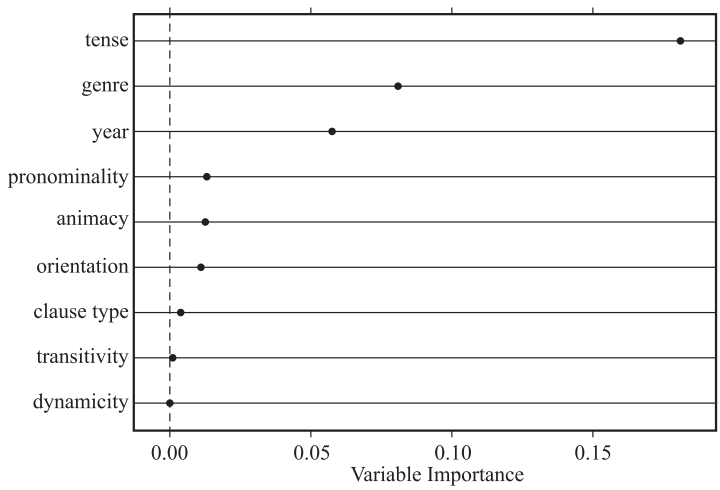


图2　影响must、have to、have got to三种构式变体选择因素的随机森林分析

由随机森林分析的统计结果可知，随机森林模型整体具有高度显著性（$P < 0.001$）。模型的分类准确率为76.6%。笔者通过分析对比两个模型后得出：在四层条件推断决策树中出现的变量，在随机森林分析模型中也位于相对重要性排序的前列，只有“动词的语法及物性”和“主语的代词性”这两个预测变量在重要性排序上变化比较大。由此可推断这两个变量对于构式变体的选择并不起到决定性作用。从图2中也可以看出，在条件推断决策树模型当中，“年代”和“体裁”两个变量均处于第二层，但在随机森林分析模型当中，“体裁”的相对重要性更大，这也说明表达“必须”含义的must、have to、have got to这三种构式变体的选择，在历时演变中受到体裁这一因素的影响更为显著。笔者认为，这可能是由于在美式英语发展的过程当中，对于新闻、报纸等体裁的语言表达有了更严格的规范，因而也就对这三种构式变体的选择更为慎重。

5.2　主要情态动词向半情态动词转变的总趋势

根据前文的分析，总体上看，若以1884年为时间节点，1884年后，must的频数及相对频数大幅增加；若以1908年为节点，1908年后，must的频数及相对频数也大幅增加。但若以1954年为节点，1954年后，must的频数及相对频数呈现出明显的下降趋势，have to的相对频数却呈现出明显的上升趋势。基于must和have to在1954年

后频数及相对频数变化呈现近似互补的态势，笔者推测，情态动词must在1954年后可能逐渐被半情态动词have to取代。笔者又通过在线版本的COHA，利用List项下的POS，分别设置verb.modal及verb.INF，对情态动词must 和半情态动词have to的标准化频数变化趋势进行验证，发现must的标准化频数（即每百万词中must的频数）确实在1900年前后呈现下降趋势（如图3所示），而have to的标准化频数却一直呈现增长趋势（如图4所示）。因此，从相对频数变化趋势近似互补的情况来看，半情态动词have to极有可能正逐渐取代情态动词must。王娟（2007：84）指出："英语情态动词的语义特征主要表现在说话者的主观性上。早期情态动词的主观性较弱，在从弱主观性演变为强主观性的过程中所产生的联想不是和某一事物的特征相关，而是和说话者所要表达的意愿、态度相关，如must 和 may 它们语义演变的过程中，主观性都得到了加强。"由此可知，must主观性加强，必定有一个词来承担其原有的语义及语用功能。have to与must意义相近，且很多研究表明，表达相同含义时，must倾向于表达"主观驱动"而have to则更强调"客观必须"。因此，笔者推测：半情态动词have to很有可能正在逐渐取代情态动词must的部分含义，使must的情态意义得以加强。这也与现代英语语法化的理论相符合，即现代英语中出现了新一代助动词，取代了情态动词的一些功能。

ALL	PER MIL	1820	1830	1840	1850	1860	1870	1880	1890	1900	1910	1920	1930	1940	1950	1960	1970	1980	1990	2000	2010
416946	1,029.50	1,230.41	1,124.55	1,192.57	1,193.03	1,203.14	1,215.37	1,184.92	1,084.75	1,123.21	1,155.65	958.97	965.54	935.85	862.45	875.68	763.84	629.45	549.51	418.99	365.19

0.416 seconds

图3　must在1820—2019年间的标准化频数分布情况

HELP	ⓘ	★		ALL	PER MIL	1820	1830	1840	1850	1860	1870	1880	1890	1900	1910	1920	1930	1940	1950	1960	1970	1980	1990	2000	2010
1	ⓘ	★	HAVE TO	172045	424.80	50.28	50.54	57.19	73.72	90.57	122.36	140.88	154.06	196.75	257.97	253.54	378.23	419.75	496.66	543.73	567.34	546.14	585.32	556.98	529.24

1.187 seconds

图4　have to在1820—2019年间的标准化频数分布情况

6　结语

本文通过变异语言学视角，探究了英语情态系统中的情态动词must和半情态动词have to、have got to在美式英语中的历时演化规律，并探究了对三种构式变体选择影响较为显著的因素。由研究结果可知：（1）美式英语中，must在1810—2009年使用频数正在不断下降，have got to虽然在小说语体中仍然占据很大的比重，但是在历时演变中其使用频数正在不断降低，而have to使用频数呈上升趋势；（2）影响三种构式变体选择的最主要因素是"时态"，随后依次为"体裁"和"年代"，除此之外，本研究选取的其他预测变量对三种构式变体选择产生的影响并不明显；（3）1954年之后，半情态动词have to正逐渐取代情态动词must的部分含义。

本研究借助条件推断决策树和随机森林分析实现了对三种构式变体历时演变

的可视化，能够直观、细致地刻画其演变规律和影响其选择的因素，但仍然存在一些不足。have to和have got to都是半情态动词，表达的含义相近，但是在历时演变中两者的变化规律却截然不同。根据Biber *et al.*（1999）的研究，have got to在英式英语对话中比较常见，本研究使用的语料均来自美式英语语料库，且不包含对话这一体裁，所以本研究中包含have got to的语料较少，研究不够充分，后续研究可以结合英式英语语料库进行深入探讨。另外，半情态动词have to正逐渐取代情态动词must部分含义的总趋势仍需搭配词转移研究的印证。若后续研究能够通过大型语料库，探究must与have to后接的搭配词存在由must转移至have to的趋势，研究结果便能更加让人信服。

注释

1　关于本文中的半情态动词，学界有不同称法，Rachel & Traugott（1997）采用quasi-modal（准情态动词）这一术语，Quirk *et al.*（1985）采用semi-auxiliaries（半情态动词）这一术语。本文采用Biber *et al.*（1999）建议的术语，即semi-modal（半情态动词）。这几种术语的区别与联系详见Krug（2000）。

参考文献

BIBER D, JOHANSSON S, LEECH G, et al. Longman grammar of spoken and written English [M]. London: Longman, 1999.

BOD R, HAY J, JANNEDY S. Probabilistic linguistics [M]. Cambridge, MA.: The MIT Press, 2003.

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.

CAPPELLE B, DEPRAETERE I, LESUISSE M. The necessity modals *have to*, *must, need to* and *should*: using n-grams to help identify common and distinct semantic and pragmatic aspects [J]. Constructions and Frames, 2019, 11(2): 220-243.

CLOSE J, AARTS B. Current change in the modal system of English [C]//LENKER U, HUBER J, MAILHAMMAR R. English historical linguistics 2008. Amsterdam: John Benjamins, 2010: 165-182.

DEPRAETERE I, REED S. The handbook of English linguistics [M]. Oxford: Blackwell, 2006.

GRIES S. Multifactorial analysis in corpus linguistics: a study of particle placement [M]. London: Continuum, 2003.

KRUG M G. Emerging English modals: a corpus-based study of grammaticalization [M].

Berlin: Mouton de Gruyter, 2000.

LEECH G. Modality on the move: the English modal auxiliaries 1961-1992 [C]// FECCHINETTI R, KRUG M G, PALMER F R. Modality in contemporary English. Berlin: Mouton de Gruyter, 2003: 223-240.

PEREK F. Argument structure in usage-based construction grammar [M]. Amsterdam: John Benjamins, 2015.

QUIRK R, GREENBAUM S, LEECH G, et al. A comprehensive grammar of the English language [M]. London: Longman, 1985.

RACHEL N, TRAUGOTT E C. Scope and the development of epistemic modality: evidence from ought to [J]. English Language and Linguistics, 1997(1): 295-317.

SMITH N. Changes in modals and semi-modals of strong obligation an epistemic necessity in recent British English [C]//FECCHINETTI R, KRUG M, PALMER F R. Modality in contemporary English. Berlin: Mouton de Gruyter, 2003: 241-266.

SZMRECSANYI B, GRAFMILLER J, HELLER B, et al. Around the world in three alternations: modeling syntactic variation in varieties of English [J]. English World-Wide, 2016(37): 109-137.

高秋萍. 英语论说文中的情态问题[J]. 课程 · 教材 · 教法，2009（10）：60-64.

金婷茹. 中国英语中必然类情态动词must和have to变异形式的多因素分析[D]. 大连：大连海事大学，2018.

孟庆楠，罗卫华. 变异语言学视角下英语边缘情态动词构式多元定量研究——以情态need和实义need为例[J]. 外语教学与研究，2020（5）：688-700.

王娟. 英语情态动词语义演变的动因及合理性探析[J]. 苏州大学学报（哲学社会科学版），2007（6）：82-84.

王振华. 法庭交叉质询中的人际关系——系统功能语言学“情态”视角[J]. 外语学刊，2004（3）：51-59.

许家金. 基于语料库的历时语言研究述评[J]. 外语教学与研究，2020（2）：200-212.

通信地址：116026　辽宁省大连市　大连海事大学外国语学院

贸易冲突话语中英语指责表述的局部功能研究*

安徽工程大学 刘运锋

提要：不同于描写语言宏观系统的整体功能研究，局部功能研究聚焦于受限文本中语言使用的具体功能或意义，是语料库语言学语境下功能研究的发展趋势。本文采用语料库驱动方法，从局部功能范畴及其构成序列入手，对比分析2001—2017年中美反倾销贸易冲突文本中的英语指责表述。分析显示，中美反倾销指责表述呈现出规律性的特征及发展变化，体现出不同话语社团对反倾销实践的认识差异，可为我国应对反倾销贸易冲突提供启示。同时，本研究证实了局部功能视角在揭示文本细微特征方面的可行性和有效性，可为文本分析提供参考。

关键词：贸易冲突文本、指责表述、局部功能、局部语法

1 引言

功能主义是语言学研究的重要流派，对语言功能有多种分类。Jakobson（1960）将其分为指称功能、诗歌功能、情感功能、意动功能、寒暄功能、元语言功能等六类。Halliday（1970）将其分为概念功能、人际功能、语篇功能等三类纯理功能。如果按照人与世界、人与人、人与文化的关系来划分，上述分类也可总括为表达功能、交际功能、创造功能等三类基本功能。语言功能分类取决于语言满足人类需要的使用方式，但这些分类都是对宏观语言系统的概括，远离具体语言使用（卫乃兴 2015）。Austin（1962）的言语行为理论描写具体言语的行事功能，认为有多少使用方式就有多少功能分类。可见，功能分类存在不同程度的范畴化。卫乃兴（2015：13）根据抽象程度将上述功能分类纳入一个由语言系统和语言使用构成的连续统，纯理功能分类最为宏观，靠近语言系统一侧，言语行为功能分类最为微观，靠近语言使用一侧。言语行为理论关注语言使用，在语言学研究历史上具有重要价值，但自上而下的依言行事功能研究更多地符合Wittgenstein（1953）提出的“意义即使用”思想（meaning is use），还没有将言语行为研究真正落点于

* 本文系安徽省省级质量工程新文科、新医科研究与改革实践项目“新文科背景下贸易冲突话语教学与研究实践”（2020wyxm045）的阶段性成果。本文得到李文中教授和许家金教授的指导，特此致谢。

文本。Sinclair(1987,1991 ）利用语料库驱动方法在文本中研究词语意义（ meaning in texts ），将“意义即使用”从概念层面落实到文本操作，为局部的、文本的语言功能研究提供了理论和实践指导。在此基础上，“局部文本主题”（ Sinclair 1999 ）、“局部文本功能”（ Mahlberg 2005；卫乃兴 2015 ）、“局部主题词”（ Scott & Tribble 2006 ）、“局部语义韵”（ Tribble 2000；Partington 2004；李文中 2019 ）等一系列研究为较细颗粒度地描写语言使用提供了新的探索方向。

由此可知，言语行为理论和语料库语言学为局部功能研究提供了两条可选路径。目前，基于言语行为理论的语言功能研究已成为经典做法，而采用语料库语言学方法的则相对较少。有鉴于此，本文采用语料库语言学局部功能视角探讨专业文本中的指责功能，文本类型限定为中美反倾销贸易冲突文本，以期为局部功能和文本分析的结合提供有益借鉴。

2 文献综述

语言整体功能和局部功能是互补而非对立的关系，只是观察视角不同。由系统层面的元功能进入使用层面的局部功能是功能研究的重要发展标志，揭示了一种必然的探索趋势（ 卫乃兴 2015：14 ）。但视角转换或发展背后必然有重要理论支持，语料库语言学局部功能思想可追溯至Firth的描写语言学（ Descriptive Linguistics ）理论。描写语言学的主要任务是陈述意义（ Firth 1951：118 ），通过特定情景语境中“受限语言”（ restricted language ）的多层级意义分析，建立一套适用于特定语言描写的一般语言理论，而不是适用于一般语言描写的普遍理论（ Firth 1957：31-32 ）。据此可知，尝试对语言整体系统进行描写实际上是不现实的，只有考虑文类、语体、使用情景等因素才能将语言描写控制在可操作范围内。根据受限语言概念，Sinclair（ 1980：257 ）开始利用语料库方法描写类型文本“主题”（ aboutness ）来解释意义连贯。文本主题是语言事件或个人立场从一个状态到另一个状态的不间断活动，语言使用在交互层面实现成员间的意义协商，在自主层面进行个人经验的内化，两个层面的活动随着文本展开同时进行（ Sinclair 1992 ）。无论意义连贯，抑或意义协商发展，都需要在完整文本或同主题文本群中借助词语的“相互释义关系”（ paraphrase relationship ）来实现（ Sinclair 2004a：114 ）。言有尽而意无穷，参与者基于不同经验赋予意义解读的自由，但由于“内文”（ intratextual ）限制，文本意义发展会呈现向心性与一致性（ 李文中 2017 ）。Teubert（ 2010：204-207 ）从话语角度将意义限定为“一个话语社团对给定词项、短语、文本片段或完整文本的全部释义内容”，释义则是“那些试图对对应某一表达的话语对象进行解释、确定、修饰，甚至拒绝的文本片段”。Teubert的“互文”（ intertextual ）限制理论增加了意义的历时特点，否定、拒绝的释义特征在意义重述、解释的基础上丰富了意义连贯或意义发展的维度，更强调意义的临时性、协

商性和文本局部性。概言之，Firth的受限语言概念为Sinclair的语料库文本意义研究带来思想启示，Sinclair的文本意义协商性和Teubert的话语释义特征为局部功能或意义研究提供了理论基础。

在方法上，局部功能研究没有采用Firth的多层级意义分析路径，而是以词项（lexical item）为基本单位描写词语搭配意义（Sinclair 1966；Sinclair *et al.* 1970）。词项是自由的、无层级之分的单位。词项分析将词语和语法视为相互渗透和竞争的整体，利用语料库驱动方法在共文（co-text）语境中以搭配为核心进行扩展分析并描写完整意义单位（Sinclair 1996）。根据词语、语法的连续竞争态势，意义单位可分为习语、搭配框架、搭配扩展单位、单个技术词（Sinclair 2004b：5）。由于搭配框架和以搭配为核心的扩展单位充满内部变异，介于完全固定的习语与完全自由的单个技术词之间，且占语言使用的绝大多数，因而成为局部语法描写的重点（Sinclair 2010：41）。语料库语言学意义研究经历搭配分析、扩展意义单位分析、局部语法描写三个发展阶段，为局部功能研究提供了方法指导。局部语法没有数量和种类限制，我们需要大量局部语法来描写真实的语言使用（Sinclair 2004b：6），局部语法路径将局部功能研究推向了新的阶段。

近年来，在局部语法描写框架下，研究者尝试描写评价、请求、道歉、感谢等局部功能（Hunston 2011；Su 2017，2018；Su & Wei 2018），丰富了言语行为研究，显示出局部功能在文本分析中的独特优势。但这些研究大多通过计算机自动识别带有显性标记的言语表述，没有考察无标记的隐性表述。因此，本文采用语料库驱动方法，研究中美反倾销文本中的指责功能，内容涵盖显性和隐性指责表述，期望通过局部功能视角揭示中美反倾销指责表述的特征，为我国应对反倾销贸易冲突提供语言学支持。

3 研究设计

3.1 研究对象和问题

反倾销是对外国商品在本国市场上的倾销所采取的抵制措施。反倾销文本是反倾销实践的语言再现，属于典型的受限语言。反倾销指责是一种具有专门语域特征性的局部功能。本研究采用语料库语言学自下而上的方法，将反倾销文本中表达指责功能的句子称为“指责表述”，收集2001—2017年中美关于两国反倾销的英文报道，从局部功能视角研究中美反倾销指责表述，旨在回答下列问题：

（1）中美反倾销文本中英语指责表述的特征各是什么，有何异同？

（2）中美反倾销文本中英语指责表述呈现什么发展规律？

3.2 研究语料

以anti-dumping/antidumping为搜索词，从《中国日报》《环球时报》《人民日报》英文版中收集完整文本，建立中国反倾销语料库。在LexisNexis Academic中，以anti-dumping OR antidumping AND China OR Chinese为搜索项收集完整文本，建立美国反倾销语料库，语料主要来自*The New York Times*、*The Washington Post*、*The Wall Street Journal*。两个语料库均以搜索项至少出现一次为标准收集完整文本，并人工排除搜索项偶然出现一次的非中美反倾销文本，基本数据见表1。

表1 中美反倾销语料库基本信息

	中国反倾销语料库			美国反倾销语料库		
收集年限	文本数	形符数	形符数/篇	文本数	形符数	形符数/篇
2001—2007	208	140,876	677	296	239,591	809
2008—2012	281	186,209	663	198	156,386	790
2013—2017	297	172,947	582	155	111,991	723
合计	786	500,032		649	507,968	

语料收集年限分三个阶段，分别以2001年中国加入世界贸易组织、2008年全球爆发金融危机、2013年中国开始倡议建立亚洲基础设施投资银行为关键节点，三个节点事件均对中国对外贸易产生了重要影响。分阶段收集便于描写一定时期内反倾销指责表述的典型复现型式，以及深度对比反倾销指责表述的发展规律。

3.3 研究步骤

（1）提出反倾销指责表述的概念定义和操作定义，根据该定义人工标注显性和隐性指责表述。

指责表述的概念定义限定为指责者对受责者应负责的行为、状况、后果给予否定评价的言语行为，目的是指出受责者应负责的消极行为、状况或不良后果，并要求受责者作出改正或弥补。本研究将“有句子终端符号或结束符号的表述”作为完整句，在实际操作中，句子层级内语言形式如果包含下列某个核心概念，即视为指责表述。

a. 有指责者明确表述受责者行为或状态存在问题的消极评价词语；

b. 有明确表述受责者行为或状态给指责者带来不良后果的消极评价词语；

c. 有表述指责者因不良后果而采取惩罚、威胁受责者等行为的词语；

d. 有表述指责者要求受责者采取改正、补救行为的词语。

中美反倾销指责表述可归为两类。一是含有消极评价词语的显性表述，如例（1）利用消极评价词语unfair指责美国采取不公平贸易政策，例（2）通过消极评价词语have been declining指责美国反倾销措施对中国出口的影响，例（3）通过表示惩罚的词语imposed anti-dumping duties指责美国的商品倾销行为；二是没有消极评价词语，但含有委婉指责的隐性表述，如例（4），中国政府敦促美国采取公平贸易政策，建议中含有指责。按照上述分析，a、b、c属于显性指责表述，d属于隐性指责表述。

(1) The United States adopted **unfair** trade protective measures.

(2) China's exports of solar products to the US **have been declining** because of multiple rounds of anti-dumping and countervailing duties.

(3) The Chinese Ministry of Commerce **imposed anti-dumping duties** on US companies after they were found to have dumped Tertiary Butylhydroquinone products in the Chinese market.

(4) The Chinese Ministry of Commerce has urged Washington to abide by its commitment against protectionism and help maintain a free and open international trade environment.

（2）综合标注的指责表述，概括出一套反倾销指责局部功能范畴。

（3）将指责表述的词语、语法实现成分与局部功能范畴一一匹配，形成不同的局部功能范畴序列。

（4）统计局部功能范畴序列出现频数，总结中美反倾销指责表述的使用特征。

4　研究结果与讨论

经过文本细读，共标注出中美反倾销指责表述3,007句，其中美国反倾销指责1,326句，中国反倾销指责1,681句。基于这些表述，概括出7种反倾销指责局部功能范畴（见表2）。

表2　反倾销指责局部功能范畴

局部功能范畴	定义与实例
1.指责者（the accuser）	做出指责行为的人、组织、机构等 例：**The Ministry of Commerce** strongly complains that the US Government directs its protectionist fire at Chinese textile industry.
2.受责者（the accusee）	被指责的人、物、组织、行为、状态等 例：The Ministry of Commerce strongly complains that **the US Government** directs its protectionist fire at Chinese textile industry.
3.指责（accusing）	指责者的指责动作 例：The Ministry of Commerce strongly **complains** that the US Government directs its protectionist fire at Chinese textile industry.
4.影响对象（the affected）	受责者造成消极后果的直接影响对象 例：The Ministry of Commerce strongly complains that the US Government directs its protectionist fire at **Chinese textile industry**.
5.指责内容（specification）	引起指责的具体内容 例：The Ministry of Commerce strongly complains that the US Government directs **its protectionist fire** at Chinese textile industry.
6.链接（hinge）	同一个序列内两个功能范畴之间的链接成分 例：The Ministry of Commerce strongly complains that the US Government **directs** its protectionist fire at Chinese textile industry.
7.行动或建议（actions taken or to be taken）	指责者对受责者采取的行动、威胁、建议等 例：Nine Chinese firms **filed a complaint** against unfair charges of the US Commerce Department. China government also called on the United States **to drop the offending provision**.

局部功能范畴分类需要高度概括，指责者、受责者、影响对象、指责内容等范畴较为容易确立。除此之外，还需对如下功能范畴作出解释。

a. 指责：在反倾销文本中，指责范畴的使用存在两种情况。一是在复合句中一般将主句的报道类谓语成分（如ARGUE[1]、SAY[2]、NOTE、CLAIM、WARN、SUGGEST、ADD等）标注为指责，如例（5）中said标注为指责，引出指责命题。二是在简单句中将指责类谓语成分（ACCUSE、CRITICISE、CHARGE、COMPLAIN等）标注为指责，如例（6）将accuses标注为指责，后面的成分为具体指责内容。

(5) Zhang **said** that safeguard measures pose a great threat to Chinese exporters.

(6) He **accuses** the Bush administration of unfairly punishing China by imposing import duties on glossy paper imports.

b. 链接。Hunston & Sinclair（2000：85）认为，一个句子可以有一个或以上链接。本研究结合实际语料只考虑动词的链接功能。链接范畴使用存在三种情况：一是复合句中从句的非报道类行为动词、情态动词和BE动词，如例（7）中has erected标注为链接；二是简单句中的情态动词、BE动词、非报道类和非指责类行为动词，且动词本身不表达消极意义，如例（8）中took作为链接引出指责内容；三是简单句中这些动词本身含有消极意义，整个动词短语构成指责内容，则出现零个链接，如例（9）中violates the WTO rules标注为指责内容，无链接使用。

(7) He added the US side **has erected** barriers against China.

(8) The United States **took** a series of protectionist moves against Chinese exports, ranging from textiles to television sets.

(9) The US **violates the WTO rules**.

c. 行动或建议。该功能范畴看似与指责表述关系不太紧密，但分析发现该范畴均指向反倾销指责，而且使用频数较高。如例（10）中complaint to the WTO表示指责者对受责者消极情状的一种回击，该行动由后面的指责内容引起。又如例（11）中to immediately cancel its safeguard measures表示指责者建议受责者对其消极情状作出改正。

(10) China filed **complaint to the WTO** about controversial US tariffs of 8 per cent to 30 per cent on steel imports which took effect March 20.

(11) They also called on the United States **to immediately cancel its safeguard measures**.

根据上述局部功能范畴，本研究将指责表述实现的词语、语法成分与局部功能范畴一一对应，形成不同的局部功能范畴序列（Local Functional Category Sequence，简称LFCS）。如表3所示，该指责表述的局部功能范畴序列可描写为“受责者+链接+指责内容”。该序列是一个形式序列，类似于一个抽象的话语单

位，序列内可实现多种结构相似、语义相近的指责表述，其中词语、语法成分丰富多变。

表3　局部功能范畴序列例示

功能	受责者		链接	指责内容	
语法	n. phrase	prep. phrase	v. phrase	n. phrase	prep. phrase / to-inf.
	Anti-dumping and safeguard measures	in the U.S.	have already become	a weapon	for trade protectionism.
词语	Washington		is resorting to	trade protectionism	to pressure China.
	The US's moves		expose	a dangerous rise	in protectionism.

按照该工作思路，本研究从中美反倾销指责表述中总结出19种局部功能范畴序列（见表4），大致分为显性和隐性指责表述两大类。

显性指责表述使用表示消极评价的词语，在序列中均有明确的“指责内容”，可再分为三类。第一类为受责者凸显类指责表述，包括LFCS1、LFCS2、LFCS3、LFCS4、LFCS5，通常将受责者置于主语位置作为指责靶子，指责者不出现，属于直接的、主观的指责表述。第二类为影响对象凸显类指责表述，包括LFCS6、LFCS7、LFCS8，通常将影响对象做主语，强调影响对象因受责者的消极情状而遭受损失，也属于直接的、主观的指责表述。第一类和第二类属于“自言类”（averral）（Sinclair 1986）表述，即文本作者通过个人的解释表达态度及立场。第三类为指责者凸显类表述，包括LFCS9、LFCS10、LFCS11、LFCS12、LFCS13、LFCS14、LFCS15。与前两类相比，该类表述通过指责者的介入将指责内容向后迁延，使指责语气相对缓和，属于间接的、客观的指责表述。指责者凸显类表述属于“借言类”（attribution）表述（Sinclair 1986），即文本作者通过援引他人观点增加表述的言据性。

隐性指责表述没有使用表示消极评价的词语，但均通过提出建议的方式进行指责。本研究将其作为第四类：建议类，具体包括LFCS16、LFCS17、LFCS18、LFCS19。该类指责表述通过提出建议的方式表达指责，指责方式委婉、间接。需要说明的是，建议类表述在形式序列中也出现了受责者凸显类和指责者凸显类表述，与第一类和第三类有交叉现象。但与之相比，建议类指责表述有两个区别特征：一是没有消极评价词语，二是提出建议内容。因此，将其单独划为一类。总

体而言，第一类和第二类指责语气最为直接、主观，第四类指责语气最为间接、委婉，第三类指责语气相对缓和，居于第一类和第二类和第四类中间。

表4　中美反倾销文本中指责表述的局部功能范畴序列

类别	特征		局部功能范畴序列
显性指责表述	有明确“指责内容”	受责者凸显	LFCS1：受责者+链接+指责内容
			LFCS2：受责者+指责内容
			LFCS3：受责者+指责内容+影响对象
			LFCS4：受责者+链接+指责内容+影响对象
			LFCS5：指责内容1+受责者+链接+指责内容2
		影响对象凸显	LFCS6：影响对象+指责内容
			LFCS7：影响对象+指责内容+受责者
			LFCS8：影响对象+链接+指责内容+受责者
		指责者凸显	LFCS9：指责者+指责+受责者+链接+指责内容
			LFCS10：指责者+指责+受责者+指责内容
			LFCS11：指责者+指责+指责内容
			LFCS12：指责者+指责+受责者+链接+指责内容+影响对象
			LFCS13：指责者+指责+影响对象+指责内容+受责者
			LFCS14：指责者+采取行动+受责者+指责内容
隐性指责表述	无“指责内容”，但“提出建议”		LFCS15：指责者+指责+影响对象+采取行动+指责内容
			LFCS16：指责者+链接+受责者+提出建议
			LFCS17：受责者+链接+提出建议
			LFCS18：指责者+指责+受责者+链接+提出建议
			LFCS19：指责者+采取行动+提出建议

4.1　美国反倾销指责表述特征

经统计，美国反倾销指责表述共有1,326句，呈现16种局部功能范畴序列（见表5）。

表5　美国反倾销指责表述局部功能范畴序列频数统计

类别	局部功能范畴序列	2001—2007年		2008—2012年		2013—2017年	
		原始频数	标准频数（百万词次）	原始频数	标准频数（百万词次）	原始频数	标准频数（百万词次）
1	LFCS1	19	79.3	75	480.0	35	312.5
	LFCS2	7	29.2	24	153.5	18	160.7
	LFCS3	10	41.7	8	51.2	8	71.4
	LFCS4	15	62.6	10	64.0	10	89.3
2	LFCS6	6	25.0	20	127.9	3	26.8
	LFCS7	10	41.7	21	134.3	5	44.6
	LFCS9	96	400.7	109	697.0	55	491.1
	LFCS10	82	342.4	53	338.9	65	580.4
3	LFCS11	37	154.4	32	204.6	3	26.8
	LFCS12	33	137.7	36	230.2	10	89.3
	LFCS13	41	171.1	34	217.4	18	160.1
	LFCS14	61	254.6	76	486.0	57	509.0
	LFCS16	36	150.3	7	44.8	11	98.2
4	LFCS17	10	41.7	3	19.2	0	0
	LFCS18	17	71.0	7	44.8	0	0
	LFCS19	14	58.4	15	96.0	4	35.7
合计		494		530		302	

在以上全部指责表述中，受责者凸显类有239句，占比为18.0%；影响对象凸显类有65句，占比为4.9%；指责者凸显类有898句，占比为67.7%；建议类有124句，占比为9.4%。可见，美国反倾销指责表述使用最多的是指责者凸显类，其次是受责者凸显类，再次是建议类，最后是影响对象凸显类。

为具体描写美国反倾销指责表述，现分三个阶段利用原始频数观察其主要使用特征。

第一阶段共出现494句反倾销指责表述，其中使用较多的是指责者凸显类LFCS9、LFCS10、LFCS11、LFCS13、LFCS14；使用较少的主要有受责者凸显类LFCS2、LFCS3，影响对象凸显类LFCS6、LFCS7，建议类LFCS17（见图1）。

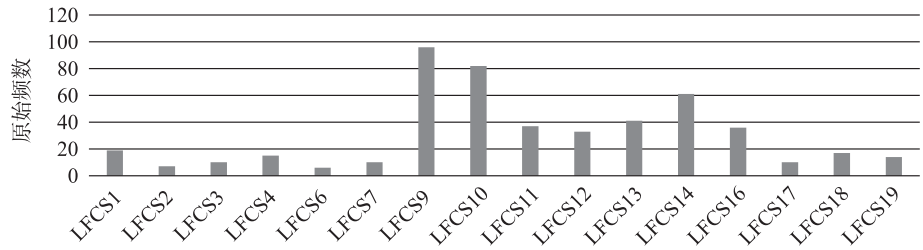


图1　美国反倾销指责表述（2001—2007年）局部功能范畴序列原始频数

在第二阶段，指责表述共出现530句，使用较多的有指责者凸显类LFCS9、LFCS10、LFCS12、LFCS14，受责者凸显类LFCS1；使用相对较少的有受责者凸显类LFCS3、LFCS4，建议类LFCS16、LFCS17、LFCS18（见图2）。

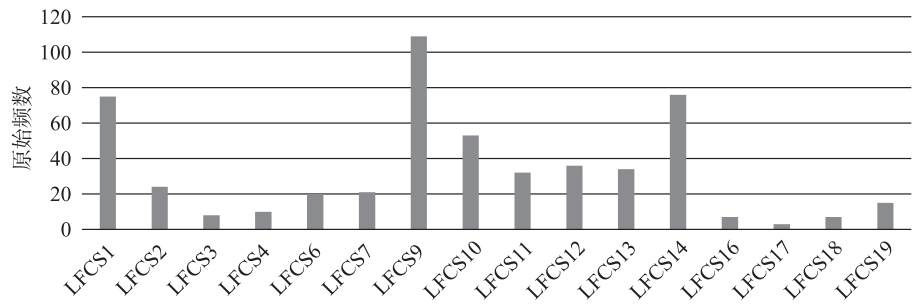


图2　美国反倾销指责表述（2008—2012年）局部功能范畴序列原始频数

在第三阶段，指责表述共出现302句，使用较多的有指责者凸显类LFCS9、LFCS10、LFCS14，受责者凸显类LFCS1、LFCS2；使用相对较少的有受责者凸显类LFCS3，影响对象凸显类LFCS6、LFCS7，指责者凸显类LFCS11，建议类LFCS17、LFCS18、LFCS19，其中LFCS17、LFCS18均出现0次（见图3）。

为观察美国反倾销指责表述在三个阶段的变化，现根据标准频数绘制局部功能范畴序列使用对比图（见图4）。如图4所示，受责者凸显类LFCS2、LFCS3、LFCS4，指责者凸显类LFCS14在三个阶段呈现逐渐上升趋势；建议类LFCS17、LFCS18则呈现逐渐下降趋势。其他指责表述在第二阶段上升，然后下降，呈现特征比较复杂，如受责者凸显类LFCS1，影响对象凸显类LFCS6、LFCS7，指责者凸显类LFCS9、LFCS11、LFCS12、LFCS13以及建议类LFCS19。

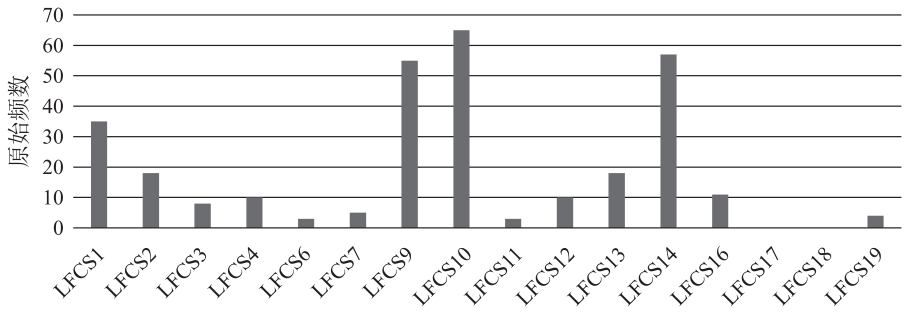


图3　美国反倾销指责表述（2013—2017年）局部功能范畴序列原始频数

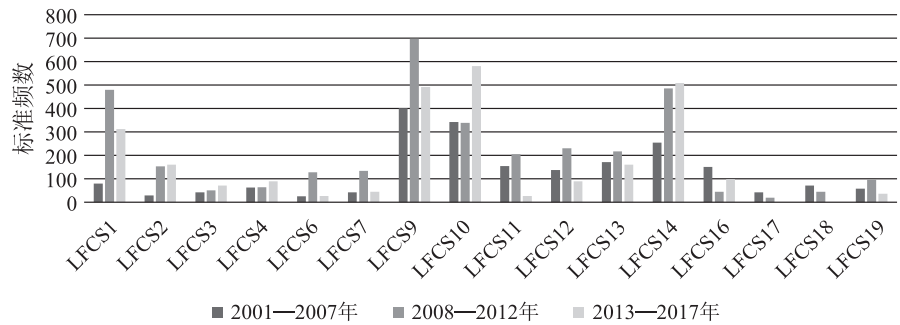


图4　美国反倾销指责表述局部功能范畴序列标准频数对比

经过上述分析发现，美国反倾销指责表述呈现如下特征及变化趋势。

（1）美国反倾销指责表述使用频数由高到低依次为指责者凸显类（67.7%）、受责者凸显类（18.0%）、建议类（9.4%）、影响对象凸显类（4.9%），其中指责者凸显类表述占绝大多数。受责者凸显类和影响对象凸显类都采用指责者隐含方式直接表述指责内容，可合并到一起（总占比为22.9%）。可见，美国反倾销指责中指责者凸显表述远多于指责者隐含表述，客观指责表述多于主观指责表述，较多由权威人士通过引语的方式间接引出指责内容。同时，显性指责表述远多于委婉的隐性指责表述，较少通过提出建议的方式进行反倾销指责表述。

（2）在三个阶段，LFCS2、LFCS3、LFCS4等采取指责者隐含方式进行的反倾销指责表述逐渐增多，说明美国对中国的反倾销指责趋于直接。

（3）LFCS14的使用在三个阶段逐渐增多，说明中国加入世界贸易组织以来，美国对中国的反倾销指责采取惩罚、威胁等行为呈逐渐增长趋势。

（4）同时，LFCS17、LFCS18的使用逐渐减少，说明美国采用建议等委婉方式对中国进行反倾销指责呈逐渐下降趋势，反过来表明，美国对中国的反倾销指责趋于直接。

4.2　中国反倾销指责表述特征

中国反倾销指责表述在三个阶段共1,681句，呈现18种局部功能范畴序列（见表6）。其中，受责者凸显类有397句，占比为23.6%；影响对象凸显类有80句，占比为4.8%；指责者凸显类有891句，占比为53.0%；建议类有313句，占比为18.6%。可以看出，中国反倾销指责表述使用最多的是指责者凸显类，其次是受责者凸显类，再次是建议类，最后是影响对象凸显类。受责者凸显类和影响对象凸显类表述突出的主题不同，但均采用指责者隐含方式进行反倾销指责表述，合并后占比为28.4%。可见，中国对美国的反倾销指责表述一半以上采用指责者凸显方式，其次采用指责者隐含方式，最后采用建议方式。

表6　中国反倾销指责表述局部功能范畴序列数据统计

类别	局部功能范畴序列	2001—2007年		2008—2012年		2013—2017年	
		原始频数	标准频数（百万词次）	原始频数	标准频数（百万词次）	原始频数	标准频数（百万词次）
1	LFCS1	91	646.0	63	338.3	50	289.1
	LFCS2	32	227.2	19	102.0	17	98.3
	LFCS3	11	78.1	10	53.7	5	28.9
	LFCS4	35	248.4	22	118.1	11	63.6
2	LFCS5	26	184.6	5	26.9	0	0
	LFCS6	13	92.3	11	59.1	9	52.0
	LFCS8	32	227.2	10	53.7	5	28.9
3	LFCS9	71	504.0	103	553.1	91	526.2
	LFCS10	41	291.0	55	295.4	40	231.3
	LFCS11	29	206.0	33	177.2	33	190.8
	LFCS12	71	504.0	71	381.3	50	289.1
	LFCS13	18	127.8	56	300.7	58	335.4
	LFCS14	10	71.0	14	75.2	27	156.1
	LFCS15	0	0	8	43.0	12	69.4

（待续）

（续表）

类别	局部功能范畴序列	2001—2007年		2008—2012年		2013—2017年	
		原始频数	标准频数（百万词次）	原始频数	标准频数（百万词次）	原始频数	标准频数（百万词次）
4	LFCS16	18	127.8	62	333.0	155	896.2
	LFCS17	10	71.0	12	64.4	14	80.9
	LFCS18	4	28.4	13	69.8	23	133.0
	LFCS19	2	14.2	0	0	0	0
合计		514		567		600	

接下来，分三个阶段观察中国反倾销指责表述的使用特征及发展变化。

第一阶段共出现514句反倾销指责表述，使用较多的有受责者凸显类LFCS1，指责者凸显类LFCS9、LFCS10、LFCS12；使用相对较少的主要有受责者凸显类LFCS3，影响对象凸显类LFCS6，指责者凸显类LFCS13、LFCS14、LFCS15，建议类LFCS17、LFCS18、LFCS19，其中LFCS15出现0次（见图5）。

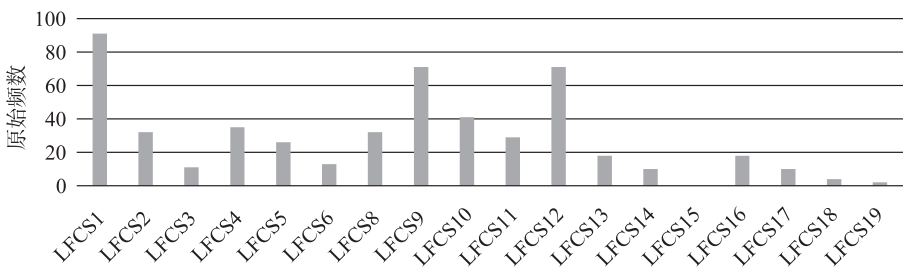


图5　中国反倾销指责表述（2001—2007年）局部功能范畴序列原始频数

第二阶段共出现567句反倾销指责表述，其中使用较多的有受责者凸显类LFCS1，指责者凸显类LFCS9、LFCS10、LFCS12、LFCS13，建议类LFCS16；使用相对较少的有受责者凸显类LFCS3、LFCS5，影响对象凸显类LFCS6、LFCS8，指责者凸显类LFCS14、LFCS15，建议类LFCS17、LFCS18、LFCS19，其中LFCS19出现0次（见图6）。

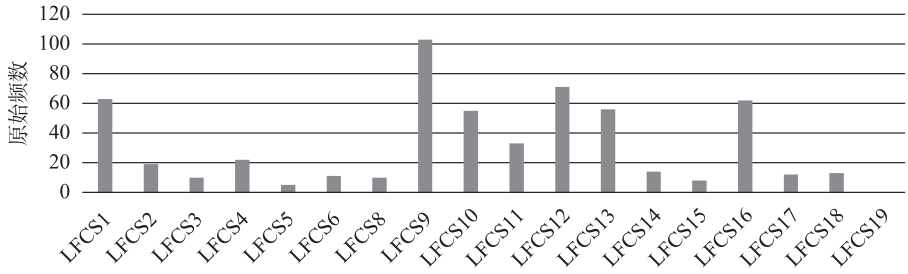


图6　中国反倾销指责表述（2008—2012年）局部功能范畴序列原始频数

第三阶段共出现600句反倾销指责表述，使用较多的有受责者凸显类LFCS1，指责者凸显类LFCS9、LFCS12、LFCS13，建议类LFCS16；使用相对较少的主要有受责者凸显类LFCS3、LFCS4、LFCS5，影响对象凸显类LFCS6、LFCS8，指责者凸显类LFCS11、LFCS14、LFCS15，建议类LFCS17、LFCS18、LFCS19，其中LFCS5、LFCS19均出现0次（见图7）。

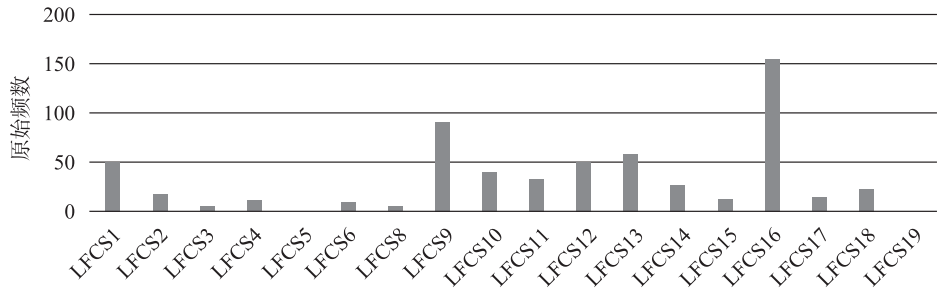


图7　中国反倾销指责表述（2013—2017年）局部功能范畴序列原始频数

根据标准频数绘制局部功能范畴序列使用对比图观察其使用变化（见图8）。如图8所示，指责者凸显类LFCS13、LFCS14、LFCS15，建议类LFCS16、LFCS18在三个阶段呈现逐渐上升趋势；受责者凸显类LFCS1、LFCS2、LFCS3、LFCS4、LFCS5，影响对象凸显类LFCS6、LFCS8，指责者凸显类LFCS10、LFCS12则呈现逐渐下降趋势。其他如指责者凸显类LFCS9在第二阶段上升，然后逐渐下降；指责者凸显类LFCS11和建议类LFCS17则在第二阶段下降，然后逐渐上升，呈现出复杂变化特征。

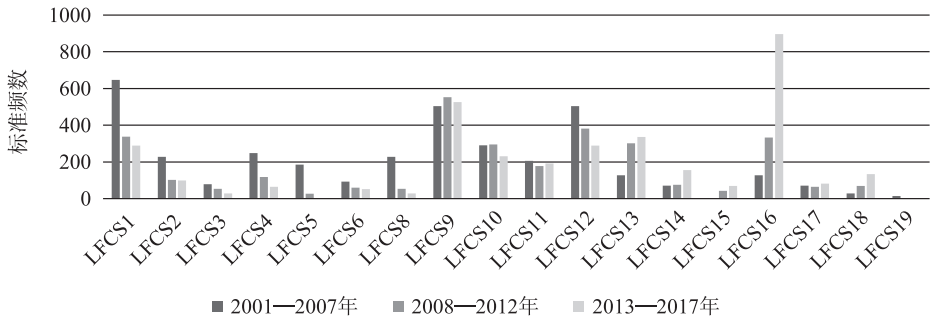


图8　中国反倾销指责表述局部功能范畴序列标准频数对比

通过上述分析，中国反倾销指责表述呈现如下特征和发展趋势。

（1）中国反倾销指责表述使用频数由高到低依次为指责者凸显类（53.0%）、受责者凸显类（23.6%）、建议类（18.6%）、影响对象凸显类（4.8%），其中一半以上采用指责者凸显方式，其次采用指责者隐含方式，最后采用委婉建议方式。总体上，中国反倾销文本中指责者凸显表述远多于指责者隐含表述，客观指责表述多于主观指责表述，显性指责表述远多于委婉的隐性指责表述。

（2）综观三个阶段，受责者凸显类LFCS1、LFCS2、LFCS3、LFCS4、LFCS5和影响对象凸显类LFCS6、LFCS8等采取指责者隐含方式进行反倾销指责表述呈逐渐下降趋势，说明中国对美国的反倾销指责趋于间接。

（3）LFCS14、LFCS15的使用在三个阶段逐渐增多，说明中国对美国反倾销指责采取反倾销申诉、征收反倾销税等活动呈增长趋势，逐步与美国对中国的反倾销惩罚、威胁等行为保持对等。

（4）建议类表述LFCS16、LFCS18的使用在三个阶段呈明显上升趋势。建议类指责属于隐性指责表述，态度较为委婉，说明中国更趋于以合作协商的非对抗方式处理中美反倾销贸易冲突。

4.3　中美反倾销指责表述特征及发展对比

结合以上分析，中美反倾销指责表述的特征及变化异同可总结如下。

中美反倾销指责表述特征的相同点有：

（1）中美反倾销指责表述共呈现19种局部功能范畴序列。这说明在局部功能限制下英语指责表述呈现一定的规律性，同时也显示局部功能范畴组合的灵活性。

（2）中美反倾销指责表述使用频数从高到低均依次为指责者凸显类、指责者隐含类（受责者凸显类、影响对象凸显类）、建议类，其中指责者凸显类表述均占一半以上，中国语料部分为53.0%，美国语料部分为67.7%。中美反倾销文本中，指责者凸显表述多于指责者隐含表述，客观指责表述多于主观指责表述，显性指

责表述多于委婉的隐性指责表述。指责者凸显方式表明指责信息来源和指责者身份，增加信息的言据性，同时突出指责者的态度及立场。指责者凸显类表述占总表述的一半以上，说明国家间交流的正式性和客观性，这也符合新闻报道的“借言”使用特征。大量显性指责表述利用消极、否定的评价词语突出贸易冲突的尖锐性，而隐性指责表述利用建议、劝说方式寄予未来，又体现出贸易冲突的合作性。值得注意的是，“指责+建议”类隐性表述成为冲突类文本的一个重要特征，这一方面揭示了受限文本中的“复合言语行为”现象（Clyne 1994），另一方面也支持了语言功能多重性的研究结论（Mahlberg 2003，2005；张毓、卫乃兴 2017）。

（3）综观三个阶段，中美反倾销均采用反倾销税、反倾销申诉、补贴调查、进口限制等行为。其中，美国对中国的反倾销指责中采取惩罚、威胁、警告等行为呈逐渐增长趋势。同时，中国对美国的反倾销指责表述采取反倾销申诉、征收反倾销税等行为也呈增长趋势，逐步与美国对中国的反倾销惩罚等行为保持对等。反倾销惩罚一直是美国重要的贸易保护主义措施，在反倾销贸易冲突中，我们要认识到贸易“零和游戏”竞争心态的危险，对于美国贸易保护的“大棒政策”要给予有力还击。

结合中美反倾销指责表述局部功能范畴序列的标准频数和频数分布（分别见图9和表7），发现中美反倾销指责表述特征存在如下差异：

（1）如表7所示，除LFCS2、LFCS3、LFCS6、LFCS9、LFCS11外，中美反倾销指责表述的局部功能范畴序列在使用频数上均呈现显著性差异。这表明中美话语社团对反倾销实践存在很大认识差异，这种差异既有两国经济发展水平的客观差异，也有基于利益、战略等因素的主观差异。中美贸易冲突已成为我国长期面临的一种社会现象。

（2）综观三个阶段，美国采取指责者隐含方式进行反倾销指责表述逐渐上升，对中国的反倾销指责趋于直接。而中国采取指责者隐含方式进行反倾销指责表述呈逐渐下降趋势，对美国的反倾销指责趋于间接。指责者隐含表述属于“自言类”主观指责，从某种程度上缺少信息来源的可靠性，不符合新闻报道要求。因此，对于美国的合理指责，我们应给予解释或改善，对于无理指责，应给予有力驳斥。

（3）美国采用建议方式对中国进行反倾销指责呈逐渐下降趋势，对中国的反倾销指责趋于直接，以竞争者角度处理贸易争端。中国采用建议方式进行反倾销指责则呈明显上升趋势，对美国的反倾销指责趋于间接，更愿意以合作者态度处理中美反倾销贸易冲突。中国以和为贵、和合共赢的处世理念有助于建立和维护公正、合理的国际贸易秩序，但对于国际上不合理的贸易诉求，我们试图用道德感化去解决贸易争端的做法值得进一步反思。

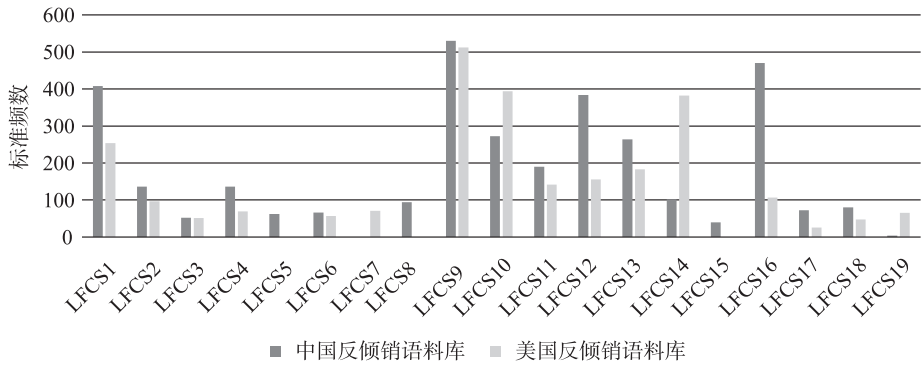


图9　中美反倾销指责表述局部功能范畴序列标准频数对比

表7　反倾销指责表述局部功能范畴序列在两个语料库中的频数分布

中国反倾销语料库			美国反倾销语料库			
局部功能范畴序列	原始频数	标准频数（百万词次）	原始频数	标准频数（百万词次）	对数似然比	显著度
LFCS1	204	408.0	129	254.0	18.24	0.000*[3]
LFCS2	68	136.0	49	96.5	3.41	0.065
LFCS3	26	52.0	26	51.2	0.00	0.955
LFCS4	68	136.0	35	68.9	11.29	0.001*
LFCS5	31	62.0	0	0	43.47	0.000*
LFCS6	33	66.0	29	57.1	0.33	0.569
LFCS7	0	0	36	70.9	–48.55	0.000*
LFCS8	47	94.0	0	0	57.55	0.000*
LFCS9	265	530.0	260	511.8	0.16	0.690
LFCS10	136	272.0	200	393.7	–11.28	0.001*
LFCS11	95	190.0	72	141.7	3.55	0.060
LFCS12	192	384.0	79	155.5	50.38	0.000*
LFCS13	132	264.0	93	183.1	7.42	0.006*
LFCS14	51	102.0	194	381.9	–86.76	0.000*
LFCS15	20	40.0	0	0	28.04	0.000*

（待续）

（续表）

中国反倾销语料库			美国反倾销语料库			
局部功能范畴序列	原始频数	标准频数（百万词次）	原始频数	标准频数（百万词次）	对数似然比	显著度
LFCS16	235	470.0	54	106.3	125.13	0.000*
LFCS17	36	72.0	13	25.6	11.60	0.001*
LFCS18	40	80.0	24	47.2	4.30	0.038*
LFCS19	2	4.0	33	65.0	–32.70	0.000*
合计	1681		1326			

5 结语

本文在语料库语言学局部语法描写框架下，利用局部功能范畴序列细颗粒度剖析中美反倾销文本中的英语指责表述。研究表明，从局部功能视角，采用自下而上的语料库对比方法可有效揭示中美话语社团对反倾销实践的认识特征及差异。此外，本文对“指责+建议”类指责表述的研究结果揭示了语言使用的多功能复杂特征，更印证了从局部功能视角进行意义描写和文本分析的必要性。

以往的语料库语言学局部功能研究聚焦于特定语境下显性表述的文本意义和话语策略，加深了我们对受限语言使用规律的认识。与前人研究相比，本文将显性和隐性表述一起纳入分析范围，更全面地揭示了具体言语行为的局部功能，在方法上与言语行为理论研究互为补充，为文本或话语分析提供了新的研究路径。在实践层面，反倾销指责局部功能研究能够揭示习焉不察的细微特征，为我国应对国际贸易冲突提供有益启示。在外语教学中，相同功能的不同表述序列以及具体实现成分可引入课堂教学，帮助学习者构建商务冲突类型文本。同时，语言表述、专业知识和态度立场的结合有助于学习者提高外语学习的自信心和归属感。

简而言之，从局部功能视角系统地描写受限语言，贴近语言使用并注重个性的社会体验，其理论和实践意义皆不应低估，值得进一步探索。本研究仅从局部功能范畴序列入手，探索性地描写新闻报道文本中的指责功能，后续研究可以：（1）扩大语料收集范围，结合局部功能范畴序列内实现的词语、语法成分，分析中美反倾销指责表述在主题、内容上呈现的特征及规律；（2）描写反倾销指责表述局部语法型式，并利用有限状态实现意义单位的自动识别与提取。

注释

1 文中所有大写动词表示词元（lemma），如ARGUE 包括 argue、argues、argued、has argued、have argued等不同形式。

2 语料中表示信息来源的“according to+指责者”“指责者+SAY”“SAY+指责者”等使用位置灵活，可以出现在句首、句中、句末。为描写方便，本研究将上述使用一并描写成“指责者+SAY”，并置于句首。

3 *表示显著，显著性水平设为0.05。

参考文献

AUSTIN J. How to do things with words [M]. Oxford: Clarendon Press, 1962.

CLYNE M. Intercultural communication at work: cultural values in discourse [M]. Cambridge: Cambridge University Press, 1994.

FIRTH J R. Modes of meaning [C]//TILLOTSON G. Essays and studies 1951: being volume four of the new series of essays and studies collected for the English Association by Geoffrey Tillotson. London: John Murray, 1951.

FIRTH J R. A synopsis of linguistic theory, 1930-1955 [C]//FIRTH J R. Studies in linguistic analysis: special volume of the philological society. London: Basil Blackwell, 1957.

HALLIDAY M. Language structure and language function [C]//LYONS J. New horizons in linguistics. Harmondsworth: Penguin, 1970: 140-165.

HUNSTON S. Corpus approaches to evaluation: phraseology and evaluative language [M]. London: Routledge, 2011.

HUNSTON S, SINCLAIR J M. A local grammar of evaluation [C]//HUNSTON S, THOMPSON G. Evaluation in text: authorial stance and the construction of discourse. Oxford: Oxford University Press, 2000.

JAKOBSON R. Linguistics and poetics [C]//SEBEOK T. Style in language. Cambridge, MA.: The MIT Press, 1960.

MAHLBERG M. The textlinguistic dimension of corpus linguistics: the support function of English general nouns and its theoretical implications [J]. International Journal of Corpus Linguistics, 2003, 8(1): 97-108.

MAHLBERG M. English general nouns: a corpus theoretical approach [M]. Amsterdam: John Benjamins, 2005.

PARTINGTON A. ‘Utterly content in each other’s company’: semantic prosody and semantic preference [J]. International Journal of Corpus Linguistics, 2004, 9(1): 131-156.

SCOTT M, TRIBBLE C. Key words and corpus analysis in language education [M].

Amsterdam: John Benjamins, 2006.

SINCLAIR J. Beginning the study of lexis [C]//BAZELL C, CATFORD J, HALLIDAY M, et al. In memory of J. R. Firth. London: Longman, 1966: 410-430.

SINCLAIR J. Some implications of discourse analysis for ESP methodology [J]. Applied Linguistics, 1980, 1(3): 253-261.

SINCLAIR J. Fictional worlds [C]//COULTHARD R. Talking about text: studies presented to David Brazil on his retirement. Birmingham: University of Birmingham, 1986.

SINCLAIR J. Looking up: an account of the COBUILD project in lexical computing and the development of the Collins COBUILD English language dictionary [M]. London: Collins ELT, 1987.

SINCLAIR J. Corpus concordance collocation [M]. Oxford: Oxford University Press, 1991.

SINCLAIR J. Priorities in discourse analysis [C]//COULTHARD M. Advances in spoken discourse analysis. London: Routledge, 1992: 79-88.

SINCLAIR J. The search for units of meaning [J]. Textus: English Studies in Italy, 1996, 9(1): 75-106.

SINCLAIR J. The computer, the corpus and the theory of language [C]//AZZARO G, ULRYCH M. Transiti letterari e culturali (Volume II). Trieste: EUT Edizioni Università di Trieste, 1999: 1-15.

SINCLAIR J. Trust the text: language, corpus and discourse [M]. London: Routledge, 2004a.

SINCLAIR J. Language and computing, past and present [C]//AHMAD K, ROGERS M. New directions in LSP studies proceedings of the 14th European Symposium on Language for Special Purposes: communication, culture, knowledge, 18-22 August 2003. Guildford: The University of Surrey, 2004b.

SINCLAIR J. Defining the definiendum [C]//G SCHRYVER. A way with words: recent advances in lexical theory and analysis, a festschrift for Patrick Hanks. Kampala, Uganda: Menha Publishers, 2010.

SINCLAIR J, JONES S, DALEY R. English lexical studies: final report for period January 1967-September 1969 [M]. Birmingham: The University of Birmingham, 1970.

SU H. Local grammars of speech acts: an exploratory study [J]. Journal of Pragmatics, 2017, 111: 72-83.

SU H. “Thank bloody God it’s Friday”: a local grammar of thanking [J]. Corpus Pragmatics, 2018, 2(1): 83-105.

SU H, WEI N. “I’m really sorry about what I said”: a local grammar of apology [J]. Pragmatics, 2018, 28(3): 439-462.

TEUBERT W. Meaning, discourse and society [M]. Cambridge: Cambridge University Press, 2010.

TRIBBLE C. Genres, keywords, teaching: towards a pedagogic account of the language of project proposals [C]//BURNARD L, MCENERY A. Rethinking language pedagogy from a corpus perspective: papers from the Third International Conference on Teaching and Language Corpora. Frankfurt: Peter Lang, 2000.

WITTGENSTEIN L. Philosophical investigation [M]. Oxford: Blackwell, 1953.

李文中. 老子“道可道”及英译的内文性解读与验证[J]. 语料库与跨文化研究，2017（1）：91-107.

李文中. 局部语义韵与话语管理[J]. 外国语，2019（4）：81-91.

卫乃兴. 简论局部功能[J]. 外国语文研究，2015（3）：12-20.

张毓，卫乃兴. 学术文本概指名词双重功能研究及对英语教学的启示[J]. 中国外语教育，2017（2）：56-63.

通信地址：241000　安徽省芜湖市　安徽工程大学外国语学院

基于语言学英文学术论文可比语料库的复杂名词短语研究*

北京航空航天大学 高 霞

提要：本文基于语料库研究方法，对语言学英文论文中17类复杂名词短语的使用进行了横向中西学者间的比较和纵向不同写作水平作者间的对比（硕士生和中西作者）。研究结果显示，中国硕士生、中国学者、西方学者三个群体的论文中复杂名词短语使用频数无显著差异，高频使用的复杂名词短语型式相似，均为形容词+名词、名词+名词和名词+介词短语（抽象意义）型式。三类复杂名词短语型式在三组文本间使用差异显著，五类复杂名词短语型式在中国硕士生文本中与中西学者论文中的使用差异显著。学术写作水平似乎是影响学术论文复杂名词短语使用的一个重要因素。本文还对学术英语（English for Academic Purposes，简称EAP）教学和学术论文写作提出了建议。

关键词：复杂名词短语、语料库、学术论文、中西学者、硕士生

1 引言

学术文本常被认为是复杂的、难以理解的专业文体（Biber & Gray 2016：1）。以往众多研究视学术文本复杂性为句法层面问题，并基于此提出了各种文本复杂度指数，如T单位[1]均长、T单位从句数和平均句长等（如Casanave 1994；Stockwell & Harrington 2003；Ellis & Yuan 2004；Elder & Iwashita 2005；Larsen-Freeman 2006；Beers & Nagy 2009；Jiang 2012等）。这些基于句子复杂度的参数被广泛接受并使用，然而甚少有研究验证这些参数的适切性。Biber及其同事所做的一系列基于语料库的多特征多维度跨文体对比研究和学术语篇历时发展研究发现，短语复杂度（phrasal complexity）是学术语篇复杂度的重要参数，句子复杂度（clausal complexity）是口语交际的标志性特征（Biber & Clark 2002；Biber & Gray 2010，2016；Biber *et al.* 2011）。Ravid & Berman（2010） 和Staples *et al.*（2016）讨论了大学阶段本族语学习者写作文本中短语复杂度的发展。Lu（2011）和Parkinson & Musgrave（2014）的研究也指出，随着写作水平提高，二语学习者写作文本中的短语复杂度增加。鉴于此，本文拟基于语言学中国学者、西方学

* 本文系教育部人文社会科学研究一般项目“中国英语能力等级量表不同水平学习者区别性语言特征研究”（20YJA740011）的阶段性成果。

者及中国硕士生学术论文可比语料库，对比分析三个群体的学术语篇中复杂名词短语的使用，以求发现中国学者和硕士生学术语篇短语复杂度的典型特征，验证Biber *et al.*（2011）提出的短语复杂度发展假说。

2 文献综述

学界及EAP教师和学习者一直认为与其他文体相比，学术文本语法结构复杂、意义明确（Biber & Gray 2016：14）。Hughes（1996：33-34）指出，口语交际多用“简短的句子，少用复杂的嵌套从句”，书面文本则使用“更长、更复杂、富含嵌套结构的短语和句子，以及显性的句间关系指示语，清楚表达意义、呈现篇章结构”。Hyland（2007：284）也指出，细化和扩展是学术语篇重要的修辞功能，不同学科文本间的使用差异显著。高水平学习者的写作文本也被认为具有同样特征。如Myhill（1999）发现，使用复杂的语法结构和从属句是高水平学习者学术写作文本的典型特征。学习者修改后的习作中会使用更多扩展句子结构和复杂的语法结构（Keen 2004：96）。

基于以上认知，二语习得和学术写作的众多研究视文本复杂度为句法复杂度，视句长、从属从句和并列从句的数量为文本复杂度的参数（如Willis 2003：192；Purpura 2004：91）。Wolfe-Quintero *et al.*（1998）对比分析39项二语发展研究所使用的100多个参数，指出T单位从句数量和完整句子中依存从句个数是最准确的文本复杂度参数。Ortega（2003）关于27个复杂度研究的对比发现，T单位均长和T单位从句数量是最常用的文本复杂度参数。然而这两个综述性研究均指出，现有研究结果并不一致。这或表明基于句子复杂度的参数并不能全面反映文本复杂度。Bardovi-Harlig（1992：391）指出，采用T单位分析评价二语高级学习者文本的复杂度似乎不能准确反映学习者的语言水平。其他学者，如Ravid（2005）、Rimmer（2006，2008）、Norris & Ortega（2009）、Ravid & Berman（2010），也发现使用基于从属关系的参数研究写作文本复杂度存在一定问题。

早在1960年，Rulon Wells就对比分析了书面文本的名词化特征和口语交际的动词化特征，提出学术语篇中名词比动词更重要。Biber（1985，1986）的多特征多维度文体对比分析也显示，从属从句是口语的典型特征，学术语篇更依赖短语修饰语而非从句。Biber *et al.*（1999）系统描述了英语口笔语语法特征，指出口语中多种类型从属从句的使用频数显著高于笔语。Halliday（1979）的理论层面解析也指出，信息高度集中、词汇密集的书面文本语法结构简单，书面语的复杂度在于词汇和短语。Biber *et al.*（2011）基于大型学术文本语料库对28个语法特征使用的对比研究验证了这一点——复杂名词短语是学术语篇语法复杂度更恰当的参数。他们指出，基于T单位的复杂度参数缺点在于未包含名词短语构建中的非从

句嵌套。Taguchi *et al.*（2013）的研究也发现，写作水平与从句层面的复杂度不相关，复杂名词短语修饰语才是写作文本质量的重要参数。另有相关研究表明，二语学习者写作文本日趋依赖名词短语而非从句或动词短语（Crossley & McNamara 2014），文本中的短语结构随其学术写作水平提高而趋于复杂（Lu 2011；Biber *et al.* 2014；Staples & Reppen 2016；Staples *et al.* 2016）。

然而短语复杂度与学术写作发展的实证研究甚少。Parkinson & Musgrave（2014）对比EAP课程组和在读硕士生两组国际学生学术语篇中复杂名词短语的使用，发现低水平学习者（EAP组）显著多用形容词+名词型式，高水平学习者（在读硕士组）显著多用其他名词短语型式。Ansarifar *et al.*（2018）分析波斯语为母语的语言学硕士、博士学习者论文摘要和语言学国际期刊论文摘要中16类复杂名词短语的分布，发现三个群体最常用的前置名词修饰语均为形容词和名词，后置名词修饰语则为介词短语；随着写作水平提高，论文摘要中的名词+名词型式使用增多；期刊论文摘要中名词+多重介词短语修饰语型式的使用频数显著高于其他两个群体。Ruan（2018）关于语言学中英学者期刊论文摘要名词短语使用的对比研究发现，本族语学者论文中简单名词短语（（限定词）+名词）的使用是中国学者的两倍；中国学者论文摘要多用形容词+名词和名词+名词型式，本族语学者则更多使用名词+of引导的介词短语型式。Lan & Sun（2019）对比分析二语学习者写作文本与期刊论文中11类名词修饰语的使用，发现后者的短语型名词修饰语（如名词+名词型式，名词+介词短语型式）使用频数显著高于前者。

囿于名词短语结构的复杂性和手工标注的工作量，已有实证研究数据样本较小，且多基于论文摘要和学习者文本。论文摘要信息高度集中，主题相关的名词短语复现较多，与论文全文的名词短语使用存在一定差异。鉴于此，本文拟基于语言学中西学者及硕士生学术论文可比语料库，探究三个群体对不同类型复杂名词短语的使用情况。本研究选择语言学学科作为研究对象主要是因为：（1）论文中名词短语的使用存在显著的系统性学科差异——人文学科论文多使用从句修饰，社会和自然科学学科论文则更依赖于名词和短语修饰语（Biber & Gray 2016），而应用语言学论文多被归为社会科学学科；（2）中国仅英语专业硕士生使用英文撰写毕业论文。我们选择应用语言学硕士生论文，以对比不同学术写作水平的作者群体对名词短语的使用情况。选择对比中西学者和硕士生学位论文，一是因为汉英名词短语结构差异较大——汉语名词修饰语均前置（Kirkpatrick & Xu 2012），或许会影响中国学者、硕士生英文学术语篇中名词短语的使用；二是对比新手（硕士生）和专家（中西学者）学术语篇中不同类型名词短语的使用，可进一步验证复杂名词短语分级参数与学术写作经验、水平的相关性。

3 研究设计

3.1 研究问题

本文旨在研究语言学中国学者及硕士生英文论文中复杂名词短语的使用情况，拟回答以下具体问题：（1）语言学中西学者和中国硕士生英文论文中复杂名词短语的使用整体上是否存在显著差异？哪些名词短语特征的使用存在显著差异？（2）三个群体学术语篇复杂名词短语的使用是否与学术写作水平相关？

3.2 研究语料

本研究共使用了三组语料。首先，基于全球最大的文献摘要与科研信息引用数据库Scopus，参照Wood（2001），检索2010年后符合以下条件的应用语言学期刊论文：论文第一作者姓名具有中国大陆人士的姓名特征（排除台湾及其他海外华裔学者姓氏，其特点为语音清化），且工作单位在中国大陆（第一作者姓名为大陆人士，但工作单位为港澳台或海外者，论文排除）。网络检索拟选论文第一作者的简历，确认其为中国大陆人士，且在中国大陆获得博士（硕士）学位。下载满足条件、且有全文链接的论文共10篇，组成中国语言学学者论文库。检索相同期刊2010年后西方学者发表的论文，采用同样的遴选标准和方法（姓名和工作单位）筛选、下载10篇论文，构成西方学者论文库。检索中国知网博硕士学位论文库，下载5篇2010年后应用语言学方向硕士论文。所有文本均以TXT格式存储，删除摘要、表格、图、例子及参考文献。语料库的详细构成见表1。

表1 语料库的构成

群体	形符	类符	类符/形符比	标准化类符形符比
硕士生	116,801 (5)	5,890	5.21	34.50
中国学者	101,405 (10)	6.281	6.51	33.33
西方学者	90,621 (10)	6,905	7.82	36.97

3.3 研究分类框架

本研究基于Biber *et al.*（2011），选择17类复杂名词短语作为研究对象（详见表2）。该框架比较完善，已有多个研究采用（Parkinson & Musgrave 2014；Ansarifar *et al.* 2018；Ruan 2018；Lan & Sun 2019），研究结果可进行多维度的对比。

表2　Biber *et al.*（2011）名词短语修饰语发展阶段假设

发展阶段	复杂名词型式	赋码	例子
2	Attributive adjective+Noun	AN	**significant** difference
	-ed participle+Noun	DPN	a **given** perspective
	-ing participle+Noun	IPN	the **drafting** process
	Noun+Relative clauses	NRC	responses **that did not match the correct pronunciation**
3	Noun+Noun	NN	**manuscript** reviewers
	Possessive noun+Noun	PNN	**participants'** self-report
	Noun+Of phrases as postmodifiers (concrete/locative meaning)	NOC	the Declaration **of Helsinski**
	Noun+Preposition phrases with prepositions other than of (concrete/locative meaning)	NPC	the results **in Table 3**
	Noun+-ed participial clauses as postmodifiers	NDP	the product **advertised**
	Noun+-ing participial clauses as postmodifiers	NIP	a question **concerning Theophanous' future**
4	Attributive adjectives+Noun+Noun	ANN	a **small effect** size
	Noun+Of phrases as postmodifiers (abstract meaning)	NOA	the process **of collaborative writing**
	Noun+Preposition phrases with prepositions other than of (abstract meaning)	NPA	research **in business genres**
5	Noun+Complement clauses	NCC	The finding **that newness does not exhibit clear F0 manifestations independent of focus and topic**_
	Noun+Appositive noun phrases as postmodifiers	NAN	Poland, **an EFL country**
	Noun+Multiple prepositional phrases as post modifiers, with levels of embedding	NMP	questions **about the efficacy of instructor feedback on student writing**
	Noun+To-clauses as postmodifiers	NTC	the pressure **to publish quickly and frequently in competitive journals**

3.4 分析方法

首先，两位标注者（具有语言学博士学位和语料库数据手工标注经验）熟悉17类名词短语的标注方案，增加一类简单名词，试标注小段样本。核对标注结果，讨论两人标注不一致的短语，达成一致。而后随机抽取一篇论文，两人分别独立标注，再次讨论标注不一致的短语，统一认识。然后，抽取6篇论文（24%），两人分别独立标注，剩余的18篇则由作者一人标注。基于双人标注文本，计算kappa系数为0.91，表明一致性很高。最后统计三个子库17类复杂名词短语和简单名词的频数，并进行标准化。使用SPSS 22.0进行单因素方差分析，验证三组语料间的显著差异性，进行对比分析。鉴于要在同一数据集上同时检验17类复杂名词短语的使用是否存在差异，单因素方差分析事后检验选择Bonferroni校正，显著水平设为0.002（0.05/17）。

4 结果与讨论

表3为中西学者及中国硕士生学术论文中简单名词和复杂名词短语的原始频数和百分比。三个群体学术语篇中，中国硕士生论文中复杂名词短语使用最多（74.25%），西方学者论文中简单名词占比最高（28.57%）。这与Ruan（2018）的发现部分一致——本族语学者论文摘要中简单名词的使用频数显著高于中国学者，中国学者论文摘要中复杂名词短语的使用频数显著高于本族语学者。三个子库70%以上的名词均含修饰语，显著高于Biber *et al.*（1999）的发现（60%）。这或是因为他们所用的学术子库包括教材、书籍和期刊论文，而本研究聚焦期刊论文和学位论文。另一个可能的原因是，他们所用的学术子库采样早于1990年，而本研究所用文本均为2010年后。Biber & Gray（2016：142）的英语历时发展研究（1750—1990年）显示，学术语篇中定语从句的数量日趋减少，复杂名词短语的使用日趋增加。三个子库文本中简单名词频数在48.68次/千词—52.98次/千词，复杂名词短语频数在132.43次/千词—140.40次/千词。单因素方差分析结果显示，三个群体学术语篇中简单名词（$F = 0.524$，$P = 0.599$，$\eta^2 = -0.041$）与复杂名词短语（$F = 0.79$，$P = 0.466$，$\eta^2 = -0.018$）的使用均无显著差异。中西学者及中国硕士生学术语篇均体现了名词性这一典型特征。这与Lan & Sun（2019）的发现——中国学习者习作中11类名词短语频数（95.42次/每千词）仅为期刊论文的一半（187.30次/每千词）相悖。原因可能在于他们收集的学习者文本为大一学生课程习作，本研究则为硕士学位论文，两者文体差异较大；且写作和语言水平有一定差异，语言学硕士生作为高级学习者，复杂名词短语的使用更接近专业学者。

表3　简单名词和复杂名词短语的频数

群体	简单名词	复杂名词短语	总数
中国硕士生	5,686（25.75%）	16,399（74.25%）	22,085
中国学者	5,157（27.51%）	13,589（72.49%）	18,746
西方学者	4,801（28.57%）	12,001（71.43%）	16,802

具体到各类复杂名词短语的使用（见图1），三个群体学术语篇均高频使用的四类型式是：形容词+名词、名词+名词、名词+of引导的介词短语（抽象意义）和名词+非of引导的介词短语（抽象意义）。Parkinson & Musgrave（2014）、Ansarifar *et al.*（2018）、Lan & Sun（2019）也有同样的发现。Biber *et al.*（1999）指出，尽管在口语中较少出现，形容词、名词作名词前置修饰语，以及介词短语作名词后置修饰语仍是学术语篇的典型特征。四类型式在三个群体文本中的使用存在细微差别：中国硕士生论文中使用频数由高至低分别为形容词+名词、名词+非of引导的介词短语（抽象意义）、名词+of引导的介词短语（抽象意义）和名词+名词；中国学者论文中则为形容词+名词、名词+名词、名词+非of引导的介词短语（抽象意义）和名词+of引导的介词短语（抽象意义）；西方学者论文中则为形容词+名词、名词+of引导的介词短语（抽象意义）、名词+非of引导的介词短语（抽象意义）和名词+名词。这与Ruan（2018）对中西学者论文摘要中复杂名词短语使用的研究发现稍有出入：西方学者论文摘要中，最高频使用的是名词+of引导的介词短语，其次是形容词+名词和名词+名词；中国学者论文摘要中则依次为形容词+名词、名词+of引导的介词短语和名词+名词。摘要和论文全文对名词短语的使用存在差异，再次凸显学术文本次类研究的必要性。

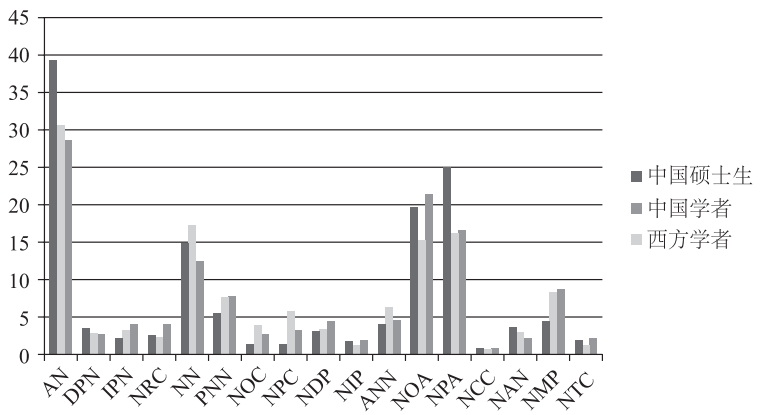


图1　中西学者及硕士生学术论文中复杂名词短语的使用频数（次/每千词）

形容词+名词型式的使用频数在三个子库中均最高，在复杂名词短语总体中，中国硕士生论文中该型式占比最高（28.77%），其次是中国学者论文（23.38%）和西方学者论文（22.05%）。Ansarifar *et al.*（2018）指出，硕士生、博士生和期刊论文中形容词+名词型式占比为22.05%—26.74%。然而EAP学习者习作中复杂名词短语半数以上为形容词+名词型式（57.1%）（Parkinson & Musgrave 2014）。究其原因可能在于Parkinson & Musgrave（2014）研究中的EAP学习者属于新手作者，多用初级复杂名词短语型式。名词+介词修饰语（抽象意义）型式在三个子库文本中的高频使用，验证了Parkinson & Musgrave（2014）、Biber *et al.*（2011）的发现——学术文本中名词+介词修饰语（抽象意义）型式的频数远高于名词+介词修饰语（具体意义）型式。这或是由学术语篇固有特质所决定的，即学术语篇多讨论分析抽象的概念、意义和关系，较少涉及具体的地点和意义。

单因素方差分析的结果见表4。三组文本使用频数差异均显著的三类复杂名词短语为：名词+名词、名词+of引导的介词短语（具体意义）和名词+非of引导的介词短语（具体意义）。中国学者论文中名词+名词型式的使用频数显著高于其他两个群体，中国硕士生论文中名词+名词型式的使用频数显著高于西方学者。Ansarifar *et al.*（2018）基于波斯语博士、硕士和期刊论文的研究则发现，经验丰富的学术论文作者更多使用名词+名词型式。本研究中，中国学者和硕士生论文中名词+名词的使用频数显著高于西方学者，很可能是母语语序迁移的影响——汉语中名词只有前置修饰语（Kirkpatrick & Xu 2012）。Cao & Xiao（2013）的研究也证实了这一点——中国作者倾向于使用形容词和名词做名词前置修饰语。中国学者论文中名词+介词短语（具体意义）型式的使用频数显著高于其他两个群体，西方学者对该类型式的使用又显著高于中国硕士生。Ansarifar *et al.*（2018）则发现，波斯语博士、硕士和学者间该类型式使用无显著差异。名词+介词短语（具体意义）型式是否可以视为不同水平作者的区别性特征还需进一步验证。本研究中，中西学者使用该型式表示具体位置（如下例所示），显著多于硕士生。

（1）the participant in the following excerpt (ES 011001)
the center of the table (CS 011101)
the results in the current study (ES 021201)
the topic in the second sentence (CS 031101)

中国硕士生与中西学者论文中的使用频数差异显著，而中西学者文本间使用频数无显著差异的五类复杂名词短语为：形容词+名词、动名词+名词、所有格

名词+名词、名词+非of引导的介词短语（抽象意义）和名词+多重介词短语型式。这一发现证实了Biber *et al.*（2011）、Parkinson & Musgrave（2014）、Ansarifar *et al.*（2018）及Lan & Sun（2019）的结论——复杂名词短语的使用可区分不同写作水平的作者群体。中国硕士生文本中形容词+名词和名词+非of引导的介词短语（抽象意义）型式的使用频数显著高于中西学者，动名词+名词、所有格名词+名词和名词+多重介词短语型式的使用频数则显著低于中西学者。该发现部分佐证了Biber *et al.*（2011）提出的名词修饰语发展假设——学术写作水平较低的中国硕士生文本中显著多用初级的形容词+名词型式，显著少用较高级的所有格名词+名词和名词+多重介词短语型式。然而必须指出，不同等级、不同类型复杂名词短语在不同水平学术写作文本中并非呈线性增长。如初级的动名词+名词型式在中国硕士生文本中的使用频数显著低于中西学者，而高级的非of引导的后置介词修饰语的使用频数却显著高于中西学者。Ansarifar *et al.*（2018）也指出，Biber *et al.*（2011）提出的名词修饰语具体分级在多等级波斯语学习者语料中未能得到证实。

表4 中西学者及中国硕士生学术论文中各类复杂名词短语的使用差异

发展阶段	复杂名词短语	F值	显著性	Eta方(η^2)	中国硕士生与中国学者	中国硕士生与西方学者	中国学者与西方学者
2	AN	35.048	0.000	.739	*	*	–
	DPN	10.258	0.001	.436	–	*	–
	IPN	27.518	0.000	.688	*	*	–
	NRC	32.626	0.000	.725	–	*	*
3	NN	101.802	0.000	.894	*	*	*
	PNN	23.112	0.000	.648	*	*	–
	NOC	46.943	0.000	.793	*	*	*
	NPC	568.428	0.000	.987	*	*	*
4	NDP	294.700	0.000	.972	–	–	–
	NIP	126.192	0.000	.938	–	–	–
	ANN	488.758	0.000	.983	*	–	*
	NOA	3908.00	0.000	.998	*	–	*
	NPA	3865.5	0.000	.998	*	*	–
5	NCC	57.383	0.000	.871	–	–	–
	NAN	140.616	0.000	.944	–	–	–
	NMP	596.117	0.000	.986	*	*	–
	NTC	95.394	0.000	.919	-	-	*

注：*表示差异显著（P < 0.002），–表示差异不显著（P > 0.002）。

名词+不定式动词短语型式是唯一一类中国学者及硕士生的使用频数无显著差异，且两群体使用频数显著低于西方学者的复杂名词短语。作为高级、习得较晚的复杂名词型式，该型式适用的名词有限（如pressure、responsibility、right、opportunity、willingness、tendency、likelihood、motivation、effort、ability、attempt、decision、choice等），且不符合汉语名词修饰语前置的语序，中国学者和硕士生文本中均少用。Lan & Sun（2019）对学习者文本中各类名词短语的使用频数和学习者托福成绩的相关性研究显示，学习者托福成绩与初级名词修饰语相关度更高。本研究中三个群体使用差异显著的复杂名词短语也集中于初级（2、3等级）。高级（4、5级）复杂名词短语难以区分不同水平的群体，其原因还需进一步探究。

5 结论

本研究基于语言学中西学者和中国硕士生学术论文可比语料库，标注、统计三个群体学术论文中17类复杂名词短语的使用情况，探究不同母语背景、不同学术写作水平群体间复杂名词短语使用的异同。研究结果表明：三个群体英文论文中复杂名词短语使用的总体频数无显著差异，高频使用的复杂名词短语相似，均为形容词+名词、名词+名词和名词+介词短语（抽象意义）型式，即三个群体均高频使用形容词、名词做名词前置修饰语及介词短语做后置修饰语，这支持了Biber *et al.*（2011）、Biber & Gray（2016）的发现——学术文本中，意义多被压缩、以复杂名词短语来表达而非从句，即学术文本的复杂性源于短语复杂性而非从句的嵌套。EAP教学中教师或许应该更关注复杂名词短语的理解和使用，基于学科规范和真实文本，引导学生认识学术语篇的名词化特征，理解复杂名词短语的构成，正确使用复杂名词短语压缩、传递信息。

单因素方差分析结果表明，三类复杂名词短语在三组文本间使用差异显著，五类复杂名词短语在中国硕士生文本中与中西学者论文中的使用差异显著。作者的学术写作水平是影响文本中复杂名词短语使用的一个重要因素，文本中复杂名词短语的使用可区分不同学术写作水平的群体。这一发现与Biber *et al.*（2011）、Parkinson & Musgrave（2014）、Ansarifar *et al.*（2018）及Lan & Sun（2019）的结论一致。然而具体到每类复杂名词短语的使用差异，我们发现不同等级、不同类型名词短语在不同水平学术写作文本中并非呈线性增长。Biber & Clark（2002）提出的名词修饰语从限定性从句修饰语到短语修饰语的发展连续统和Biber *et al.*（2011）提出的复杂名词修饰语发展等级指数的正确性需要二语习得实验数据的进一步验证。

文本文体的类型化不是一个瞬时产物，而是一个由不确定态到范式化的历时沉淀过程，是具体的、个人的话语或语篇经反复运用所形成的特定语言社群单位全体成员共识的约定俗成的范式。囿于语料手工标注工作量，本研究仅对比了25篇论文，采样偏小，未能全面揭示中国学者、学习者英文学术论文中复杂名词短语的使用特征，特别是学科间的使用异同。在标注过程中我们发现，Biber *et al.*（2011）的名词修饰语分类框架未考虑复杂名词短语的长度和复杂度影响。如在形容词+名词型式中，一个与多个形容词、副词+形容词+名词型式在信息压缩的程度及理解和产出的难度上显然存在差异，且多类修饰语嵌套的型式如何归类（如a Nepalese student she called “Surya”到底应归为形容词+名词，还是名词+定语从句），也需进一步探讨。如何正确使用复杂名词短语压缩、传递信息，在文本经济性与信息传递的清楚、准确性间达到平衡，是EAP教学新的关注点，也是我们研究的重点。如何开发程序，实现复杂名词短语的自动和半自动标注，提高标注的准确性，是开展该类研究的关键，也是下一步研究拟突破的难点。

注释

1 T单位包含一个主句和属于这个句子的一切从句。

参考文献

ANSARIFAR A, SHAHRIARI H, PISHGHADAM R. Phrasal complexity in academic writing: a comparsion of abstracts written by graduate students and expert writers in applied linguistics [J]. Journal of English for Academic Purposes, 2018, 31: 58-71.

BARDOVI-HARLIG K. A second look at T-unit analysis: reconsidering the sentence [J]. TESOL Quarterly, 1992, 26(2): 390-395.

BEERS S, NAGY W. Syntactic complexity as a predictor of adolescent writing quality: which measures? Which genre? [J]. Reading and Writing, 2009, 22: 185-200.

BIBER D. Investigating macroscopic textual variation through multifeature/multi-dimensional analyses [J]. Linguistics, 1985, 23(2): 337-360.

BIBER D. Spoken and written textual dimensions in English: resolving the contradictory findings [J]. Language, 1986, 62(2): 384-414.

BIBER D, CLARK V. Historical shifts in modification patterns with complex noun phrase structures: how long can you go without a verb? [C]//FANEGO T, JOSÉ LÓPEZ-COUSO M, PÉREZ-GUERRA J. English historical syntax and morphology. Amsterdam: John Benjamins, 2002: 43-66.

BIBER D, GRAY B. Challenging stereotypes about academic writing: complexity, elaboration, explicitness [J]. Journal of English for Academic Purposes, 2010, 9(1):

2-20.

BIBER D, GRAY B. Grammatical complexity in academic English: linguistic change in writing [M]. Cambridge: Cambridge University Press, 2016.

BIBER D, GRAY B, PONPOON K. Should we use characteristics of conversation to measure grammatical complexity in L2 writing development? [J]. TESOL Quarterly, 2011, 45(1): 5-35.

BIBER D, GRAY B, STAPLES S. Predicting patterns of grammatical complexity across language exam task types and proficiency levels [J]. Applied Linguistics, 2014, 37(5): 639-668.

BIBER D, JOHANSSON S, LEECH G, et al. Longman grammar of spoken and written English [M]. London: Longman, 1999.

CAO Y, XIAO R. A multi-dimensional contrastive study of English abstracts by native and non-native writers [J]. Corpora, 2013, 8(2): 209-234.

CASANAVE C. Language development in students' journals [J]. Journal of Second Language Writing, 1994, 3(3): 179-201.

CROSSLEY S, MCNAMARA D. Does writing development equal writing quality? A computational investigation of syntactic complexity in L2 learners [J]. Journal of Second Language Writing, 2014, 26: 66-79.

ELDER C, IWASHITA N. Planning for test performance: does it make a difference? [C]// ELLIS R. Planning and task performance in a second language. Amsterdam: John Benjamins, 2005: 219-238.

ELLIS R, YUAN F. The effects of planning on fluency, complexity, and accuracy in second language narrative writing [J]. Studies in Second Language Acquisition, 2004, 26(1): 59-84.

HALLIDAY M A K. Differences between spoken and written language: some implications for language teaching [C]//PAGE G, et al. Communication through reading: proceedings of the 4th Australian Reading Conference. Adelaide: Australian Reading Association, 1979.

HUGHES R. English in speech and writing: investigating language and literature [M]. London: Routledge, 1996.

HYLAND K. Applying a gloss: exemplifying and reformulating in academic discourse [J]. Applied Linguistics, 2007, 28(2): 266-285.

JIANG W. Measurements of development in L2 written production: the case of L2 Chinese [J]. Applied Linguistics, 2012, 34(1): 1-24.

KEEN J. Sentence-combining and redrafting processes in the writing of secondary school students in the UK [J]. Linguistics and Education, 2004, 15(1-2): 81-97.

KIRKPATRICK A, XU Z. Chinese rhetoric and writing: an introduction for language teachers [M]. South Carolina: Parlor Press, 2012.

LAN G, SUN Y. A corpus-based investigation of noun phrase complexity in the L2 writings of a first-year composition course [J]. Journal of English for Academic Purposes, 2019, 38: 14-24.

LARSEN-FREEMAN D. The emergence of complexity, fluency, and accuracy in the oral and written production of five Chinese learners of English [J]. Applied Linguistics, 2006, 27(4): 590-619.

LU X. A corpus-based evaluation of syntactic complexity measures as indices of college-level ESL writers' language development [J]. TESOL Quarterly, 2011, 45(1): 36-62.

MYHILL D. Writing matters: linguistic characteristics of writing in GCSE English examinations [J]. English in Education, 1999, 33(3): 70-81.

NORRIS J, ORTEGA L. Towards an organic approach to investigating CAF in instructed SLA: the case of complexity [J]. Applied Linguistics, 2009, 30(4): 555-578.

ORTEGA L. Syntactic complexity measures and their relationship to L2 proficiency: a research synthesis of college-level L2 writing [J]. Applied Linguistics, 2003, 24(4): 492-518.

PARKINSON J, MUSGRAVE J. Development of noun phrase complexity in the writing of English for academic purposes students [J]. Journal of English for Academic Purposes, 2014, 14: 48-59.

PURPURA J. Assessing grammar [M]. Cambridge: Cambridge University Press, 2004.

RAVID D. Emergence of linguistic complexity in written expository texts: evidence from later language acquisition [C]//RAVID D, SHYLDKROT H. Perspectives on language and language development. Boston: Springer, 2005: 337-355.

RAVID D, BERMAN R. Developing noun phrase complexity at school age: a text-embedded cross-linguistic analysis [J]. First Language, 2010, 30(1): 3-26.

RIMMER W. Measuring grammatical complexity: the Gordian knot [J]. Language Testing, 2006, 23(4): 497-519.

RIMMER W. Putting grammatical complexity in context [J]. Literacy, 2008, 42(1): 29-35.

RUAN Z. Structural compression in academic writing: an English-Chinese comparison study of complex noun phrases in research article abstracts [J]. Journal of English for Academic Purposes, 2018, 36: 37-47.

STAPLES S, REPPEN R. Understanding first-year L2 writing: a lexico-grammatical analysis across L1s, genres, and language ratings [J]. Journal of Second Language Writing, 2016, 32: 17-35.

STAPLES S, EGBERT J, BIBER D, et al. Academic writing development at the university level: phrasal and clausal complexity across level of study, discipline, and genre [J]. Written Communication, 2016, 33(2): 149-183.

STOCKWELL G, HARRINGTON M. The incidental development of L2 proficiency in NS-NNS email interactions [J]. CALICO Journal, 2003, 20(2): 337-359.

TAGUCHI N, CRAWFORD W, WETZEL D. What linguistic features are indicative of writing quality? A case of argumentative essays in a college composition program [J]. TESOL Quarterly, 2013, 47(2): 420-430.

WELLS R. Nominal and verbal style [C]//SEBEOK T. Style in language. Cambridge: Cambridge University Press, 1960.

WILLIS D. Rules, patterns and words: grammar and lexis in English language teaching [M]. Cambridge: Cambridge University Press, 2003.

WOLFE-QUINTERO K, INAGAKI S, HAE-YOUNG K. Second language development in writing: measures of fluency, accuracy, and complexity [M]. Honolulu, HI.: Second Language Teaching & Curriculum Center, University of Hawaii, 1998.

WOOD A. International scientific English: the language of research scientists around the world [C]//FLOWERDEW J, PEACOCK M. Research perspectives on English for academic purposes. 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

基于语料库的“主义”词译出译入对比研究*

西安外国语大学 石欣玉 黄立波

提要：毛泽东著作中“主义”词使用频率很高，这一现象不仅体现了毛泽东言语的个性化特征，也成为毛泽东政治话语体系的重要组成成分。本文以《毛泽东选集》中的“主义”词为切入点，以译出和译入版本为研究对象，借助语料库考察两种译本对各类“主义”词的翻译方法，探究不同翻译方向译本对毛泽东个性化言语在处理方式上的异同。考察发现，就“主义”词翻译而言，译出版本和译入版本均未表现出完全的篇内规律性，二者存在一定的篇际相似性和差异性。本文认为，一方面，两种译本不具有完全的篇内规律性，是可接受性与充分性相互妥协、相互调和的结果；另一方面，两种译本存在篇际相似性和差异性，是由于毛泽东著作及其译出版本具有更高的权威性。政治文献外译不仅要确保正确性、及时性和权威性，而且在保证可接受性的基础上还应保留中国特色，制定中国标准。

关键词：“主义”词、翻译方向、语料库

1 引言

近现代汉语“主义”一词从日语“主義”（罗马音 shugi）借用而来（高名凯、刘正埮 1958：96；刘正埮等 1984：408），其用法和意义在汉语中得到进一步发展，进而衍生出大量“主义”词[1]。在中国近现代政治、文化、思想和语言发展过程中，“主义”词以具体的话语实践方式促进了各种思潮、观念、学说和主张的产生、传播、竞争与发展，具有重要的社会价值和历史意义。“主义”词在毛泽东著作中使用频率很高，这不仅展现出毛泽东的个性化言语特征，也带有特定的时代特色和历史烙印。本文从“主义”词切入，以毛泽东著作的译出和译入文本为研究对象，借助语料库考察两种译本对各类“主义”词的翻译，比较两种译本对毛泽东个性化言语的处理方式。

* 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）的阶段性成果。黄立波为本文通讯作者。
作者贡献：
石欣玉：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。

2 “主义”词的来龙去脉

“主义”二字在古汉语中已有使用，其用法和意义与现代汉语存在较大差别。《逸周书·谥法解》中有“主义行德曰元”一句，此处“主义”属动宾结构，意指“谨守仁义”。《史记·太史公自序》中有“敢犯颜色以达主义，不顾其身，为国家树长画”一句，此处“主义”指“对事情的主张”[2]。但现代汉语中“主义”并未沿袭古汉语的用法和意义，而是从日语借用而来。

日本哲学家西周在明治五六年间用古汉语“主義”一词来意译英语principle（余又荪 1935：14），这是“主義”在日语中的最早使用[3]，属独立用法，取“原理”“原则”义（陈力卫 2012：145）。而后“主义”词义进一步扩展，用来表示“系统的理论学说或思想体系”（刘凡夫 2012：15）。同期，“主义”还用以翻译英语词缀“-ism”，其词缀用法开始不断增多。日本哲学家井上哲次郎等人1881年编译出版的《哲学字彙》中，以“主义”为词缀的词语多次用以翻译西方学术术语，如“altruism（爱他心、利他主义）”“federalism（联邦主义）”“egoism（主我学派、自利主义）”等[4]，使“主义”词缀得以正式确立并广泛使用（Spira 2015：125-126；陈力卫 2012：146）。1887年后，“主义”词广泛出现在各种日语译著、教科书、新闻、杂志中（王汎森 2018：142），1920年前后迎来小高潮，1934年左右达到顶峰，1945年后又开始被大量使用（陈力卫 2012：146）。

日语“主義”的用法及意义大约在19世纪80年代传入中国，在近现代汉语中先后产生了独立用法和词缀用法。近现代汉语中“主义”独立用法的最早用例见于晚清外交家、史学家、思想家黄遵宪1880—1887年撰写的《日本国志》：“总理举其立会之主义以告于众”（Spira 2015：122；香港中国语文学会 2001：351），此处“主义”意指“主张”。梁启超（1896）在《论师范》中也用到“主义”一词：“以上诸事，皆以深知其意、能以授人为主义”，此处“主义”属偏正结构，意指“宗旨”。但“主义”独立用法的这两种意义并未延续下来。与之相比，“主义”词缀用法出现于19世纪末20世纪初，演化出三种常用意义并沿用至今：（1）指一定社会制度或政治经济体系，可追溯至1896年《时务报》第十二册：“氏为近世社会主义（学派之名）之泰山北斗也”；（2）指对客观世界、社会生活及学术问题等所持有的系统理论和主张，可追溯至1898年《清议报》第二册：“极东之新木爱罗主义者，……”；（3）指思想作风，可追溯至1901年《清议报》第九十册：“利己主义之对，有爱他主义”（黄河清 2010：981）。

19世纪90年代后，“主义”词在中国逐渐流行并呈泛滥趋势，在报纸、期刊、图书甚至官书中广为使用，随后也引起了一些不满和批评。20世纪早期，报刊文章题目中“主义”词不断增多，其中许多借自日语，一些则为近现代汉语独创；诸如“奋斗”“结婚”“作业”“图书馆”等一般性动词/名词也被冠以“主义”，说明“主义”词在近现代汉语中开始呈泛化趋势（陈力卫 2012：149）。五四运动之后，“主义”词逐渐与政治挂钩，并从20世纪20年代起成为政治“时髦语”（王汎

森 2018：182-185），与政治立场和政治选择紧密相关，成为政治话语中不可或缺的表述形式。但“主义”词的过度流行引发了时人的反思与批判。

尽管“主义”词遭遇批评，但其在19世纪90年代至20世纪40年代的使用频次依然呈显著增长趋势。我们以“主义”为关键词检索全国报刊索引数据库（https://www.cnbksy.com），发现从19世纪90年代到20世纪30年代，“主义”词的使用频次整体上呈攀升趋势，在20世纪30年代达到最盛（15,721次），在20世纪40年代稍有回落但使用频次仍高达11,100次（见图1）。

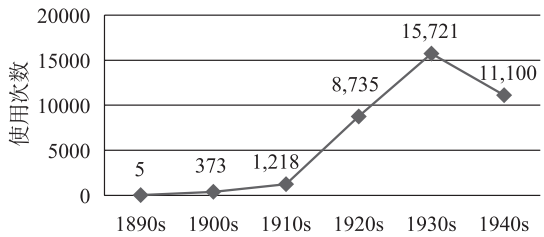


图1　19世纪90年代至20世纪40年代“主义”词的使用频次

总而言之，古汉语“主义”经日语延展后回到近现代汉语[6]，最终产生了近现代汉语中“主义”的用法和意义（见图2），在此过程中，该词的用法和意义发生了语际和语内“变异”。在中国近现代政治、文化、思想和语言的演化发展中，“主义”词以具体的语言形式促进了各种思潮、观念、学说和主张的产生、传播、竞争与发展，因而具有重要的社会价值和历史意义。

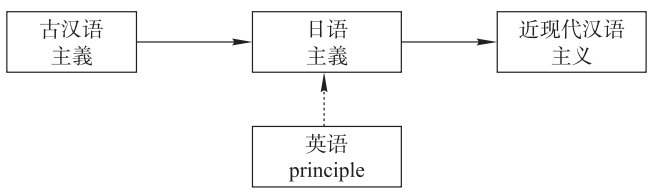


图2　近现代汉语“主义”词的来源

3　毛泽东著作“主义”词的使用

在毛泽东著作中，“主义”词具有举足轻重的意义，并表现出鲜明的时期性特征。毛泽东早期使用的“主义”词均属借用外来概念，且使用密度很低：1916年《致萧子升信》中使用了“们罗主义”一词，1917年《夜学日志首卷》中使用了“严格主义”一词（中共中央文献研究室 1990：49，102）。1917年后，毛泽东开始集中使用“主义”词。在非公开发表的文字中，初次集中使用“主义”词是

在《〈伦理学原理〉批注》一文，其中有“利己主义”“利他主义”“自利主义”等8个“主义”词（中共中央文献研究室：112-257）；在公开发表的文字中，毛泽东初次集中使用“主义”词是在《〈湘江评论〉创刊宣言》一文，其中先后出现“平民教育主义”“劳获平均主义”“实验主义”等8个“主义”词（中共中央文献研究室：264-267）。

本文聚焦于毛泽东著作“主义”词的翻译，研究对象为毛泽东著作译出文本*Selected Works of Mao Tse-tung*（简称SWM）[7]和译入文本*Mao's Road to Power: Revolutionary Writings, 1912-1949*（简称MRP）[8]，研究范围限定在两种译本的重合篇目（共86篇）。需要指出的是，尽管两种译本存在重合篇目，但它们所参照的源文本实际上存在版本差异：SWM以《毛泽东选集》（1950—1961年版，1—4卷）为底本，属于毛泽东著作的修改版本；而MRP主要以《毛泽东集》（竹内实 1983）和《毛泽东集补卷》（竹内实1986）为底本，基本属于毛泽东著作的早期版本。从文本生产和传播的角度看，毛泽东著作的修改版本和早期版本属于“同源文本”，二者之间存在同源关系（石欣玉、黄立波 2021：76），从而为毛泽东著作英译研究提供了可能性。基于两种译本的重合篇目及其对应的源文本，本研究建成毛泽东著作同源文本汉英复合语料库，为考察“主义”词的翻译情况打下基础。

“主义”词在两种源文本中的使用频次较高，并且分布较为广泛（见表1）。在SWM源文本中，“主义”词使用频次为2,737次，标准化频率为0.96%，分布于74个文本（占文本总量的86.05%）；在MRP源文本中，“主义”词使用频次为2,736，标准化频率为0.86%，分布于74个文本（占文本总量的86.05%）。相比而言，在国家语委现代汉语语料库中（库容约1,073万词），“主义”词使用频次为19,318次，标准化频率为0.18%。这说明，与国家语委现代汉语语料库相比，“主义”词在两种源文本中的使用偏多。

表1 “主义”词在两种源文本中的使用情况

源文本	库容	文本总量	“主义词”的使用			
			频次	标准化频率	文本频数	相对文本频率
SWM	286,179	86	2,737	0.96%	74	86.05%
MRP	319,593	86	2,736	0.86%	74	86.05%

我们以中国社会科学院语言研究所词典编辑室编纂的《现代汉语词典》试用本（1973）及其第七版（2016）为标准，一一查证两种源文本中的“主义”词，收录到词典条目的认定为固定短语，未收录到词典条目的则认定为临时短语。以下将对这两类“主义”词进行具体描述。

3.1 “主义”固定短语

SWM源文本和MRP源文本中的“主义”固定短语如表2所示。

表2　两种源文本中的“主义”固定短语

类别	具体词语
固定短语	爱国~、保守~、报复~、本位~、辩证唯物~、大汉族~、帝国~、法西斯~、分散~、封建~、改良~、个人~、个人英雄~、功利~、共产~、官僚~、国际~、机会~、教条~、经验~、军国~、历史唯物~、列宁~、马克思~、马克思列宁~、马列~、冒险~、民权~、民生~、民族~、命令~、三民~、社会~、事务~、投降~、唯物~、尾巴~、文牍~、现实~、享乐~、形式~、修正~、虚无~、主观~、~、资本~、自由~、宗派~

从表2可见，这些固定短语主要是在现代汉语中较为常见的词语。同时，它们在两种源文本中均具有较高的使用次数。例如，“帝国主义”一词在SWM源文本中使用681次，在MRP源文本中使用698次。再如，“社会主义”一词在SWM源文本中使用203次，在MRP源文本中使用182次。

3.2 “主义”临时短语

SWM源文本和MRP源文本中的“主义”临时短语主要属于创造性言语使用，需要注意的是，此处“创造性言语使用”并不专指毛泽东个人的创造性言语使用，还包括源自他处的创造性言语使用。这些词语在两种源文本中的使用次数偏低。例如，“阿Q主义”在两种源文本中均使用2次；“平均主义”在SWM源文本中使用9次，在MRP源文本中使用8次。考察发现，“主义”临时短语的使用方式可进一步划分为借用、改造和创造三种类型，如表3所示。

表3　两种源文本中“主义”临时短语的分类

类别		产生方式	具体词语
临时短语	借用型	（音译/意译）外来概念	多神~、悲观~、佛教~、国家~、基督~、基马尔~、平均~、侵略~、取消~、托洛茨基~、议会~、专制~、自由放任~
	改造型	形容词修饰语+“主义”的固定短语	半三民~、旧三民~、老教条~、伪三民~、新教条~、新三民~、真三民~
	创造型	专有名词+主义	阿Q~、布哈林~、陈独秀~、大波兰~、（李）立三~、张国焘~、章乃器~

（待续）

（续表）

类别		产生方式	具体词语
临时短语	创造型	比喻+主义	阿Q ~、风头 ~、关门 ~、军阀 ~、两个拳头 ~、流寇 ~、奴隶 ~、山头 ~、上山 ~、土匪 ~、一个拳头 ~、自流 ~
		普通词汇+主义	按劳分配 ~、堡垒 ~、不承认 ~、不抵抗 ~、惩办 ~、大后方 ~、单纯军事 ~、地方 ~、东方文化 ~、放任 ~、分裂 ~、革命 ~、个人第一 ~、公开 ~、行会 ~、机械 ~、极端 ~、集中 ~、家族 ~、“精诚团结” ~、空谈 ~、历史 ~、盲动 ~、旧/新民主 ~、排外 ~、拼命 ~、拼一下 ~、迁就 ~、失败 ~、逃跑 ~、屠杀 ~、退却 ~、唯生 ~、小团体 ~、一党 ~、一个 ~、应付 ~、游击 ~、战争绝对 ~、战争相对 ~、自大 ~

借用型“主义”临时短语是以音译或意译方式引入并使用其他语言文化中的概念。例如，“基马尔主义”和“托洛茨基主义”分别音译自土耳其语Kemalizm（英语作Kemalism）和俄语Троцкизм（英语作Trostkysm），“多神主义”和“自由放任主义”则分别是从英语“polytheism”和法语“laissez-faire”意译而来。

改造型“主义”临时短语是在现有“主义”词基础上添加形容词修饰语，构成新的“主义”词。例如，在“三民主义”前冠以不同修饰语，产生“半三民主义”“伪三民主义”“真三民主义”“旧三民主义”“新三民主义”等。

创造型“主义”临时短语是在“主义”前冠以不同搭配词，如“主义”前加上“陈独秀”，产生“陈独秀主义”，用以表示陈独秀的理论和主张。创造型“主义”临时短语具体可分为三个次类：（1）专有名词+主义，如“阿Q主义”“佛教主义”等；（2）比喻+主义[9]，如“自流”一词比喻“在缺乏约束、引导的情况下自由发展”（中国社会科学院语言研究所词典编辑室 2016：1738），后面加上“主义”，产生“自流主义”，表示“放任自由地发展的思想倾向”（诸丞亮、栾培琴 1993：219）；（3）普通词汇+主义，如在经济学概念“按劳分配”后加上“主义”，产生“按劳分配主义”，表示按照劳动者劳动数量和质量分配个人消费品，多劳多得、少劳少得的经济体系。

需要指出的是，以上“主义”临时短语的分类并非界线分明，个别“主义”词实际上同时属于两个类别。如“阿Q主义”不仅属于“专有名词+主义”类，也属于“比喻+主义”类，表示阿Q式精神胜利者的行为或主张。

整体上，“主义”固定短语和临时短语可以说明，“主义”词在两种源文本中呈高频次、多样化使用特征，体现了毛泽东的个性化言语特征。基于毛泽东著作同源文本汉英复合语料库，下文将考察“主义”词在SWM译本和MRP译本中的翻译方法，探究不同翻译方向的译本在毛泽东个性化言语特征处理方式上的异同。

4　毛泽东著作“主义”词的英译

本节将先后考察对“主义”固定短语和临时短语的英译，比较两种译本对这两类“主义”词的翻译方法。

4.1　“主义”固定短语的英译

对于“主义”固定短语，两种译本基本使用以-ism或-ist为后缀的词语进行翻译。例如，“帝国主义”在两种译本中均主要译作imperialism。再如，“自由主义”在译出文本SWM中主要译作liberalism（巫和雄 2013：81-86，283-287），在译入文本MRP中亦是如此。

但也存在特殊情况。例如，对“官僚主义”“民权主义”“三民主义”等词，SWM译本和MRP译本均采用多样化翻译方法。限于篇幅，下文将以“官僚主义”的翻译为例，对比两种译本的处理方法。

“官僚主义”指“脱离实际，脱离群众，不关心群众利益，只知发号施令而不进行调查研究的工作作风和领导作风”（诸丞亮、栾培琴 1993：369），在毛泽东著作中最早出现于《必须注意经济工作》（1933）一文。关于“官僚主义”是否属于日语借词，学术界存在争议。顾江萍（2011）研制的日语借词专题语料库中未收录该词，而胡毅美（2012）、常晓宏（2014：207）等则将该词视作日语借词。不过，无论在日语还是汉语中，“官僚主义”一词与英语bureaucracy均存在密切关联。我们在全国报刊索引数据库中以“官僚主义”为关键词进行检索，发现早在1911年，甘永龙（1911：4）就在其节译文章《日本在高丽之进步》[10]中使用了“官僚主义”[11]翻译bureaucracy[12]一词。此外，我们进一步检索20世纪早期的一些英华词典[13]，发现“官僚主义”属于bureaucracy[14]在汉语中的后期译词之一。由此可以推测，“官僚主义”属于翻译词。

统计发现，在所考察的重合篇目中，译出文本SWM对“官僚主义”一词基本完全使用bureaucracy进行翻译，仅有1次使用bureaucratic practices；而译入文本MRP则使用了bureaucracy、bureaucratism和bureaucratic practices三种译法，其中bureaucratic practices仅使用1次，bureaucracy与bureaucratism使用次数大致相当。说明在考察范围内，两种译本在bureaucracy与bureaucratism的选用上存在一定规律性差异。

为进一步检验该差异，我们将考察范围扩大到两种译本的整体，在其全部译文中分别检索bureaucracy与bureaucratism，发现SWM使用bureaucracy共计18次，而未使用bureaucratism；MRP使用bureaucratism共计54次，使用bureaucracy共计30次。这说明，整体上两种译本在bureaucracy与bureaucratism的选用上确实存在一定差异。

考察发现，bureaucracy与bureaucratism在词义上存在差别。根据*Oxford English Dictionary*（OED），bureaucracy[15]一词最早出现于1815年，初期指“官僚体制”，1818年产生“官僚体制中的官员”义，1843年产生“实行官僚体制的国家/机构/组织等”义，1861年产生“官僚主义/作风”义。相比之下，bureaucratism[16]意义相对简单，主要指“官僚体制”或“官僚体制的主张/实践”，该词源自意大利物理学家、作家Augustus Granville1837年的科普性论著*The Spas of Germany*（第1卷）。

同时，bureaucracy与bureaucratism在使用频率上亦存在差异。OED词典网站显示，bureaucracy使用频次等级为6，在现代英语中出现频次为10—100次/100万词；bureaucratism使用频次等级为4，在现代英语中出现频次为0.1—1次/100万词。此外，在历时美国英语语料库COHA中，bureaucracy使用频次高达1,769次，从19世纪30年代到20世纪80年代之间基本呈增长趋势，20世纪80年代后呈降低趋势（见图3）；bureaucratism使用频次仅有13次，在20世纪30年代、50年代和80年代使用稍多（见图4），表现出较明显的时代特征。

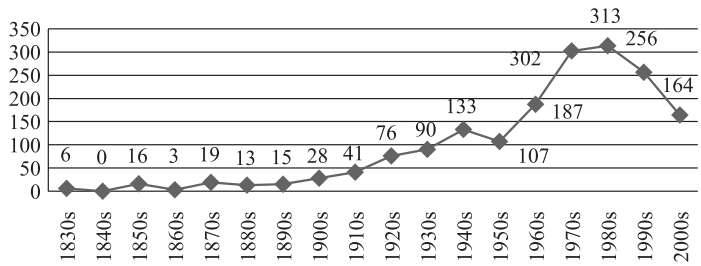


图3　COHA语料库中bureaucracy使用频次历年变化

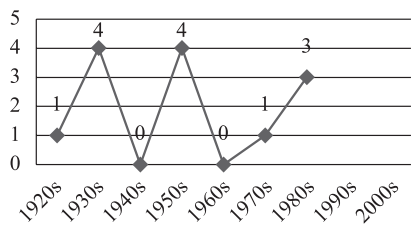


图4　COHA语料库中bureaucratism使用频次历年变化

综合OED和COHA语料库的检索结果可知，英语中bureaucracy使用频次较高，意义较为丰富，包含“官僚主义”之义；bureaucratism使用频次很低，意义较为单一，与“官僚主义”含义之间存在更强、更直接的对应性。

整体上，在翻译“官僚主义”时，两种译本作出了不同的选择：译出文本SWM采用目标语文化中使用频率较高的bureaucracy，在一定程度上降低了翻译文本的异质性，同时增加其可接受性；译入文本MRP倾向于采用目标语文化中使用频率较低、与“官僚主义”对应性更强的bureaucratism一词，在一定程度上保留了翻译文本的异质性，同时增加其充分性。

4.2 “主义”临时短语的英译

本小节将按照表3“主义”临时短语的分类，探讨译出文本SWM和译入文本MRP对各类“主义”临时短语的翻译。

4.2.1 借用型“主义”临时短语的英译

借用型“主义”临时短语来自音译或意译的外来概念，在翻译该类“主义”词时，两种译本均基本使用外来概念的英文对应表述进行翻译，但也有一些例外情况（见表4）。如在翻译“多神主义”“悲观主义”“平均主义”等词时，两种译本均直接使用英文对应词汇；而在翻译“侵略主义”“专制主义”“自由放任主义”等词时，SWM译本仅使用单一译词，MRP译本则采用两种译法。整体上看，在借用型“主义”临时短语的翻译上，两种译本间存在一定的篇际相似性，其中译入文本MRP灵活性较高。

表4　两种译本对借用型“主义”临时短语的英译

“主义”词	SWM	MRP
多神~	polytheism	polytheism
悲观~	pessimism	pessimism
佛教~	Buddhism	Buddhism
国家~	étatistes	étatistes / étatisme
基督~	Christianity	Christianity
基马尔~	Kemalism	Kemalism
平均~	equalitarianism	egalitarianism
侵略~	aggression	aggression / expansionism
取消~	liquidationism	liquidationism

（待续）

（续表）

“主义”词	SWM	MRP
托洛斯基/托洛茨基~	Trotskyism	Trotskyism
议会~	parliamentarism	parliamentarism
专制~	despotism	despotism / autocracy / despotic
自由放任~	laissez-faire	laissez-faire / to follow one’s own inclination

4.2.2 改造型“主义”临时短语的英译

改造型“主义”临时短语由形容词修饰语加“主义”固定短语构成，对该类“主义”词，两种译本的处理方式完全一致（见表5），均采用两种方法进行翻译。方法一是在“主义”固定短语的固定译法前添加英语形容词修饰语：以“老教条主义”为例，两种译本均译作old dogmatism，即在“教条主义”的固定译词dogmatism前添加old。方法二是采用灵活性译法，例如对“半三民主义”和“伪三民主义”，两种译本均将这两个临时短语处理成句子的形式。总体而言，两种译本对改造型“主义”临时短语的翻译呈现出高度篇际相似性。

表5 两种译本对改造型“主义”临时短语的英译

“主义”词	SWM & MRP
半三民~	(Three People’s Principles) become incomplete
旧三民~	old Three People’s Principles
老教条~	old dogmatism
伪三民~	(Three People’s Principles) become false
新教条~	new dogmatism
新三民~	new Three People’s Principles
真三民~	genuine Three People’s Principles

4.2.3 创造型“主义”临时短语的英译

4.2.3.1“专有名词+主义”类

基于“专有名词+主义”形式产生的“主义”词包括两种：（1）“人名+主义”（阿Q主义、陈独秀主义、章乃器主义）；（2）“修饰语+国别/民族+主义”（大波兰主义）。在翻译该类“主义”词时，两种译本也存在一定的相似性（见表6）。

表6　两种译本对“专有名词+主义”类临时短语的英译

“主义”词	SWM	MRP
阿Q ~	Ah Q-ism	Ah Q-ism
布哈林 ~	/	Bukharinism
陈独秀 ~	Chen Tu-hsiuism	Chen Duxiuism
大波兰 ~	Greater Poland chauvinists	great Polish chauvinists
（李）立三 ~	/	(Li) Lisanism
张国焘 ~	/	Zhang Guotao line
章乃器 ~	Chang Nai-chi’s line	Zhang Naiqi’s line

对“大波兰主义”一词，两种译本均使用chauvinists（沙文主义者）进行灵活处理：SWM译本将它处理为Greater Poland chauvinists，MRP译本则将其处理为great Polish chauvinists。对于“人名+主义”形式的“主义”词，两种译本均未采用一致性的翻译方法：在翻译“阿Q主义”和“陈独秀主义”时，两种译本直接在英文人名后面添加了-ism词缀；而在翻译“章乃器主义”时，两种译本却在英文人名所有格后面添加了line（路线）一词。

考察发现，毛泽东著作中“阿Q主义”和“陈独秀主义”均使用2次，且均带有较明显的批判意味（见表7）。此外，该二词贬义色彩较重：根据《毛选》中的注释，“阿Q主义”是“精神上的胜利法”（毛泽东 1991a：270）；“陈独秀主义”是“以陈独秀为代表的右倾投降主义错误”，导致“第一次国内革命战争遭到失败”（毛泽东 1991a：238）。

表7　“阿Q主义”和“陈独秀主义”的使用语境

“主义”词	使用语境
阿Q主义	我们以后者换得前者，重新与国民党合作，为救亡而奋斗。如果说这是共产党的投降，那只是**阿Q主义**和恶意的污蔑。（《中国共产党在抗日时期的任务》，《毛泽东选集第1卷》）
	打小算盘，弄小智术，官僚主义，**阿Q主义**，实际上毫无用处。这些东西，用以对付敌人都不行，用以对付同胞，简直未免可笑。（《反对日本进攻的方针、办法和前途》，《毛泽东选集第1卷》）

（待续）

（续表）

“主义”词	使用语境
陈独秀主义	但在对于革命的性质、任务和方法的认识方面，却表现了它的幼年性，因此在这次革命的后期所发生的**陈独秀主义**能够起作用，使这次革命遭受了失败。（《矛盾论》，《毛泽东选集第1卷》）
	现在，党的政策必须与此不同，不是“一切斗争，否认联合”，也不是“一切联合，否认斗争”（如同一九二七年的**陈独秀主义**那样），而是联合一切反对日本帝国主义的社会阶层，同他们建立统一战线，……（《〈农村调查〉的序言和跋》，《毛泽东选集第3卷》）

相比之下，《毛泽东选集》中仅有1次使用“章乃器主义”。该词出现于“在上海，对‘少号召，多建议’的章乃器主义给了批评，开始纠正了救亡工作中的迁就倾向”一句，从其中“批评”二字可知，毛泽东对“章乃器主义”的态度相对缓和，不带有强烈的针对性意见。根据文末注释，章乃器提出的“少号召，多建议”主张是错误的，“后来，他已逐步地认识了这个错误”（毛泽东 1991b：398）。这亦可说明，《毛泽东选集》中“章乃器主义”一词的贬义色彩相对较轻，批判意味相对较低。

由以上可见，虽然均属于由“人名+主义”形式构成的临时短语，“阿Q主义”“陈独秀主义”和“章乃器主义”在感情色彩和具体意义上存在明显差异：“阿Q主义”和“陈独秀主义”贬义色彩较重、批判意味较强，而“章乃器主义”贬义色彩较轻、批判意味较弱。

我们在OED网站分别考察英文单词line和后缀-ism的具体意义，发现line[17]一词可作“规则/原则/准则”讲，感情色彩上属中性；而以-ism[18]为后缀的名词在一定程度上体现出语言使用者对所言内容的成见。line和-ism在意义和用法上的区别，可以解释两种译本在翻译“人名+主义”类临时短语时所采取的不同处理方式。

4.2.3.2“比喻+主义”类

“比喻+主义”类临时短语最具典型性。比喻能够增强语言表达功能和语用效果，如“奴隶主义”一词汲取“奴隶”的抽象化意义，喻指“不加思考就盲从追随的思想作风”（诸丞亮、栾培琴 1993：157），生动形象，言简意赅。

从表8可见，对“比喻+主义”类临时短语，两种译本均倾向于灵活性处理。具体处理方式包括：（1）使用以-ism为后缀的词语，如“阿Q主义”“关门主义”“军阀主义”的翻译；（2）使用范畴词，如“山头主义”的翻译；（3）使用动名词形式，如“两个拳头主义”“自流主义”的翻译。其中“流寇主义”一词的翻译最能体现两种译本的高度灵活性译法。

表8　两种译本对“比喻+主义”类临时短语的英译

“主义”词	SWM	MRP
阿Q ~	Ah Q-ism	Ah Q-ism
风头 ~	ostentation / seeking the limelight	ostentation / seeking the limelight / (self-) aggrandizement
关门 ~	closed-doorism	closed-doorism
军阀 ~	warlordism	warlordism
两个拳头 ~	striking with two “fists” in two directions at the same time	striking with two “fists”
流寇 ~	ways of roving rebels / ideology of roving rebel bands / in the manner of roving rebels / roving-rebel idea / roving rebels / roving-rebel ideology	roving-rebel-style / roving-rebelism / roving rebel ideology / ideology of roving rebel bands / roving rebel idea
奴隶 ~	slavishness	slavishness
山头 ~	mountain-stronghold mentality	mountain-stronghold mentality
上山 ~	principle of going up into the hills	doctrine of going up the mountains
土匪 ~	bandit ways	banditism
一个拳头 ~	striking with one “fist” in one direction at one time	/
自流 ~	letting things slide	letting things take their natural course

4.2.3.3“普通词汇+主义”类

“普通词汇+主义”类是数量最多、涵盖领域最广的一类“主义”临时短语。与其他类别“主义”临时短语的翻译相比，两种译本在翻译此类短语时采用的灵活性处理最多（见表9）。具体包括以下三种翻译方法：（1）其他译法，如“‘精诚团结’主义”“民主主义”“拼命主义”等词的翻译；（2）使用范畴词，如“按劳分配主义”“不承认主义”“单纯军事主义”等词的翻译；（3）使用以-ism或-ist为后缀的词语，如“逃跑主义”“退却主义”“机械主义”等词的翻译。在“普通词汇+主义”类临时短语的翻译上，两种译本呈现出较明显的篇际差异性。

表9　两种译本对“普通词汇＋主义”类临时短语的英译

“主义”词	SWM	MRP
按劳分配～	doctrine of distribution according to labour	doctrine of distribution according to labor
堡垒～	policy of blockhouse warfare / building blockhouses / blockhouse warfare / principle of blockhouse warfare /	blockhouse warfare / the blockhouse philosophy / building blockhouses / blockhouse / blockhouse-ism / principle of blockhouse warfare
不承认～	policy of non-recognition	policy of nonrecognition
不抵抗～	non-resistance	nonresistance
惩办～	punitiveness	punishmentism / to be punished
大后方～	policy of maintaining a great rear area	having a large rear area
单纯军事～	purely military approach	purely military approach
地方～	localism	localism
东方文化～	Doctrine of Oriental Culture	Doctrine of Oriental Culture
放任～	doing nothing about them	laissez-faire-ism
分裂～	/	separatist
革命～	revolution / revolutionary cause	revolution / revolutionary cause
个人第一～	doctrine of “me first”	doctrine of “me first”
经济公开～	/	economic openness
行会～	/	guild
机械～	mechanists	mechanists
极端～	/	extremism
集中～	centralism / centralization	centralism / centralization
家族～	clan system / clan authority / clan sentiment	clan system / clan influence / clan ideology
“精诚团结”～	“unity in good faith”	“sincere unity”
空谈～	phrase-mongering / windbag / idle chatter	empty talk / empty-talkism / empty-talkist / phrase-mongering
历史～	(in our) historical approach	historitists

（待续）

（续表）

“主义”词	SWM	MRP
盲动～	putschism / haste / impetuosity	putschism / adventurism
旧/新民主～	old / new democracy / old / new democratic	old / new democracy / old / new democratic
排外～	anti-foreign / to shut it out / exclusivism	anti-foreign / isolationism / antiforeignism
拼命～	desperate recklessness	desperate recklessness
拼一下～	/	desperationism
迁就～	excessive accommodation / deviation of accommodation	excessive accommodation
失败～	defeatism	defeatism
逃跑～	flightism	flightism
屠杀～	policy of massacre	butcherism
退却～	retreatism	retreatism
唯生～	vitalism	vitalism
小团体～	“small group” mentality	small group mentality
一党～	one-party doctrine	one-party doctrine
一个～	“one doctrine”	“one doctrine”
应付～	to have recourse to passive response	to react to circumstances
游击～	guerrilla-ism	guerrilla-ism
战争绝对～	idea of war as an absolute	war as an absolute
战争相对～	to become a relativist in war	to become relativists in war
自大～	/	conceited types

总体而言，两种译本对各类“主义”词的翻译均存在一定多样性，主要采用以下五种翻译方法：（1）使用以-ism或-ist为后缀的词语；（2）使用表达理论学说、思想体系、社会制度、政治经济体系、思想作风等意义的范畴词，如principle、doctrine、idea、theory、mentality、ideology、policy、approach、way、strategy、tendency等；（3）借用外来词；（4）使用“修饰语+已有固定短语的英译”形式；（5）采用其他译法。

宏观层面上，在“主义”词翻译方法的整体模式方面，两种译本均不具有完全的篇内规律性。在翻译“主义”固定短语时，两种译本均主要使用以-ism或-ist为后缀的词语，仅有个别例外情况。而在翻译“主义”临时短语时，情况则较为复杂：（1）对于借用型临时短语，两种译本均主要采用以-ism为后缀的词语；（2）对于改造型临时短语，两种译本均主要采用“修饰语+已有固定短语的英译”形式；（3）对于创造性临时短语，两种译本使用其他译法进行灵活性处理情况居多。

微观层面上，就“主义”词的具体译法而言，两种译本间存在一定的篇际相似性和篇际差异性。我们可以观察到MRP译本“借鉴”SWM译本的多处痕迹（如“章乃器主义”“半三民主义”“伪三民主义”“单纯军事主义”等词的翻译），亦可观察到MRP译本“偏离”SWM译本的一些地方（如“官僚主义”“大波兰主义”“土匪主义”“屠杀主义”等词的翻译）。译入文本MRP同时作出“借鉴”与“偏离”的翻译选择，造成两种译本之间存在一定的篇际相似性及差异性。

5 讨论

5.1 关于规律性、相似性及差异性的阐释

毛泽东著作中“主义”词的频繁使用，不仅是毛泽东个性化言语的突出特征，也带有一定的时代和历史烙印，还是特定历史时期中国政治、经济、文化、军事等多个领域的“关键词”。整体而言，在“主义”词的翻译上，毛泽东著作译出文本SWM和译入文本MRP均未呈现完全的篇内规律性，同时二者之间存在一定的篇际相似性和差异性。两种译本表现出的这些特征可从两方面进行解释。

第一，两种译本均不具有完全的篇内规律性是可接受性和充分性共同作用的结果。Toury分别从“可接受性”和“充分性”两个方面描述“翻译”的含义：一方面，翻译是特定文化/语言中的文本生产，在目标语文化中占据一定地位或填补一定空缺；另一方面，翻译是源文本的一种表征，该源文本已存在于另一不同文化并在其中占据一定地位（Toury 2012：69-70）。可接受性和充分性是两种不同的观察视角，前者涉及翻译与目标语文化间的关系，后者则涉及翻译与源语言文化间的关系。在翻译过程中，需要不断地就充分性和可接受性作出选择：选择充分性会使翻译文本较多地呈现源语言文化特征或传统，同时较为偏离目标语文化中的规范；选择可接受性则会使翻译文本较多地偏离源文本，同时更加符合目标语文化中的规范（Toury 2012：79-80）。由于微观层面上具体的翻译选择基本不会呈现百分之百的规律性，翻译无法实现完全的充分性或可接受性，而是二者相互妥协、相互融合的结果（Toury 2012：70-71，80-81）。可以说，可接受性与充分性之间并不存在清晰明确的界线，二者仿佛分别处于某个连续统的两极，任何一个翻译文本都能在这个连续统内找到一定的位置。就毛泽东著作中“主义”词的翻

译而言，根据“主义”词的构成方式、词汇意义和使用语境，译出文本SWM和译入文本MRP分别就可接受性和充分性作出具体的翻译选择，这些选择整体上不会呈现绝对的规律性特征，两种译本亦均不会达到完全的充分性或可接受性。

第二，两种译本的篇际相似性和差异性是由毛泽东著作及其译出文本作为权威型文本的特殊性决定的。一方面，根据Newmark提出的文本类型和文本功能理论，毛泽东著作属权威性文本，翻译过程中应格外重视源文本中的“个人成分”（如特殊搭配、原创隐喻、“不可译”词汇、特殊句法、新词、方言古语等）（Newmark 1988：39-44）。另一方面，毛泽东著作译出文本SWM在毛泽东著作的海内外众多译本中处于典范地位。毛泽东著作对外英译是一次国家翻译实践行为，由中共中央对外联络部牵头，参与者既有国内人员又有国外专家，既有专家学者又有党内干部（巫和雄 2013：36-51），毛泽东本人亦参与其中（程镇球 2002：213），最终产生的译出文本SWM代表着新中国成立初期官方层面的对外发声，其中涉及的翻译选择具有较高参考价值。以“官僚主义”为例，译出文本SWM在翻译时使用英语中出现频次很高、意义较为丰富的bureaucracy，译入文本MRP则倾向于使用英语中出现频次较低、与“官僚主义”含义对应性更强的bureaucratism，MRP译本的翻译选择虽然“偏离”SWM译本，却能更加“靠近”源文本，使源文本的权威地位不受动摇，同时保证翻译文本具有较高准确性。再以“章乃器主义”为例，译出文本SWM中该词译法不同于“阿Q主义”和“陈独秀主义”的译法，应是出自多方面考虑，译入文本MRP对此加以借鉴，能够保证翻译文本的正确性。整体而言，译入文本MRP作出的种种翻译选择，都是基于毛泽东著作及其译出文本作为权威型文本的特殊性。

5.2　对政治文献外译的启示

作为国家翻译实践行为，毛泽东著作对外翻译是新中国成立初期官方层面对外发声的有效尝试，对当前政治文献对外翻译具有一定借鉴意义。

首先，要保证政治文献对外翻译的正确性、及时性和权威性，从而掌握中国国家话语权的主动性和主导性。作为传播中国声音的载体，政治文献对外翻译是引导国际话语的重要方式（黄友义等 2014：5），因此要保证其能够准确传达源文本的信息，并及时地树立权威性。从毛泽东著作译入文本MRP借鉴译出文本SWM“主义”词的翻译方法可知，译出文本“先发制人”，以其较高的权威性获得了海外学界的关注和认可，成为重要的参考标准和借鉴对象。

其次，政治文献对外翻译不仅应与国际话语体系和表达方式对接，保证翻译的可接受性，还应保留中国特色，制定中国标准，完善发布机制，保证翻译的充分性。从历史、文化、语言等方面看，政治文献通常包含许多中国特色术语（如本文研究的毛泽东著作中的“主义”词），翻译时应当体现鲜明的中国特色。

6 结语

随着中国综合国力不断提升，国际社会对中国政策、立场和观点的关注亦不断增多。政治文献对外翻译应具有及时性和权威性，体现鲜明的中国特色，制定并传播中国标准，从而更好地回应国际关切，掌握国家话语权的主动性和主导性，完善国家对外话语体系建设。

注释

1 本文将“主义”独立用法和词缀用法构成的词汇（如“主义”“帝国主义”“社会主义”等）统称为“‘主义’词”。

2 “敢犯颜色以达主义”一句，学界对其中“主义”二字有多种理解和阐释，详见Spira（2015），参见罗竹风（1986：704）。

3 日本政治家、文学家、记者福地源一郎于1878年使用“主義”意译英语principle一词，斋藤毅（1977）和《日本国语大词典》（1974）均将此视为“主義”在日语中的首次使用。但根据余又荪（1935）考证结果，福地源一郎使用“主義”一词的时间实际上晚于西周。

4 Spira（2015）指出，1881年版《哲学字彙》收录13个以“主義”为词缀的词语，但据笔者考证，其中3个条目（“absolutism（專制主義）”“asceticism（嚴肅主義）”“indifferentism（局外主義）”）并未收录其中。

5 香港中国语文学会所编《近现代汉语新词词源词典》中将《日本国志》误作《日本杂事史》，特此说明。

6 顾江萍（2011：105）指出，20世纪90年代，中国语言学界将此类原本借自古汉语、后又以借词身份回到汉语的词汇称为“回归词”。

7 译出文本SWM实际上包括“初版稿”“旧改稿”和外文社版，参见徐永焕（2006），潘卫民、卜海丽（2013）。本研究考察的是1960—1965年的外文社版。

8 译入文本MRP由美国毛泽东研究专家斯图亚特·施拉姆（Stuart Schram）主持翻译，计划出版十卷，目前已出版八卷，涵盖1912—1945年间毛泽东的著作和讲话。

9 本文以《毛泽东言语辞典》（诸丞亮、栾培琴 1993）对“主义”词的释义为参照，将使用比喻的“主义”临时短语划分到“比喻+主义”类。

10 节译自1911年*The North China Daily News*（《字林西报》）“Progress in Korea”一文。

11 全句为：“夫官吏之增多，自由于邦国之发达，然欲求殖民地之进步，则实无取于官僚主义。”

12 英文原文为：“No doubt the development of the country is answerable for some

of this enormous advance, but it is a sound maxim of colonial progress that the less bureaucracy has to do with it the better.”

13 包括以下词典：罗存德.英华音韵字典集成：第六版[M].上海：商务印书馆，1906；颜惠庆.英华大辞典[M].上海：商务印书馆，1920；严恩椿，沈宇.世界英汉汉英两用辞典[M].上海：世界书局，1933；汪倜然.综合英汉新辞典[M].上海：世界书局，1935.

14 该词其他译法包括以下三种：官僚政治、分部政治、专权之政、部曹繁设制度；官僚政府；官僚、有司、官吏等。

15 见https://www.oed.com/view/Entry/24905?redirectedFrom=bureaucracy&。

16 见https://www.oed.com/view/Entry/365523?redirectedFrom=bureaucratism&。

17 见https://www.oed.com/view/Entry/108603?isAdvanced=false&result=2&rskey=uGFMdO&。

18 见https://www.oed.com/view/Entry/100006?isAdvanced=false&result=2&rskey=T6RePl&。

参考文献

NEWMARK P. A textbook of translation [M]. New York: Prentice-Hall International, 1988.

SPIRA I. A conceptual history of Chinese-isms: the modernization of ideological discourse, 1895-1925 [M]. Leiden: Brill, 2015.

TOURY G. Descriptive translation studies and beyond [M]. Amsterdam: John Benjamins, 2012.

常晓宏. 鲁迅作品中的日语借词[M]. 天津：南开大学出版社，2014.

陈力卫.“主义”概念在中国的流行及其泛化[J]. 学术月刊，2012（9）：144-154.

程镇球. 翻译论文集[C]. 北京：外语教学与研究出版社，2002.

甘永龙. 日本在高丽之进步[J]. 东方杂志，1911（5）：4-6.

高名凯，刘正埮. 现代汉语外来词研究[M]. 北京：文字改革出版社，1958.

顾江萍. 汉语中的日语借词研究[M]. 上海：上海辞书出版社，2011.

胡毅美. 鲁迅作品中日语借词的个案研究——以“主义”为例[J]. 长春理工大学学报（社会科学版），2012（5）：93-95.

黄河清. 近现代辞源[M]. 上海：上海辞书出版社，2010.

黄友义，黄长奇，丁洁. 重视党政文献对外翻译，加强对外话语体系建设[J]. 中国翻译，2014（3）：5-7.

梁启超. 论师范[N]. 时务报（第十五册），1896.

刘凡夫. 以黄遵宪《日本国志》（1895）为语料的日语借词研究[J]. 日语学习与研

究，2012（3）：10-18.

刘正埮，高名凯，麦永乾，等. 汉语外来词词典[M]. 上海：上海辞书出版社，1984.

罗竹风. 汉语大词典：第一卷[M]. 上海：上海辞书出版社，1986.

毛泽东. 毛泽东选集：第一卷[C]. 北京：人民出版社，1991a.

毛泽东. 毛泽东选集：第二卷[C]. 北京：人民出版社，1991b.

潘卫民，卜海丽.《毛泽东选集》英译过程与价值研究[J]. 湘潭大学学报（哲学社会科学版），2013（6）：17-19.

石欣玉，黄立波. 毛泽东著作英译与国家形象建构：基于语料库的考察[J]. 外语教学，2021（3）：75-81.

王汎森. 思想是生活的一种方式[M]. 北京：北京大学出版社，2018.

巫和雄.《毛泽东选集》英译研究[M]. 北京：中国社会科学出版社，2013.

香港中国语文学会. 近现代汉语新词词源词典[M]. 上海：汉语大词典出版社，2001.

相贺徹夫. 时国语大词典[M]. 东京：小学馆，1974.

徐永焕. 关于英语毛泽东选集稿再次修改问题[J]. 财经，2006（24）：144-145.

余又荪. 日译学术名词沿革（续）[J]. 文化与教育旬刊，1935（70）：14.

中共中央文献研究室. 毛泽东早期文稿[C]. 长沙：湖南出版社，1990.

中国社会科学院语言研究所词典编辑室. 现代汉语词典（第七版）[Z]. 北京：商务印书馆，2016.

斎藤毅，明治のことば：東から西への架け橋[M]. 东京：讲坛社，1977.

诸丞亮，栾培琴. 毛泽东言语辞典[Z]. 济南：山东人民出版社，1993.

竹内实. 毛泽东集：1—10卷[C]. 东京：苍苍社，1983.

竹内实. 毛泽东集补卷：1—10卷[C]. 东京：苍苍社，1986.

通信地址：710061　陕西省西安市　西安外国语大学外国语言文学研究院

国内外财经文本分析研究综述*

北京外国语大学 牛华勇 窦一轩 夏晓雪

提要：财经领域的过往研究基于计量模型的因果推断，随着大数据与计算机算法的日渐成熟，以文本信息为代表的非结构化数据已可量化并应用到财经领域的研究中。文本信息中的语言特征，如文本可读性、文本语调和文本相似度等逐渐成为学者研究量化的重点，从而使文本分析技术应用到财经领域的研究中。本文从文本语言特征和财经领域的不同研究问题着手，对当前国内外文献进行了梳理，分析文本信息与财务信息之间的关系并指出文本分析技术未来在财经领域的发展方向，为相关领域研究者提供参考。

关键词：文本分析、财经领域、文本信息、文本语言特征

1 引言

财经领域的研究问题与变量设定多围绕数据库统计整理的结构化数据展开，进行以计量模型为基础的因果推断。随着计算机编程算法与大数据技术的发展，非结构化数据中蕴含的信息逐渐得到挖掘，如文本、音视频等。相比结构化数据，非结构化数据来源和形式更加多样，文本信息更是具有体量增长速度极快、时频高等特点（沈艳等，2019），而中文文本在语法与语义上与外国语言文本有较大区别，由于具有“听话听音，听锣听声”的语境特点（林乐、谢德仁，2016），中文文本信息的可挖掘性更强，涌现出许多值得研究的问题，但也具备较大的解构难度。借助计算机语言分析技术，如词典法、词袋法、主题分析法、自然语言处理技术等，将文本信息与文本特征量化为文本数据，是解决研究问题的主要途径。

文本分析研究可追溯至文本语言特征的建构，如文本可读性的定义（Dale & Chall 1949）和基于词频统计和手工计算构建的分析指标，如迷雾指数（Fog Index）（Gunning 1952），Kincaid指数（Flesch 1948）等。随着计量经济学的发展和统计技术的完善，文本分析开始应用于因果推断，但受限于技术水平，多为统计词频进行分析，如通过统计股吧留言与评论，预测股票价格走势（Antweiler & Frank 2004），进行上市公司年报信息的研究综述与相关性分析（Jones & Shoemaker

* 窦一轩为本文通讯作者。
作者贡献：
牛华勇：选题构思、字数占比（40%）、修改润色；
窦一轩：数据分析、讨论结论、初稿撰写、字数占比（40%）；
夏晓雪：研究方法、数据收集、字数占比（20%）。

1994），利用市场变量构建文本指数（Baker & Wurgler 2006；Qiu & Welch 2006）。计算机技术的发展以及大数据应用的成熟，使机器学习方法逐渐成为文本分析的重要工具。词典法、词袋法、自然语言处理技术等利用人工智能手段进行不同特征语言归类的手段开始显现，如Harvard-IV词典对年报情绪语调的整理归类（Tetlock 2007；Tetlock *et al.* 2008），Loughran & McDonald（2011）根据“10-K”文本整理出金融学领域的情绪词典——LM词典。国内学者也在LM词典的基础上，结合中文的语法特征与情绪表达方式，创建了适合中国金融市场使用的情绪词典（姜富伟等 2021；姚加权等 2021）。其他算法应用，如Campbell（2014）首次将LDA模型应用在文本分析当中，从相关文本中提取并构建风险指标，Mikolov *et al.*（2013）提出的Word2vec技术也逐渐应用到文本分析中来（Gentzkow *et al.* 2019）。

无论是词频统计还是模型算法，在财经文本分析的研究中，都需要提取一定语言特征进行因果推断或进行预测。本文将简要梳理财经文本分析中常用的文本语言特征，并探讨不同的研究领域中语言特征如何扮演变量角色。

2　文本语言特征

财经领域文本分析的已有文献多从文本语言特征展开，包括文本可读性、文本情绪语调、文本相似度、文本语义特征等，综述类文献也一般以文本语言特征为线索进行整理分类（Gentzkow *et al.* 2019；姚加权等 2020）。

2.1　文本可读性

文本可读性是描述文本信息的阅读难易程度的指标，反映了受众获取文本信息并能够复现的程度（Dale & Chall 1949；McLaughlin 1969）。同一信息下，文本可读性的不同会使读者的理解产生分歧，从而影响最终决策。当前，衡量文本可读性的指标一般有三种：迷雾指数（Gunning 1952）、文件大小和平实英语指标（Loughran & McDonald 2014；Bonsall *et al.* 2017；马长峰等 2020）。

2.2　文本情绪语调

文本情绪语调反映了文本呈现的作者态度，包括观点、喜好、情感等。根据一定的情感词典建立情绪语调指标，可用以判断作者态度是积极还是消极。当前研究中，文本情绪语调主要用于预测研究对象未来价值走向，文本来源较为丰富，相比文本可读性指标，研究关注的范围更广，包括媒体报道文本情绪（Antweiler & Frank 2004；Tetlock 2007；Jegadeesh & Wu 2013；汪昌云、武佳薇 2015）、电视电话会议音视频语调（Larcker & Zakolyukina 2012；林乐、谢德仁 2017；王

靖一、黄益平 2018）、社交网络文本（Chen *et al.* 2014；Renault 2017）、年报语调（Loughran & McDonald 2011；曾庆生等 2018；底璐璐等 2020）。

2.3　文本相似度

文本相似度指两个文本在遣词造句或表达含义上的相似程度（姜富伟等 2021），在财经领域多用于分析不同企业之间披露文本，如财务报告之间的相似程度（Hoberg & Phillips 2010；宋建波、冯晓晴 2022），或者同一企业不同年份的披露文本增量信息的多寡（Brown & Tucker 2011；葛锐等 2020）。文本相似度的衡量指标一般可分为三种：N-gram相似度（王贤明等 2013）、余弦相似度（Hoberg & Phillips 2016；张勇、殷建 2022）和深度学习模型（Suprapto & Polela 2020）。

2.4　文本语义特征与情感倾向

文本的语义特征体现为不同语言中词法句法的不同使用。例如文本采用的不同时态可以作为时间标记判断文本主体对于当下和未来在认知上的距离（Chen 2013；Kim *et al.* 2021），人称代词的使用体现出性别特征，从而反映出公司是否存在性别角色区分（Santacreu-Vasut *et al.* 2014；Abdelfattah *et al.* 2021）、人称代词的省略与否能够表明文本正式程度与权力的距离，在经济决策中表现出不同的参考价值（Licht *et al.* 2007）。

文本情感倾向主要表现为文本的语气强度和用词的使用语境，如具体性用词（Elliott *et al.* 2015）、不确定性用语（Loughran & McDonald 2013）、极端性（Bochkay *et al.* 2020）、生动性（Hales *et al.* 2011）、自发性（Lee 2016）、正式性（Rennekamp & Witz 2021）、自我指示性（Asay *et al.* 2018）等。虚拟语气、祈使句等句式的使用强度反映了文本主体对现实事务的掌握程度，当语言中存在虚拟语气时，个体将会因为感知到更多的不确定性而增加风险规避的倾向（Kovacic & Orso 2018）。

3　文本分析在财经领域的应用

基于以上文本语言特征，文本分析在财经领域的应用可以细分为宏观经济与政策分析、金融市场分析、会计信息分析与组织行为分析等。现有针对具体研究问题进行综合整理的文献（沈艳等 2019；刘云菁等 2021）未能将经济金融与会计应用领域完整梳理，而文本语言特征与研究问题也并非一一对应，故本文将在前人的基础上根据不同领域的研究问题进行进一步整理和扩充。

3.1　宏观经济与政策分析

财经领域的已有文献，多采用文本词频统计、文本语调、文本可读性等语言

特征进行政策分析，并拓展到宏观经济运行趋势。从具体研究问题角度可以分为政策不确定性研究和预测经济周期两方面。

3.1.1 政策不确定性指数构建与应用

Baker *et al.*（2016）利用美国新闻媒体数据，通过相关词典进行统计，建立描述经济政策不确定性的EPU指数，并进行领域拓展。EPU指数构建方法的优化与实践成为应用文本分析技术进行政策不确定性研究的重要领域。Tobback *et al.*（2018）基于SVM模型构建了比利时经济不确定性指数。Azqueta-Gavaldón（2017）基于LDA模型简化了构建EPU指数步骤。中国政策不确定性指标方面，已有文献多是基于《南华早报》构建EPU指数，Jurado *et al.*（2016）构建了中国金融不确定性指数。Bakas *et al.*（2016）根据欧洲市场文本研究了政策不确定性如何影响劳动力部门转移；其他也有诸如利用EPU指数研究政策不确定性对股票价格（Brogaard & Detzel 2015）、企业投资决策（Gulen & Ion 2016）等的影响。国内学者对于中国政策不确定性的影响分析较少，将EPU指数作为研究指标构建模型，研究变量关系的文献更为丰富。顾夏铭等（2018）研究政策不确定性对企业创新的影响，丁亚楠、王建新（2021）发现经济政策不确定性整体上降低了企业信息披露质量。

3.1.2 预测经济周期

文本分析技术为经济周期的预测提供了新的解决思路。现有文献对于利用文本语言特征预测经济周期主要从两个方面展开，包括构建经济周期指数进行预测和探索文本语调与经济周期的相关性。Thorsrud（2019）结合挪威新闻文本与GDP增长率构建经济增长指数并使用LDA模型进行分类预测；Kelly *et al.*（2021）提出HDMR模型，从央行沟通文本库中提取央行沟通测度指标，预测经济核心变量；文本情绪语调与经济运行周期的关系方面，Shapiro *et al.*（2020）基于美国16家主流经济金融媒体的新闻数据构建情绪指数，并展开文本情绪语调与经济运行情况的相关性分析。

3.2 金融市场分析

文本分析在金融市场研究中的应用主要在央行政策沟通、股票价格波动与投资决策以及金融市场指数构建与应用三个方面。

3.2.1 央行政策沟通

中央银行向市场传递货币政策目标与规则、经济形势判断以及前瞻性指引等信息由于在预期管理中具有重要作用，因此具备预测宏观经济的潜力。Hansen & McMahon（2016）应用LDA模型，对美国联邦公开市场委员会（FOMC）会议内容进行文本分类并提取文本语调变量，探索其对金融市场是否存在持续性影响；Cieslak *et al.*（2019）对FOMC公告效应的研究发现其对股票市场超额收益率

的影响是周期性的。国内对于央行政策沟通的文本分析研究比较深入，林建浩等（2021）基于文本数据的高维稀疏建模，引入央行沟通文本进行经济预测，发现能够提高预测精度。王琳、刘宏雅（2022）的研究发现，央行沟通、投资者情绪与股市波动之间存在动态时变关系，央行沟通能够有效调节投资者情绪，投资者情绪与股市波动之间呈现明显正向效应。

3.2.2　股票价格投资者决策

政策信息、新闻等文本信息能够反映股票市场的价格波动，而企业财务报告和社交媒体文本能够体现投资者的情绪与决策倾向。已有文献中，利用文本语言特征构建风险指数分析预测股票价格波动和利用文本语调分析投资者情绪的研究较为丰富。Kumar *et al.*（2022）基于Twitter评论的情绪语调，建立DFA-DBN模型进行估价预测，取得了更好的预测效果；顾文涛等（2020）将财经新闻文本加入金融情绪词典，改善了金融市场收益率预测效果；崔炎炎、刘立新（2022）利用情感分类模型提取金融科技相关股票投资者情绪指标，发现投资者情绪对金融科技类股票收益率预测具有重要作用。

3.2.3　金融市场指数构建

基于文本分析的金融市场指数主要包括以下几类：关注度指数、情绪指数、金融风险指数等。关注度指数构建主要从投资者和新闻媒体两个角度进行。Da *et al.*（2011）首次使用网页搜索次数衡量关注度，以股票代码为关键字构建投资者关注度指数；Tsukioka *et al.*（2018）基于雅虎财经论坛各公司下的发帖数量构建投资者关注度指数，发现投资者关注度与日本公司IPO抑价现象有关。石勇等（2017）基于股吧评论数据构建不同平台的关注度指数，包括投资者与新闻媒体，并建立VAR模型进行关注度与沪深300指数的相关性分析，结果表明投资者关注度对股市影响较大，新闻媒体关注度影响较小。

构建情绪指数的文本来源与关注度指数较为接近，对基于公司财务报告的文本语调构建出的情绪指数，本文将在第三部分详述。金融情绪指数构建的一个重要工具是金融情绪词典。哈佛大学通用调查词典（GI）在2000年公开之后，学者们不断创新金融情绪词典，包括Harvard-IV词典、LM词典等。Loughran & McDonald（2011）还提出了TF-IDF法，为计算量化金融文本特征提供了一种新的思路。Garcia（2013）应用LM词典构建情绪指数，发现其在经济下行时预测股票价格效果较好。Huang *et al.*（2015）使用偏最小二乘法（PLS）构建了投资者情绪指数，Petropoulos & Siakoulis（2021）基于XGBoost算法和NLP方法构建金融情感指数分析央行政策语调对经济波动的影响。

3.3　会计信息分析

关于会计信息的文本分析，文本来源基本为公司财务报告。针对不同的文本

语言特征，对于会计信息的研究方向也各不相同，其中较为突出的是会计信息形式质量。会计信息形式质量是以特定语言和呈报方式准确、清晰、简明地传递会计信息的程度，故文本可读性、情绪语调、文本相似度等文本语言学特征成为会计信息形式质量的重要体现（杨丹等 2018）。整体来看，文本分析主要集中在会计信息形式质量与财务欺诈、投资与预测等方面。

3.3.1　会计信息形式质量与财务欺诈

上市公司的财务报告可能利用不同程度的用词，达到报表粉饰的效果，出现管理层操纵乃至财务欺诈，获得更多盈利的倾向。而随着文本挖掘与大数据技术的发展，财务欺诈的手段也更为丰富（Amani & Fadlalla 2017），对识别管理层操纵与财务欺诈提出了更高挑战。会计信息形式质量的一个重要体现是文本可读性。具有较低信息质量的财务报告不仅信息披露不全，文字上也会采取隐晦的描述方式来粉饰意图。Li（2008）和Biddle *et al.*（2009）使用迷雾指数度量企业会计信息披露质量，得到一组对偶结论，发现可读性高、信息质量高的企业更能提高投资效率，具有较低盈利水平的企业，通过降低可读性进行掩饰。Lo *et al.*（2017）发现管理层会降低年报文本可读性以迷惑读者，且该特征与盈余水平高度相关。徐巍等（2021）则在中文环境下构建了衡量中文年报可读性的指标，用以分析我国上市企业的信息是否存在财务欺诈现象。

3.3.2　会计信息形式质量、投资与预测

体现会计信息形式质量的文本语言特征不仅影响投资者的投资决策，对分析师预测公司价值，构建和改进企业风险指标亦有较强的参考价值。有学者发现企业财务报告的可读性影响着投资者的情绪反应与决策结果（Miller 2010；Rennkamp 2012），可读性越强，反衬了管理层乐观积极的情绪（Li 2010），对于会计信息形式质量更高的企业，散户投资意愿也更强。Lehavy *et al.*（2011）和Bassemir *et al.*（2013）分别通过年报和电视电话会议文本进行研究，发现可读性越差，分析师预测偏离程度越大。刘建秋等（2022）基于信号理论和迎合理论，发现企业社会责任报告正面语调能够降低分析师预测分歧和偏差，同时也降低了公司发生股价崩盘的风险。

3.3.3　其他问题

对于企业财务欺诈和投资决策影响的研究，也有学者从其他文本着手研究，强化了企业财务信息形式质量对财务欺诈和投资预测影响的结论。有学者进一步采用深度学习算法构建识别财务舞弊与财务欺诈的模型，如Lin *et al.*（2015）比较了多个人工智能算法识别企业财务欺诈的效果，发现决策树模型和人工神经网络明显更优。李哲、王文翰（2021）考察企业“多言寡行”的文本与行动特征是否影响其绿色信贷的获取，结果表明存在显著正向影响。谭建华、王雄元（2022）研究发现企业在出现财务违规后，其年报文本可读性、相似度显著下降，呈现异

常积极的情绪语调。

在风险管理指标与模型构建方面，Manela & Moreira（2017）根据《华尔街日报》的文本信息构建了新闻隐含波动率指数（NVIX），用以管理企业股价波动风险，预测股票投资回报率。李成刚等（2021）将MD&A文本语言特征信息加入信用风险评估模型，结果表明文本语言特征信息与企业信用风险显著相关。阮素梅等（2022）通过对MD&A文本的情感语调进行算法建模，发现RF与GBDT模型能够有效识别上市公司财务风险。

3.4　企业组织行为分析

国内外对组织行为领域的文本分析较多，但对该领域的整理与综述研究并不丰富，宋铁波等（2021）对文本分析在企业管理领域的研究进行了编码梳理。这一领域的文本数据来源主要是年报MD&A信息与高管公开发言，应用领域可以归纳为两个方面——管理层战略与公司内部治理。

3.4.1　管理层战略

公司年报中的情绪语调可以反映管理层对企业发展战略的态度。吴建祖、赵迎（2012）对公司年报文本进行定性分析，发现公司注意力集中在消费者时，公司倾向于选择多元化战略；MD&A的“短期视域”语言反映了企业管理层的短视主义观念，影响了企业的长期发展战略。Brochet *et al.*（2015）建立短视主义词典对美国盈余电话会议内容进行词语视域分类，发现管理层存在短视主义现象。国内学者针对中文“听话听音”的语境特点进行研究改进。胡楠等（2021）基于Word2vec技术进行改进，发现中国上市公司高管存在短视主义现象，影响了企业的长期投资。王新光（2022）进而发现企业管理层的短视主义抑制了企业的数字化转型。

3.4.2　公司内部治理

公司治理领域的文本分析应用较为分散，学界通过对企业年报的信息披露和实际采访，挖掘企业风险管理，人力资源管理与社会责任承担领域的数据信息与相关性。周婷婷、李维安（2016）分析企业年报中的非财务信息，发现信息环境变动较低时，对企业的风险评估更为准确。McKenna *et al.*（2016）则针对企业不同性别与职级的员工采访文本，量化性别特征分析不同性别员工的工作生活平衡度与性别歧视现象。张秀敏等（2016）首次将语义分析引入企业环境信息披露的研究中，尽管结论呈现出企业环境信息披露质量与环境规制和公众关注度的相关性，但缺乏财经领域与环境领域的相关词典补充，仍需要对企业环境责任承担能力进一步证明。

4 结语

本文梳理了文本分析在财经领域应用的文献，从文本语言特征来看，文本可读性、文本语调、文本相似度和文本语义特征与情感倾向是文本分析中常用来作为量化指标的参考。从不同的学科领域来看，文本分析在宏观经济、金融学、会计学和企业管理等领域均有丰富的研究，研究问题主要可以分为两方面，包括不同领域的相关性因果推断和指标构建与预测。本文从两个角度进行文献的整理，旨在帮助读者理解文本分析技术在财务领域的应用，针对不同学科领域的具体问题提供不同语言特征的应用思路。

文本分析始于文本语言特征和简单指标的构建，随着计量经济学的发展形成变量参与因果推断或形成知识图谱。随着计算机技术的发展，大数据算法逐渐成为文本分析的重要工具，参与到各学科领域的研究当中。经济学、金融学的传统研究范式基于计量经济学的因果推断模型和检验分析，有赖于研究者组织一手资料或者二手资料形成结构化的数据变量。文本、音视频等信息在计算机算法的帮助下，不仅能够量化为数据变量参与到结构化数据分析中，还能够重新验证甚至拓展原有的研究范式与理论。

当前已有文献证实，文本信息与财务信息之间存在密切的联系。财经领域研究的因果推断显示，文本语调、文本可读性等语言特征，同时也是会计信息形式质量的体现，对企业财务欺诈倾向，投资者意愿和分析师预测精度等都具有显著的影响作用。同时，提取文本信息特征构建的文本指标在财务预测模型中能够提高预测效果也体现了文本信息对财务信息挖掘的补充，能够提高企业风险管理能力。然而，文本信息由于其非结构化的特点，基于计算机算法构建的语言指标的可信度仍然值得检验，且在文本语言特征受到管理层关注下，企业财务报告粉饰作用更为突出，主观性因素反而影响到对财务信息的判断。因此，本文认为文本分析技术的发展方向仍然值得探索：一是多角度开源文本信息，达到相互补充和验证的作用，注重对文本数据的清洗和整理，同时要注重不同领域、不同国别文本语言的特殊性，比如创造不同领域的情感词典等，提高文本信息量化指标的信度和效度；二是创新文本信息指标量化的方法，将最新计算机深度学习算法模型应用到文本信息的量化中来，提高文本信息提取效率；三是更多地将文本信息等非结构化数据应用到结构化数据模型中，验证与完善原有的理论与模型，提高因果推断的准确度，扩大文本分析在财经领域的应用范围。

参考文献

ABDELFATTAH T, ELMAHGOUB A, ELAMER M. Female audit partners and extended audit reporting: UK evidence [J]. Journal of Business Ethics, 2021, 174(1): 177-197.

AMANI F, FADLALLA A. Data mining applications in accounting: a review of the literature and organizing framework [J]. International Journal of Accounting Information Systems, 2017, 24: 32-58.

ANTWEILER W, FRANK M. Is all that talk just noise? The information content of internet stock message boards [J]. The Journal of Finance. 2004, 59(3): 1259-1294.

ASAY H, LIBBY R, RENNEKAMP K. Do features that associate managers with a message magnify investors' reactions to narrative disclosures? [J]. Accounting, Organizations and Society, 2018, 68, 1-14.

AZQUETA-GAVALDÓN A. Developing news-based economic policy uncertainty index with unsupervised machine learning [J]. Economics Letters, 2017, 158: 47-50.

BAKAS D, PANAGIOTIDIS T, PELLONI G. On the significance of labor reallocation for European unemployment: Evidence from a panel of 15 countries [J]. Journal of Empirical Finance, 2016, 39(B): 229-240.

BAKER M, WURGLER J. Investor sentiment and the cross-section of stock returns [J]. The Journal of Finance, 2006, 61(4), 1645-1680.

BAKER S, BLOOM N, DAVIS S. Measuring economic policy uncertainty [J]. The Quarterly Journal of Economics, 2016, 131(4): 1593-1636.

BASSEMIR M, NOVOTNY-FARKAS Z, PACHTA J. The effect of conference calls on analysts' forecasts – German evidence[J]. European Accounting Review, 2013, 22(1): 151-183.

BIDDLE G, HILARY G, VERDI R. How does financial reporting quality relate to investment efficiency? [J]. Journal of Accounting and Economics, 2009, 48(2): 112-131.

BOCHKAY K, HALES J, CHAVA S. Hyperbole or reality? Investor response to extreme language in earnings conference calls [J]. The Accounting Review, 2020, 95(2), 31-60.

BONSALL S, LEONE A, MILLER B, RENNEKAMP K. A plain English measure of financial reporting readability [J]. Journal of Accounting and Economics, 2017, 63(2-3): 329-357.

BROCHET F, LOUMIOTI M, SERAFEIM G. Speaking of the Short-Term: disclosure horizon and managerial myopia[J]. Review of Accounting Studies, 2015, 20: 1122-1163.

BROGAARD J., DETZEL A. The asset-pricing implications of government economic policy uncertainty [J]. Management Science, 2015,61(1): 3-18.

BROWN S, TUCKER J. Large-sample evidence on firms' year-over-year MD&A modifications [J]. Journal of Accounting Research, 2011,49(2): 309-346.

CAMPBELL J, CHEN H, DHALIWAL D, et al. The information content of mandatory

risk factor disclosures in corporate filings [J]. Review of Accounting Studies, 2014,19(1): 396-455.

CHEN H, DE P, HU Y, HWANG B. Wisdom of crowds: the value of stock opinions transmitted through social media [J]. The Review of Financial Studies, 2014, 27(5): 1367-1403.

CHEN M. The effect of language on economic behavior: evidence from savings rates, health behaviors, and retirement assets [J]. American Economic Review, 2013,103(2): 690-731.

CIESLAK A, MORSE A, VISSING-JORGENSEN A. Stock returns over the FOMC cycle[J]. The Journal of Finance, 2019, 74(5): 2201-2248.

DA Z, ENGELBERG J, GAO P. In search of attention[J]. The Journal of Finance, 2011,66(5): 1461-1499

DALE E, CHALL J. Techniques for selecting and writing readable materials[J]. Elementary English, 1949,26(5):250-258.

ELLIOTT W, RENNEKAMP K, WHITE B. Does concrete language in disclosures increase willingness to invest?[J]. Review of Accounting Studies, 2015, 20(2): 839-865.

FLESCH R. A new readability yardstick[J]. Journal of Applied Psychology, 1948, 32(3): 221-233.

GARCIA D. Sentiment during recessions [J]. The Journal of Finance, 2013, 68(3): 1267-1300.

GENTZKOW M, KELLY B, TADDY M. Text as data [J]. Journal of Economic Literature, 2019,57(3): 535-574.

GULEN H, ION M. Policy uncertainty and corporate investment [J]. The Review of Financial Studies, 2016, 29(3):523–564.

GUNNING R. Technique of clear writing [M]. New York: McGraw-Hill, 1952.

HALES J, KUANG X, VENKATARAMAN S. Who believes the hype? An experimental examination of how language affects investor judgments [J]. Journal of Accounting Research, 2011, 49(1): 223-255.

HANSEN S, MCMAHON M. Shocking language: understanding the macroeconomic effects of central bank communication[J]. Journal of International Economics, 2016, 99: S114-S133.

HOBERG G, PHILLIPS G. Product market synergies and competition in mergers and acquisitions: a text-based analysis [J]. The Review of Financial Studies, 2010,23(10): 3773-3811.

HOBERG G, PHILLIPS G. Text-based network industries and endogenous product

differentiation[J]. Journal of Political Economy, 2016, 124(5): 1423-1465.

HUANG D, JIANG F, Tu J, et al. Investor sentiment aligned: a powerful predictor of stock returns [J]. The Review of Financial Studies, 2015, 28(3): 791-837.

JEGADEESH N, WU D. Word power: a new approach for content analysis [J]. Journal of Financial Economics, 2013, 110(3): 712-729.

JONES M, SHOEMAKER P. Accounting narratives: a review of empirical studies of content and readability[J]. Journal of Accounting Literature, 1994,13(1): 142.

JURADO K, NG S, LUDVIGSON S. Measuring uncertainty[J]. Operations Research: Management science, 2016, 56(3): 265-266.

KELLY B, MANELA A, MOREIRA A. Text selection [J]. Journal of Business & Economic Statistics, 2021, 39(4) : 859-879.

KIM J, KIM Y, ZHOU J. Time encoding in languages and investment efficiency [J]. Management Science, 2021, 67(4), 2609-2629.

KOVACIC M, ORSO C. Why do some individuals fear immigration more than others? Evidence from Europe[J]. Working Paper, 2018.

KUMAR S, AKEJI A, MITHUN T. Stock price prediction using optimal network based twitter sentiment analysis [J]. Intelligent Automation & Soft Computing, 2022, 33(2): 1217-1227.

LARCKER D, ZAKOLYUKINA A. Detecting deceptive discussions in conference calls [J]. Journal of Accounting Research, 2012, 50(2): 495-540.

LEE J. Can investors detect managers' lack of spontaneity? Adherence to predetermined scripts during earnings conference calls [J]. The Accounting Review, 2016, 91(1): 229-250.

LEHAVY R, LI F, MERKLEY K. The effect of annual report readability on analyst following and the properties of their earnings forecasts [J]. The Accounting Review, 2011,86(3): 1087-1115.

LI F. The information content of forward-looking statements in corporate filings: a naïve Bayesian machine learning approach [J]. Journal of Accounting Research, 2010, 48(5): 1049-1102.

LI F. Annual report readability, current earnings, and earnings persistence [J]. Journal of Accounting and Economics, 2008, 45(2-3): 221-247.

LICHT A, GOLDSCHMIDT C, SCHWARTZ S. Culture rules: the foundations of the rule of law and other norms of governance [J]. Journal of comparative economics, 2007, 35(4), 659-688.

LIN C, ANAN C, HUANG S. Detecting the financial statement fraud: the analysis of the differences between data mining techniques and experts' judgments [J]. Knowledge-

Based Systems, 2015, 89: 459-470.

LO K, RAMOS F, ROGO R. Earnings management and annual report readability [J]. Journal of Accounting and Economics, 2017, 63(1): 1-25.

LOUGHRAN T, MCDONALD B. When is a liability not a liability? Textual analysis, dictionaries, and 10-Ks[J]. The Journal of Finance, 2011, 66(1): 35-65.

LOUGHRAN T, MCDONALD B. IPO first-day returns, offer price revisions, volatility, and form S-1 language [J]. Journal of Financial Economics, 2013, 109(2): 307-326.

LOUGHRAN T, MCDONALD B. Measuring readability in financial disclosures [J]. The Journal of Finance. 2014,69(4):1643-1671.

MANELA A, MOREIRA A. News implied volatility and disaster concerns [J]. Journal of Financial Economics, 2017, 123(1): 137-162.

MCKENNA B. VERREYNNE M. WADDELL N. Locating gendered work practices: a typology [J]. International Journal of Manpower, 2016, 37(6): 1085-1107.

MCLAUGHLIN G. SMOG grading: a new readability formula [J]. Journal of Reading, 1969, 12(8): 639-646.

MIKOLOV T, SUTSKEVER I, CHEN K, et al. Distributed representations of words and phrases and their compositionality [J]. Advances in Neural Information Processing Systems, 2013: 3111-3119.

MILLER B. The effects of reporting complexity on small and large investor trading[J]. The Accounting Review, 2010, 85(6): 2107-2143.

PETROPOULOS A, SIAKOULIS V. Can central bank speeches predict financial market turbulence? Evidence from an adaptive NLP sentiment index analysis using XGBoost machine learning technique [J]. Central Bank Review, 2021, 21(4): 141-153.

QIU L, WELCH I. Investor sentiment measures [J]. National Bureau of Economic Research Working Paper, 2006, No. w10794.

RENAULT T. Intraday online investor sentiment and return patterns in the US stock market [J]. Journal of Banking & Finance, 2017, 84: 25-40.

RENNEKAMP K, WITZ P. Linguistic formality and audience engagement: investors' reactions to characteristics of social media disclosures [J]. Contemporary Accounting Research, 2021, 38(3): 1748-1781.

RENNKAMP K. Processing fluency and investors' reactions to disclosure readability [J]. Journal of Accounting Research, 2012, 50(5): 1319-1354.

SANTACREU-VASUT E, SHENKAR O, SHOHAM A. Linguistic gender marking and its international business ramifications[J]. Journal of International Business Studies, 2014, 45(9): 1170-1178.

SHAPIRO A, SUDHOF M, WILSON D. Measuring news sentiment[J]. Journal of Econometrics, 2020,228 (2): 221-243.

SUPRAPTO, POLELA J. The influence of loss function usage at SIAMESE network in measuring text similarity [J]. International Journal of Advanced Computer Science and Applications (IJACSA), 2020, 11(12): 787-792.

TETLOCK P. Giving content to investor sentiment: the role of media in the stock market [J]. The Journal of Finance, 2007, 62(3): 1139-1168.

TETLOCK P, SAAR - TSECHANSKY M, MACSKASSY S. More than words: quantifying language to measure firms' fundamentals [J]. The Journal of Finance, 2008, 63(3): 1437-1467.

THORSRUD L. Words are the new numbers: a newsy coincident index of the business cycle [J]. Journal of Business & Economic Statistics. 2019: 393-409.

TOBBACK E, NAUDTS H, DAELEMANS W. Belgian economic policy uncertainty index: improvement through text mining [J]. International Journal of Forecasting, 2018, 34(2) : 355-365.

TSUKIOKA Y, YANAGI J, TAKADA T. Investor sentiment extracted from internet stock message boards and IPO puzzles [J]. International Review of Economics and Finance, 2018, 56: 205-217.

崔炎炎，刘立新.网络舆情赋能金融科技股票收盘价预测研究[J].统计研究，2022（6）：148-160.

底璐璐，罗勇根，江伟，等.客户年报语调具有供应链传染效应吗？——企业现金持有的视角[J].管理世界，2020（8）：148-163.

丁亚楠，王建新.“浑水摸鱼”还是“自证清白”：经济政策不确定性与信息披露——基于年报可读性的探究[J].外国经济与管理，2021（11）：70-85.

葛锐，刘晓颖，孙筱蔚.审计师更换影响管理层报告信息增量了吗?——来自纵向文本相似度的证据[J].审计研究，2020（4）：113-122.

顾文涛，王儒，郑肃豪，等.金融市场收益率方向预测模型研究——基于文本大数据方法[J].统计研究，2020（11）：68-79.

顾夏铭，陈勇民，潘士远.经济政策不确定性与创新——基于我国上市公司的实证分析[J].经济研究，2018（2）：109-123.

胡楠，薛付婧，王昊楠.管理者短视主义影响企业长期投资吗？——基于文本分析和机器学习[J].管理世界，2021（5）：139-156.

姜富伟，孟令超，唐国豪.媒体文本情绪与股票回报预测[J].经济学，2021（4）：1323-1344.

李成刚，贾鸿业，赵光辉，付红.基于信息披露文本的上市公司信用风险预警——来自中文年报管理层讨论与分析的经验证据[J].中国管理科学，2021（4）：1-14.

李哲，王文翰.“多言寡行”的环境责任表现能否影响银行信贷获取——基于“言”和“行”双维度的文本分析[J].金融研究，2021（12）：116-132.

林建浩，陈良源，罗子豪，等.央行沟通有助于改善宏观经济预测吗？——基于文本数据的高维稀疏建模[J].经济研究，2021（3）：48-64.

林乐，谢德仁.投资者会听话听音吗？——基于管理层语调视角的实证研究[J].财经研究，2016（7）：28-39.

林乐，谢德仁.分析师荐股更新利用管理层语调吗？——基于业绩说明会的文本分析[J].管理世界，2017（11）：125-145.

刘建秋，尹广英，吴静桦.企业社会责任报告语调与分析师预测：信号还是迎合?[J].审计与经济研究，2022（3）：62-72.

刘云菁，张紫怡，张敏.财务与会计领域的文本分析研究：回顾与展望[J].会计与经济研究，2021（1）：3-22.

马长峰，陈志娟，张顺明.基于文本大数据分析的会计和金融研究综述[J].管理科学学报，2020（9）：19-30.

阮素梅，杜旭东，李伟，等.数据要素、中文信息与智能财务风险识别[J].经济问题，2022（1）：107-113.

沈艳，陈赟，黄卓.文本大数据分析在经济学和金融学中的应用：一个文献综述[J].经济学，2019（4）：1153-1186

石勇，唐静，郭琨.社交媒体投资者关注、投资者情绪对中国股票市场的影响[J].中央财经大学学报，2017（7）：45-53.

宋建波，冯晓晴.关键审计事项信息含量与公司债券发行定价——基于文本相似度视角[J].会计研究，2022（3）：174-191.

宋铁波，陈玉娇，朱子君.量化文本分析法在国内外工商管理领域的应用对比与评述[J].管理学报，2021（4）：624-632.

谭建华，王雄元.上市公司违规与年报文本信息操纵[J].中国软科学，2022（3）：99-111.

汪昌云，武佳薇.媒体语气、投资者情绪与IPO定价[J].金融研究，2015（9）：174-189.

王靖一，黄益平.金融科技媒体情绪的刻画与对网贷市场的影响[J].经济学，2018（4）：1623-1650.

王琳，刘宏雅.央行沟通能否有效应对突发事件“大考”——基于中国人民银行沟通事件的文本分析[J].北京理工大学学报（社会科学版），2022（1）：77-89.

王贤明，胡智文，谷琼.一种基于随机n-Grams的文本相似度计算方法[J].情报学报，2013（7）：716-723.

王新光.管理者短视行为阻碍了企业数字化转型吗——基于文本分析和机器学习的经验证据[J].现代经济探讨，2022（6）：103-113.

吴建祖，赵迎.高层管理团队注意力对企业多元化战略选择的影响——基于中国上市公司的实证分析[J].经济与管理研究，2012（9）：107-113.

徐巍，姚振晔，陈冬华.中文年报可读性：衡量与检验[J].会计研究，2021（3）：28-44.

杨丹，黄丹，黄莉.会计信息形式质量研究——基于通信视角的解构[J].会计研究，2018（9）：3-10.

姚加权，张锟澎，罗平.金融学文本大数据挖掘方法与研究进展[J].经济学动态，2020（4）：143-158.

姚加权，冯绪，王赞钧，等.语调、情绪及市场影响：基于金融情绪词典[J].管理科学学报，2021（5）：26-46.

曾庆生，周波，张程，等.年报语调与内部人交易：“表里如一”还是“口是心非”？ [J].管理世界，2018（9）：143-160.

张秀敏，汪瑾，薛宇. 语义分析方法在企业环境信息披露研究中的应用[J].会计研究，2016（1）：87-94.

张勇，殷健.会计师事务所联结与企业会计政策相似性——基于TF-IDF的文本相似度分析[J].审计研究，2022（1）：94-105.

周婷婷，李维安.信息环境波动与董事会风险功能——基于风险信息披露视角[J].经济与管理研究，2016（5）：105-112.

通信地址：100089　北京市　北京外国语大学国际商学院

学习者英语动词配价型式使用特征研究——以agree为例*

河南师范大学 孙海燕 牛文爽

提要：语料库短语学视域下的配价型式能清晰展现词项的句法、语义限制，帮助学习者习得词汇知识。本文以高频动词agree为例，基于语料库数据对比中国学习者和本族语者的配价型式使用特点。研究结果显示学习者和本族语者使用的配价型式在频数分布上差异显著，此外学习者使用的配价型式存在多种误用现象且行动元语义类型缺乏多样性，这表明学习者尚未完全掌握agree的配价型式变化。本文从母语迁移、课堂输入等角度分析学习者误用配价型式的原因，并建议通过课堂数据驱动学习模式以及编纂配价型式词典的方式将配价型式应用于词汇教学。

关键词：语料库、配价型式、词汇教学

1 引言

语料库短语学打破了传统语言认知，用大量真实数据证实词汇、语法、意义互相选择，密不可分，并取得了一系列研究成果，如搭配（collocation）（Sinclair 1991）、词束（lexical bundle）（Biber *et al.* 2004）、搭配框架（collocational framework）（Renouf & Sinclair 1991）、型式语法（pattern grammar）（Hunston & Francis 2000）、扩展意义单位（extended unit of meaning）（Sinclair 2004）等。这些成果将语言的基本意义单位由传统的单个词汇扩展至多词序列，揭示了语言的短语特征。配价型式作为语料库短语学的新成果，把型式语法和传统的配价语法结合起来，既解决了型式语法因缺少句法功能标注而产生的歧义问题，又避免了传统配价语法基于研究者个人经验，性质、分类难以统一的缺陷，是一种简洁、完善的词汇语法型式描写框架。目前，有关配价型式的研究多以本族语语料为研究对象，聚焦于型式描写（刘国兵、杜亚平 2017）、近义动词辨析（孙海燕、柳雪莹 2019）、学术英语词汇（刘国兵、张孝莲 2021）等领域，针对学习者语言的研究尚不多

* 本文系河南省高校人文社会科学研究一般项目“共选理论视域下学习者英语短语习得模式构建”（2021-ZZJH-179）的阶段性成果。孙海燕为本文通讯作者。
作者贡献：
孙海燕：选题构思、研究方法、讨论结论、字数占比（60%）、修改润色；
牛文爽：数据收集、数据分析、初稿撰写、字数占比（40%）。

见。鉴于配价型式能够清晰展现词项的句法、语义限制，帮助学习者习得词汇知识，本研究以常用动词agree为例，基于语料库数据探究中国英语学习者动词配价型式的使用特征，分析学习者和本族语者使用的配价型式在类型、频数和搭配词方面的差异，以期为外语词汇教学提供借鉴。

2　文献综述

2.1　型式语法

型式这一概念最早由Hornby（1954：v）提出，他认为较之语法规则，学习者更应关注英语的实际用法，因此他们需要掌握英语句子的型式并了解词汇与型式的对应关系。Hornby列出了25种动词型式、4种名词型式和3种形容词型式。以动词型式为例，“He proved them wrong.”中proved的型式可描写为verb+noun+adjective。Hornby用型式描写代替句法分析，为语言描写提供了一种全新视角。Sinclair（1991）基于对大量语料库数据的观察，提出型式与意义互相关联。具体来说，如果一个词项具有多种意义，那么每种意义倾向于出现在特定的型式中，因此我们可以通过型式来判定多义词项的具体意义。Francis（1993，1995）继承并发展了Sinclair的观点，她指出不仅词汇通过选择特定的型式来表达意义，一个型式也会选择共享某种意义的词汇集。Hunston & Francis（2000）基于前人研究发展出一套语料库驱动的英语词汇语法——型式语法，并列出了常见动词、名词、形容词的各种型式。她们将型式定义为与某个词汇频繁共现且影响该词汇意义的所有单词和结构（Hunston & Francis 2000：37）。型式语法的具体做法是摒弃传统的句法功能标注，在描写体系中仅保留词性标签和高频出现的单词。例如，本文研究的单词agree的常用型式是V *with* n，V *on* n等。型式语法的理论主张可概括为“词汇语法不分家；意义型式相关联”（王勇 2008：259）。作为第一部语料库驱动的英语描写语法，型式语法揭示了语言的线性特征和短语倾向，具有简明、灵活的特点，受到英语教学和自然语言处理领域学者的广泛关注（Hunston 2002；Huang *et al.* 2011；Otero & López 2011；Ma & Qian 2020）。

2.2　配价语法

配价语法由法国语言学家Tesnière首创，目的是描述词汇的行为特点，即词汇是如何结合或者要求一定数量的补充语成分共同构成较大的语言单位，如短语或短句（甄凤超 2019：49）。Tesnière 认为动词作为句子的核心能够在其周围开辟一定数量的空位并要求其他词性的补充语予以填补，以构成合乎句法规则的语句，动词这种支配力即是动词的价（Allerton 1982：2）。具体来说，名词词组构成的补充语称作“行动元”，副词词组构成的补充语称作“状态元”。状态元的数量可以

是无穷的，而行动元的数量通常不超过三个。行动元的数量对应动词的价，支配一个行动元的动词即是一价动词，以此类推（邹友昌、刘万义 2000：8）。配价语法以具体词汇为切入点，清晰展现各句子成分间的支配依存关系，将局部的、词汇的语法与普遍语法联系起来（Teubert 2007：225）。配价理论诞生后被陆续引介到德语、英语、汉语等语言的语法研究中，研究视角从句法层面扩展至逻辑和语义层面，研究对象由动词扩展到名词和形容词（Heringer 1993；Gao & Liu 2019；韩万衡 1993）。近年来，除理论建构外，配价理论还被应用于词典编纂、语言教学等领域（Herbst *et al.* 2004；Zhao & Jiang 2020；陆俭明 1997；邵菁 2002）。

2.3 配价型式

型式语法采用具体的词和词性描写语言结构，具有简洁、灵活的优势，但由于其摒弃了传统的句法功能范畴，有时会引起歧义。例如，短语the hatred of a million coolies的型式可描写为N *of* n，这一型式有两种解读，“苦力”既可以是仇恨者也可以是被仇恨的对象（Teubert 2007：224）。为解决上述问题，Teubert（2007）曾建议把配价语法中的句法功能标签引入型式语法。Reichardt（2014）采纳了这一建议，把型式语法和配价语法结合起来，对比分析了动词consider及其德语翻译对等单位的配价句子型式。甄凤超、杨枫（2015）基于Reichardt的研究发展出一套完整的语料库驱动的英语动词配价型式描写框架。该描写框架以特定词项为中心，以语料库数据为依据，在配价型式的描写框架中既运用具体的词和词性，也使用必要的句法功能标签并辅之以搭配、语义倾向等分析手段。配价语法和型式语法是两种不同的语法框架，将二者结合起来的依据在于：（1）二者都强调词汇与语法的统一，都基于语言的线性规则描述词汇的共现，在学理上是相通的（甄凤超 2019：50）；（2）型式语法仅使用词性标签，无法体现词汇间的支配依存关系，而配价语法作为一种依存语法，其中的句法范畴恰好能弥补这一缺陷（Teubert 2007）；（3）传统的配价语法研究基于内省数据并且引入了复杂的语义格，价的概念和性质难以界定（袁毓林 2010），而型式语法以大量的语料库数据为支撑，以频数为判定型式的主要依据，研究结果具有客观性。若将型式语法的判定标准引入配价语法，便可保证价的统一（甄凤超、杨枫 2015）。配价型式将型式语法和配价语法的优势结合起来，实现了词汇、语法和意义的统一，为英语词汇教学提供了一种新方法。本文以常用动词agree为例，对比分析学习者和本族语者使用的配价型式，旨在探究学习者配价型式使用中的问题，为词汇教学提供参考。

3 研究设计

本文探究以下两个问题：（1）动词agree在本族语语料库LOB和学习者语料库CLEC中分别有哪些配价型式？（2）中国学习者和本族语者使用的配价型式有

哪些差异？产生差异的原因有哪些？

本文选取agree作为研究对象的原因有两点：一是agree的使用频率较高；二是agree的用法多样，尽管词典和一些语法书对其进行了详细描述，但许多学生在实际运用中依然会出现误用介词搭配和动词不定式等问题。研究语料取自LOB 和CLEC。LOB的库容约为一百万词，包含500篇2,000词左右的英式英语书面文本。CLEC（中国学习者英语语料库）包含高中生习作、大学英语四级、六级和英语专业四级、八级考试作文，可以较为全面地反映各阶段学生的语言特征，且库容与LOB相当，与之具有可比性。具体研究步骤如下：（1）在LOB中提取以agree为谓语动词的索引行，确定agree的配价型式。（2）在CLEC中重复上述步骤，得到学习者使用的配价型式。（3）分析学习者与本族语者使用的配价型式在类型、频数、行动元等方面的异同。（4）探究学习者与本族语者存在差异的原因，讨论如何将配价型式应用于词汇教学。

4　结果与讨论

4.1　本族语语料库中的动词配价型式

利用AntConc软件在LOB中检索动词agree，经人工删除agree作非谓语动词的索引行后，共得到有效索引行85条。该动词的配价型式见表1。

表1　LOB中动词agree的配价型式

配价型式	频数	占比	例句
			一价型式
Sub V	13	15.29%	We **agree**.
			二价型式
Sub V Obj-*that*	22	25.88%	Any teacher will **agree** that it is impossible to pursue both lines effective during a single year.
Sub V *with* Obj	22	25.88%	And I **agree** with it.
Sub V vb-*to*-inf	7	8.24%	So will you not **agree** to spend the winter in Las Palmas, Miss Barclay, and dry Pepita's tears?
Sub V *to* Obj	6	7.06%	Councils **agree** to merger plan.
Sub V *on* Obj	4	4.71%	The Anglican and main Protestant communions readily **agree** on many questions.

（待续）

（续表）

配价型式	频数	占比	例句
Sub V *upon* Obj	2	2.35%	Unless the countries of the Security Council **agree** upon it.
Sub V *in* Obj	2	2.35%	This observation shows that means and medians do not necessarily **agree** in the conclusions they yield.
Sub V *at* Obj	1	1.18%	The ordinal scale means of objects DEO and AGP are 5 and 6, while the interval scale means **agree** at the value.
三价型式			
Sub V *with* Obj clause-*that*	5	5.88%	I **agree** with the Prime Minister that I do not think we are necessarily bound for federalism in Europe.
Sub V *with* Obj *in* vb (*ing*)	1	1.18%	The government **agree** with the Royal Commission in thinking that the boundaries and status of the City of London should remain unchanged.
总计	85	100.00%	

在配价型式的描写体系中，V代表核心动词；Sub和Obj分别代表主语行动元和宾语行动元；vb是verb complement的缩写，表示动词性补足语；vb-*to*-inf表示由to引导的不定式短语。用斜体标出的是配价型式中出现的具体单词或字母串，如clause-*that*表示由that引导的小句成分。此外，不同动词使用的配价型式标注信息会有所差异。从表1可以看出，动词agree有11种配价型式，其中一价型式1种，二价型式8种，三价型式2种。最常用的三种配价型式是：（1）Sub V *with* Obj（25.88%），在这一型式中agree支配两个名词性成分，一个为主语，另一个为宾语；（2）Sub V Obj-*that*（25.88%），该型式中agree支配一个作主语的名词性成分和一个由that引导的宾语从句；（3）Sub V（15.29%），此型式为一价，agree只支配一个作主语的名词性成分。值得注意的是，《牛津词典》（Hornby 2014）中描述的典型配价型式Sub V *about* Obj和Sub be V-ed *on*/*about* nom未在此语料库中出现，这一差异在一定程度上反映了语料库数据的必要性。

配价型式强调结构与意义的关联，语言研究的最终目的也在于意义的表达，因此本文分析了动词agree的行动元语义类型，结果见表2。由表2可知，一价型式的行动元多为人称代词，如I、you、we等。二价型式的行动元语义类型较为丰富，包括职业类、观点类、方法类，如 editor、contention、approach等。三价型式的行动元主要表示组织机构，如government、commission等。上述分析表明，特定的配价型式会反映其行动元的语义特征，词汇、语法和意义互相选择（卫乃兴

2012）。需要补充的是，配价型式与行动元的互选是一种倾向性，二者之间没有绝对的对应关系，不同配价型式的行动元可能在语义上重叠。

表2　动词agree在LOB中的主要行动元

配价型式	语义类型	行动元
一价	人称类	I、you、we、they
	职业类	doctor、teacher、psychologist、editor、epicure
二价	观点类	view、contention、proposal、statement、conclusion
	方法类	approach、principle、policy
三价	组织类	government、commission

4.2　学习者语料库中的动词配价型式

学习者语料库具有数据代表性强、分析效率高的优势，可帮助研究者全面了解中介语发展特点（Granger *et al.* 2015）。本节基于学习者语料库CLEC分析中国学生的动词配价型式使用特征。在CLEC中agree作为谓语动词共出现107次，其具体配价型式见表3。

表3　CLEC中动词agree的配价型式

配价型式	频数	占比	例句
一价型式			
Sub V	5	4.67%	Do you **agree**?
二价型式			
Sub V *with* Obj	52	48.60%	To me I **agree** with this view.
Sub V Obj-*that*	18	16.82%	So we all **agree** that haste makes waste.
Sub V *to* Obj	12	11.21%	I also **agree** to this opinion.
*Sub V Obj	5	4.67%	I **agree** her words.
Sub V *on* Obj	4	3.74%	I can't **agree** on this opinion, because...
Sub V vb-*to*-inf	4	3.74%	We all hope you will **agree** to come if you have time.
Sub V *at* Obj	1	0.93%	And I also don't **agree** at cigarette smoking.

（待续）

（续表）

配价型式	频数	占比	例句
*Sub V clause-*to*	1	0.93%	In my opinion, I **agree** to the haste makes waste.
*Sub V *with* vb	1	0.93%	As I maybe not **agree** with study here, or my way is…
三价型式			
Sub V *with* Obj clause-*that*	2	1.87%	I think all of you will **agree** with me that the boy is innocent.
*Sub V *with* Obj *on* nom	2	1.87%	I **agree** with a number of my fellow members on their views.
总计	107	100%	

注：*表示CLEC中独有的配价型式。

从整体上看，学习者和本族语者共享多数配价型式，二者最常使用的两种配价型式均为Sub V *with* Obj和Sub V Obj-*that*。但二者在具体型式的使用频数上存在差异，学习者呈现出过多使用二价型式，尤其是Sub V *with* Obj这一型式的倾向。笔者使用SPSS软件对LOB和CLEC中动词agree的配价型式分布情况进行卡方检验，结果显示二者存在显著差异（$x^2 = 7.781$，$P = 0.020 < 0.05$），学习者二价型式使用过度，一价型式使用不足（见表4）。造成这一差异的原因可能是教师在日常词汇教学中多以词典、语法书和自身经验为依据，注重介绍固定搭配和短语，容易忽视较为简单的一价型式。

表4　LOB和CLEC中agree的配价型式分布情况

配价型式	语料库	
	LOB	CLEC
一价型式	13（15.29%）	5（4.67%）
二价型式	66（77.65%）	98（91.59%）
三价型式	6（7.06%）	4（3.74%）
合计	85（100%）	107（100%）

逐条观察CLEC中的索引行可以发现，学习者和本族语者使用的配价型式不仅在频数分布方面具有显著差异，而且在质量方面也有较大差别。学习者使用的4种配价型式Sub V Obj，Sub V clause-*to*，Sub V *with* vb，Sub V *with* Obj *on* nom未

在本族语语料库中出现，属于误用现象，下文列出了相应的例句进行错误分析。

Sub V Obj

（1）I **agree** it, every thing has its own stages and procedures.

（2）But I am very regret that I can't **agree** the period you offer.

Sub V clause-*to*

（3）In my opinion, I **agree** to the haste makes waste.

Sub V *with* vb

（4）As I maybe not **agree** with study here, or my way is not right, or I'm very lazy, so my results fell behind other classmates.

Sub V *with* Obj *on* nom

（5）I **agree** with a number of my fellow members on their views.

（6）I don't **agree** with George Orwell on this point.

例句（1）和（2）中，动词agree后紧跟一个代词或一个由定语从句修饰的名词，属于介词缺失错误。此类错误可能由母语负迁移导致。汉语属于意合语言，借助词语或句子间的逻辑实现连贯，而英语属于形合语言，依靠词汇和形态手段实现词语和句子的连接（马绪光 2010：112）。如果学生不了解这一差异，便容易在写作中出现介词、连词等缺失现象。例（3）使用二价型式Sub V clause-*to*，在核心词项后用介词to引导宾语从句，不符合小句需用连词引导这一语法规则。例（4）在介词with后直接使用动词原形study，而本族语者一般使用现在分词。例（5）和例（6）存在介词错误，用on代替了本族语者通常使用的in。此外，在CLEC中，学习者与本族语者共享的配价型式也存在一些错误。比如，例（7）中学生用介词on引导人物类行动元，而本族语者通常使用with。例（8）中学生使用to+V-ing代替动词不定式，属于非谓语动词错误。总的来看，学习者的错误多集中在介词方面。张会平、刘永兵（2013）的研究表明汉语概念迁移是影响学习者介词习得的重要因素，他们指出大量的目的语输入能够解决由语言负迁移造成的介词偏误。

（7）Personally, I don't **agree** on those people discribling above wholely.

（8）So I don't completely **agree** to studying abroad.

依据Nation（1990）提出的词汇知识框架，外语学习者不仅要识记词汇的发音、拼写，知晓词汇的含义、使用频率和语法行为，还应掌握词汇的语域特征和

搭配习惯。因此本文分析了学习者与本族语者使用的配价型式在名词搭配词即行动元方面的差异。首先，学习者与本族语者使用的行动元语义类型不同，学习者主要使用人物、人称和观点类行动元，而本族语者使用的行动元包括人称、职业、观点、方法和组织类。其次，虽然本族语者和学习者均使用了观点类行动元，但后者的词汇丰富度低，所用词汇局限于opinion、view、viewpoint等。产生上述差异的原因可能有两点：一是CLEC中的语料为高中至大学的考试作文或限时习作，题材多以议论文为主，因此需较多使用表达个人观点或立场的语句；二是学习者在写作时为避免错误，倾向于选择形式简单、习得较早的安全词汇。最后，由表5可知，不同于本族语者把配价型式与特定行动元联结起来的特点，学生使用的三类配价型式共享大多数行动元，这表明他们缺乏型式与意义互相选择的短语意识。

表5　动词agree在CLEC中的主要行动元

配价型式	语义类型	行动元
一价	人称类	I、you、they
二价	人称类	I、you、we、me、him、them、someone、everybody
	人物类	people、Chinese、classmate、parents、readers
	观点类	idea、opinion、viewpoint、point of view
三价	人称类	I、you、me
	观点类	view、point

依据上述对LOB和CLEC中动词agree使用状况的分析，笔者发现许多学生虽然在学习初期就接触了动词agree，但并未完全习得其用法。总体上看，学习者使用的配价型式呈现以下特征。首先，在配价类型上，虽然学习者与本族语者共享多数配价型式，但部分配价型式仅在学习者语料库中出现，属于中介语特有的现象。其次，在频数分布上，学习者存在过度使用二价型式的问题。再次，在准确性方面，学习者使用的配价型式出现较多错误，包含遗漏或混淆介词、误用动词原形和非谓语动词等。最后，在行动元的语义类型方面，较之本族语者，学习者不仅使用的语义类型丰富度低，而且没有建立起型式与语义的关联。以上特征反映出学生的短语能力欠缺，而短语能力是二语学习者语用能力和语言综合能力的重要组成部分，理解语言的短语倾向对外语学习具有重要意义（黄开胜、周新平2016：27）。为促进学生习得短语知识、发展短语能力，本文建议将配价型式应用于英语词汇教学。

4.3　配价型式的词汇教学应用方法

第一，采用课堂数据驱动学习模式。课堂数据驱动学习是一种以语料库资源为基础的外语学习法。其主要内容是教师引导学生在事先准备好的语料库数据中自主发现语言规律。这种方式具有以下优点：（1）创造真实语言环境，帮助学生理解和记忆词汇；（2）强调以学生为中心的发现式学习，培养学生自主学习能力；（3）学生不直接操作计算机，不受技术限制；（4）语料经教师筛选和整理，具有系统性和针对性（刘芹、可庆宝 2020）。将配价型式引入词汇教学可采用课堂数据驱动学习模式。以agree的教学为例，备课时教师可先从语料库中提取大量索引行，接着参考配价型式词典总结出agree的典型配价型式为Sub V Obj-*that*和Sub V *with* Obj，最后筛选、整理出符合学生水平的例句作为教学材料。课堂上，教师既可以把配价型式和索引行提供给学生，让他们进行匹配练习，也可以引导学生观察索引行，自主归纳agree的配价型式及相应的搭配词，如二价型式的常见搭配词可分为职业、观点和方法类。课后教师可依据学生的水平布置相应的作业，比如用特定的配价型式造句或分析学习者语料中的配价型式误用现象等。此外，教师还可以建立小型学习者语料库，以掌握学生配价型式运用中的问题，并进行针对性教学。

第二，编写适合学习者的英语配价型式词典。配价理论在德国取得的成就最大，目前被广泛应用于对外德语教学和德语学习词典编纂。英语虽然与德语存在差异，但同样可用配价型式描写句法结构，由德国学者Thomas Herbst等人编纂的《英语配价词典》（*A Valency Dictionary of English*）就是很好的例证（甄凤超、杨枫 2016）。但这部词典所使用的标记复杂，例证中的难词多，对读者的英语水平要求高（张爱朴、周流溪 2007）。因此，为把配价型式引入词汇教学，可编写一部适合学习者的配价型式词典。词典可收录大型通用语料库中的高频动词、名词和形容词。每个词条中不仅列出该词的所有配价型式以及对应的常用搭配词，而且提供相应例句和必要的中文释义。通过查阅这种词典，学习者不仅能够系统了解词汇的配价型式，而且可以通过具体的词汇语境和语法语境理解词汇意义。

5　结语

本文基于LOB和CLEC中的真实语料，以常用动词agree为例考察中国学习者英语动词配价型式的使用特征。研究分析了学习者和本族语使用的配价型式在类型、频数分布和行动元语义方面的异同。结果表明，学习者和本族语者使用的配价型式虽然在类型上相似，但是在使用频率方面差异显著，学习者存在过度使用二价型式的问题。在行动元语义方面，学习者不仅使用的行动元语义类型较为单一，而且没有建立起配价型式与意义的关联。此外，通过深入分析索引行可以发

现，学生使用配价型式的准确性较低，存在多种误用现象，比如遗漏或混淆介词、误用非谓语动词形式等。上述差异和误用可能由母语负迁移和短语知识欠缺等因素导致。为帮助学生习得短语知识，本文建议通过采用课堂数据驱动学习模式和编写适用于学习者的配价型式词典的方式将配价型式应用于英语词汇教学。

参考文献

ALLERTON D. Valency and the English verbs [M]. New York: Academic Press, 1982.

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.

FRANCIS G. A corpus-driven approach to grammar: principles, methods and examples [C]//BAKER M, FRANCIS G, TOGNINI-BONELLI E. Text and technology: in honor of John Sinclair. Amsterdam: John Benjamins, 1993: 137-156.

FRANCIS G. Corpus-driven grammar and its relevance to the learning of English in a cross-cultural situation [C]//PAKIR A. English in education: multicultural perspectives. Singapore: Unipress, 1995: 7-34.

GAO J, LIU H. Valency and English learners' thesauri [J]. International Journal of Lexicography, 2019, 32(3): 326-361.

GRANGER S, GILQUIN G, MEUNIER F. The Cambridge handbook of learner corpus research [M]. Cambridge: Cambridge University Press, 2015.

HERBST T, HEATH D, ROE I, et al. A valency dictionary of English: a corpus-based analysis of the complementation patterns of English verbs, nouns and adjectives [M]. Berlin: Mouton de Gruyter, 2004.

HERINGER H. Dependency syntax – basic ideas and the classical model [C]//JACOBS J, VON STECHOW A, STERNEFELD W, et al. Syntax (Volume 1). Berlin: Walter de Gruyter, 1993: 298-316.

HORNBY A. A guide to patterns and usage in English [M]. Oxford: Oxford University Press, 1954.

HORNBY A. Oxford advanced learner's English-Chinese dictionary (8th edition) [Z]. Oxford: Oxford University Press, 2014.

HUANG C, CHEN M, HUANG S, et al. EdIt: a broad-coverage grammar checker using pattern grammar [C]//Proceedings of the ACL-HLT 2011 System Demonstrations. Portland: Association for Computational Linguistics, 2011: 26-31.

HUNSTON S. Pattern grammar, language teaching, and linguistic variation [C]// REPPEN R, FITZMAURICE S, BIBER D. Using corpora to explore linguistic variation. Amsterdam: John Benjamins, 2002: 167-183.

HUNSTON S, FRANCIS G. Pattern grammar: a corpus-driven approach to the lexical

grammar of English [M]. Amsterdam: John Benjamins, 2000.

MA H, QIAN M. The creation and evaluation of a grammar pattern list for the most frequent academic verbs [J]. English for Specific Purposes, 2020, 58: 155-169.

NATION I. Teaching and learning vocabulary [M]. New York: Newbury House, 1990.

OTERO P, LÓPEZ I. A grammatical formalism based on patterns of part of speech tags [J]. International Journal of Corpus Linguistics, 2011, 16(1): 45-71.

REICHARDT R. Valency sentence patterns and meaning interpretation [D]. Birmingham: University of Birmingham, 2014.

RENOUF A, SINCLAIR J. Collocational frameworks in English [C]//AIJMER K, ALTENBERG B. English corpus linguistics: studies in honour of Jan Svartvik. Harlow: Longman, 1991: 128-144.

SINCLAIR J. Corpus, concordance, collocation [M]. Oxford: Oxford University Press, 1991.

SINCLAIR J. Trust the text: language, corpus and discourse [M]. London: Routledge, 2004.

TEUBERT W. Sinclair, pattern grammar and the question of hatred [J]. International Journal of Corpus Linguistics, 2007, 12(2): 223-248.

ZHAO Q, JIANG J. Verb valency in interlanguage: an extension to valency theory and new perspective on L2 learning [J]. Poznań Studies in Contemporary Linguistics, 2020, 56(2): 339-363.

韩万衡. 配价论的基本概念与研究方法[J]. 天津外国语学院学报，1993（00）：22-32.

黄开胜，周新平. 基于语料库的中国英语学习者词块输出能力的趋势研究[J]. 外语界，2016（4）：27-34.

刘国兵，杜亚平. 语料库驱动视角下的英语动词配价结构研究——以APPOINT为例[J]. 天津外国语大学学报，2017（6）：1-7.

刘国兵，张孝莲. 语料库驱动视角下学术英语动词搭配配价研究[J]. 外语电化教学，2021（1）：105-111.

刘芹，可庆宝. 数据驱动学习在学术英语词汇教学中的应用[J]. 当代外语研究，2020（1）：58-67.

陆俭明. 配价语法理论和对外汉语教学[J]. 世界汉语教学，1997（1）：4-14.

马绪光. “形合”、“意合”与英汉翻译的句法策略[J]. 上海师范大学学报（哲学社会科学版），2010（1）：112-117.

邵菁. 配价理论与对外汉语词汇教学[J]. 语言教学与研究，2002（1）：43-49.

孙海燕，柳雪莹. 配价结构视角下动词近义词对比分析[J]. 外语与翻译，2019（2）：80-86.

王勇. 行走在语法和词汇之间——型式语法述评[J]. 当代语言学，2008（3）：257-266.

卫乃兴. 共选理论与语料库驱动的短语单位研究[J]. 解放军外国语学院学报，2012（1）：1-6.

袁毓林. 汉语配价语法研究[M]. 北京：商务印书馆，2010.

张爱朴，周流溪.《英语配价词典》的特色与不足[J]. 辞书研究，2007（3）：81-91.

张会平，刘永兵. 英语介词学习与概念迁移——以常用介词搭配与类联接为例[J]. 外语教学与研究，2013（4）：568-580.

甄凤超. 语料库驱动的短语配价型式研究[M]. 上海：上海交通大学出版社，2019.

甄凤超，杨枫. 语料库驱动的学习者英语动词配价研究：以CONSIDER为例[J]. 外国语，2015（6）：57-67.

甄凤超，杨枫. 配价结构及搭配配价在英语词汇教学中的应用：思想和方法[J]. 外语界，2016（4）：35-42.

邹友昌，刘万义. 配价语法与配价语义补偿[J]. 解放军外国语学院学报，2000（3）：8-11.

通信地址：453000　河南省新乡市　河南师范大学外国语学院

基于语料库的美国媒体中国人口话语建构研究*

佛山开放大学 王 琴

提要：本文运用话语历史分析与语料库语言学相结合的研究方法，以美国媒体关于中国人口问题的相关报道为语料，从报道主题、话语策略、历史和社会语境三个维度进行分析。研究发现：美国媒体运用了命名、谓述、视角化和辩论策略建构“自我”和“他者”两大对立阵营，其对我国人口问题的态度经历了从“全盘否定和批评”，过渡到“观望和质疑”，最后演变为“勉强认可”的演变过程。本研究可为我国本土政策的对外传播提供借鉴。

关键词：人口话语、话语历史分析、语料库

1 引言

2021年5月31日，中共中央政治局召开会议指出，进一步优化生育政策，实施一对夫妻可生育三个子女政策及配套支持措施。这标志着我国正式迈入“三胎时代”。人口生育政策的调整以及鼓励生育配套措施的相继出台引起社会各界热议，也成为国内外学者的研究热点。

近年来，我国人口政策研究特点可以大致概括如下：研究数量呈井喷式上升；研究领域广泛，以社会学、公共管理学为主，还涉及法学、经济学、新闻传媒学等；从单一学科的理论研究发展到跨学科交叉融合研究，定量和定性相结合的研究也不断涌现；研究质量有所提升。但是，目前从批评话语分析视角探讨我国人口政策话语建构的研究还不多见。

本研究旨在运用语料库语言学（简称CL）和批评话语分析（简称CDA）的三大主流流派之一——话语历史分析方法（简称DHA），考察美国主流媒体过去九年间（2013年5月至2021年12月）涉及“中国人口政策”的相关报道，探究美国媒体如何建构我国人口政策，如何建构正面积极的“自我”和负面消极的“他者”，并结合历史和社会语境，描述话语的动态演变轨迹，揭示美国媒体的政治阴

* 本文系佛山市社科联2022年度社科项目“基于语料库的中国人口政策外媒话语建构历时研究”（2022-GJ151）的阶段性成果。

谋和话语霸权。本课题拟回答以下研究问题：

（1）美国媒体话语的主要话题是什么？

（2）美国媒体使用的主要话语策略是什么？在文本中如何体现？

（3）美国媒体对于“中国人口政策”的态度在九年间经历了怎样的历时变化？导致态度变化的主要原因是什么？

2 研究背景

2.1 文献概述

2.1.1 政治话语研究

目前，政治话语有三条主要研究路径。Chilton（2004）从认知科学和认知语言学的理论视角研究语言与政治的关系。他将政治行动看作是言语行为，聚焦行为者表征现实的方式。此外，他尤其关注合作与冲突的关系问题。Fairclough & Fairclough（2012：21）吸收了亚里斯多德的“论辩术”思想，认为政治的本质就是“在通过协商取得的决策基础上实现最高利益的追求”。他们主张通过论证和协商理论来解释“合作与冲突”之间的关系。Reisigl & Wodak（2009：91）主张运用DHA方法分析政治话语，认为政治是众多行动场域的集合，每一个独立的行动场域都与一组政治亚体裁相关。尽管DHA对政治话语的分类法被指具有“微观原子论”的弊病，但它是唯一提出连贯统一分析框架的话语分析流派，而且分析工具可操作性强，因而也成为政治话语研究者广泛使用的一种研究方法。

2.1.2 语料库语言学与批评话语分析方法的融合研究

欧洲学者们对CL和CDA两种研究方法的互补性和各自优势已经作了较为详尽的阐释（Partington 2003；O’Halloran & Coffin 2004；Baker *et al.* 2008；Mautner 2009）。Partington（2003：12）认为，语料库技术有助于强化、修正或驳斥研究者的直觉。Baker *et al.*（2008：285）也指出，采用语料库技术有助于对CDA关注的语言现象进行量化分析；CDA的理论框架可用于阐释CL的量化数据分析结果和研究发现。

2008年，Baker和兰卡斯特大学研究团队以英国媒体中的移民话语建构为例，概括梳理了语料库辅助批评话语“九步分析法”的分析步骤。他们指出，CL和CDA两种方法都可以作为研究的切入点，形成良好的循环通路（Baker *et al.* 2008：295）。“三角验证法”使研究者行走于数据、文本和语境之间，定量和定性方法相互渗透和融通，边界变得模糊，有助于展示话语的“全貌”，从而降低CDA的主观性，避免了研究者对文本过度解读或是阐释不够的情况发生（O’Halloran & Coffin 2004）。

“三角验证”研究方法历经十多年的发展，已逐渐被广泛应用于政治话语研究。Baker *et al.*（2008）调查了英国报纸中有关移民和难民（RASIM）的话语建构，比较大报和小报的话题倾向和差异以及话语的历时变化。Liu & Lin（2021）以“语境重构”理论为切入点，比较中、英、美三国主流媒体关于“一国两制”的报道，分析了“语境重构”的话语表征方式的差异及动因。

以上研究聚焦社会问题、政治体制或弱势群体，但研究对象均未涉及具体政策。Gomez-Jimenez（2018）是为数不多的触及具体政策的应用研究，他借鉴语料库技术和van Dijk的话语语义学相关理论，比较了产假制度在不同时期英国报纸中的话语表征和态度变化，揭示了英国社会不平等的阶级关系。但是他的研究仅侧重话语命题的语义，缺少对具体词汇语法资源和话语策略的分析，且未充分考虑话语的语境。

2.1.3 语料库语言学与话语历史分析方法的融合研究

迄今，将语料库的量化统计分析方法与话语历史分析的定性研究方法融合的研究为数不多。Engstrom & Paradis（2015）以英国政党官网上的文章和政策文件为语料，针对英国各政党在建构“群内”和“群外”方面的异同点进行共时和历时分析。杨敏、符小丽（2018）以“希拉里邮件门事件”为例，验证了语料库语言学和话语历史分析在分析政治话语中的价值和有效性。杨敏、侍怡君（2021）聚焦美国政府官网上“中美贸易战”的报道，分析事件的历史渊源和社会情境，揭示美方将贸易战“合法化”的手段和策略。

综上，前人研究均借鉴语料库语言学和话语历史分析的理论和分析框架，聚焦西方国家的政治事件，但对中国的政治事件或具体政策关注不足，开展我国人口政策的话语建构研究显得十分必要。

2.2 理论框架

2.2.1 话语历史分析学派的话语观

Wodak倡导的话语历史分析方法与van Dijk的社会认知模型以及Fairclough提出的辨证关系法，并称为当今批评话语分析领域的三大主流学派。话语历史分析是由Wodak带领其团队成员在分析1986年奥地利总统选举运动的反犹话语后创建的流派。Reisigl & Wodak（2009：95）认为话语是“处在特定行动场域中，由语境决定的意义集合”。话语在社会实践中形成，反映并建构社会现实。与宏观话题的相关性、多视角和论证性被视为话语的三个核心构成要素。话语不是一个封闭的单位，而是不断被话语分析者重新阐释的动态意义实体。

话语历史分析继承了法兰克福学派和Harbermas的语言哲学，以及Bourdieu的

社会学思想，发展了社会批判理论，提出将“文本内部批评、社会诊断性批评、前瞻性批评”三维度整合的批判话语观（Reisigl & Wodak 2009：94）。

话语历史分析方法的研究框架包括三个维度：语篇的主题、话语策略维度、互文和互语关系维度。话语策略包括命名策略、谓述策略、辩论策略、视角化策略、增强/削弱策略。话语历史分析方法重视对话语历史语境的考察，追踪一段时期内话语的历时变化，因而具有跨学科性、历史性、文本性和实践性的特征（杨敏、侍怡君 2021）。

2.2.2　话语、权力和意识形态

所谓权力，就是处于不同社会地位、归属于不同社会团体的社会行为者之间的不平等关系。权力在话语中被“合法化”或“去合法化”。权力在话语中不仅通过语法形式和结构得以实现，还可以通过人们对文本体裁和社会情境的控制，或是对某种公共场域的准入进行规范和管控来实现（Reisigl & Wodak 2009：95）。

Fairclough（1989）发现，媒体权力具有隐蔽性，被统治阶级所操纵。Thompson（1990）认为，意识形态是霸权符号在社会中得以传播的社会过程和形式。话语历史分析学派主张，意识形态是由特定社会团体成员共享的，带有偏向性的世界观或价值观，它由与此相关的心理表征形式，如观点、态度、评价和信念等构成。意识形态通过话语建立和维持不平等的权力关系，譬如在话语中表现为霸权身份叙事或是充当公共话语的“把关人”。此外，它还不同程度地改变着现有的权力关系。语篇是社会斗争的场所，是不同利益取向的意识形态争相夺取霸权和支配地位的场所。话语历史分析旨在揭示话语中的霸权、歧视和排外等问题，通过解码那些建立、维持或反对压迫的意识形态，实现对特定话语霸权“去神秘化”，提升大众对话语中强行嵌入的政治操纵、霸权思想和价值理念等的敏感性（Reisigl & Wodak 2009：94）。

3　研究方法

3.1　研究语料

本研究语料来自LexisNexis Academic新闻数据库，以population policy、*child policy为关键词搜索自“单独两孩”政策出台（2013年年底）至“全面三孩”政策出台（2021年5月31日）期间，《纽约时报》（NYT）、美联社（AP）、美国有线电视新闻网（CNN）和美国新闻和世界报导杂志（U.S. News & World Report）的报道。为了更全面地收集数据，将搜索时间跨度前后各延长半年，并人工删除其他国家人口政策的相关报道，最终搜索到2013年5月至2021年12月的报道共计128篇，基于此建立一个134,841词次的目标语料库，简称AMC（American Media

Corpus）。依据参照语料库必须大于目标语料库五倍大小的原则，选取Lancsbox V6.0软件自带的American English Corpus（AME06）作为参照语料库，简称AEC，共500个文件，1,001,024词次。依据报道生成的年份将目标语料库分为9个子库，由此得出美媒近9年报道数量的分布趋势（如图1所示）。

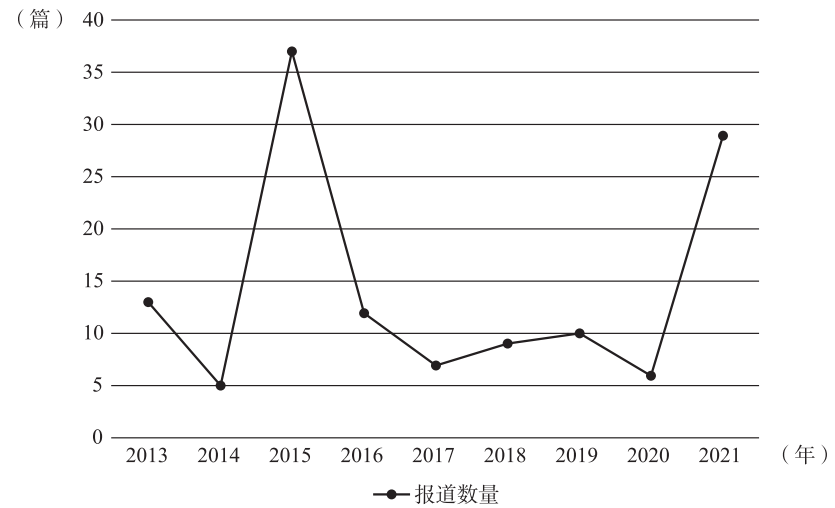


图1 AMC中每年中国人口问题报道数量分布情况

如图1，折线图线条的高低起伏表示美媒对中国人口政策关注程度的变化。两个凸起峰值期分别出现在2015年和2021年。2015年是“全面两孩”政策的转折点，媒体的关注度较前期的两孩政策调整试点期呈现显著上升趋势，接下来五年内关注度明显回落，并徘徊在较低水平；2021年由于“三孩政策”正式落地，报道数量又出现明显上升的趋势。

3.2 研究方法

为了提高研究效度，本研究采用基于语料库的方法，即定量和定性相结合的“三角验证法”。本研究使用的语料库分析软件是兰卡斯特大学研究团队开发的Lancsbox V6.0版本，可利用该软件的词频统计和主题词（Words）、索引行（KWIC）、主题词搭配网络可视化（Graphcoll）、主题词的语料库分布（Whelk）、文本上下文等技术工具，借助智慧检索功能，对语料文本进行高效的量化统计分析。

3.3 研究步骤

研究步骤分为以下四步。首先，通过主题词确定话语主题；其次，考察主题词的搭配网络；接着，人工分析索引行所处的单个文本的话语策略，并结合语料

库的KWIC和Graphcoll功能，分析特定词汇语法或者话语策略的使用实例；最后，结合语料库检索和统计结果考察话语的动态演变轨迹，并结合社会历史语境，阐释媒体话语动态演变的动因。

4 分析与讨论

4.1 确定报道主题

利用Lancsbox V6.0软件对目标语料库和参照语料库进行对比，提取目标语料库中位居前20位的高频词（功能词除外）。位居前20位的高频主题词，按照语义域可分为四类：（1）政策行为者，包括国家、政党、机构或个人等行为者，譬如China、Chinese、government、communist、women、child；（2）政策内容本身，如one-child、policy、population、fertility、birth、planning、family、aging、birthrate、demographic等；（3）政策影响的范围或地点，如Beijing、China；（4）其他类，如姓氏Wang。

总体而言，高频主题词分析表明美国媒体主要围绕中国政府生育政策调整的内容，以及对中国人口出生率和人口老龄化趋势的效果和影响进行报道。

4.2 话语策略分析

4.2.1 命名策略

DHA主张，命名策略就是用来指称社会行为者、物体、现象、事件、过程和行为的语言手段，语篇层面表现为指称语、成员范畴化名词或代词、隐喻、表示过程和行为的动词、名词等（Reisigl &Wodak 2009：93-94）。本文将着重讨论两类名词，即政策指称名词、与政策制定相关的社会行为者名词。

美国媒体对我国人口政策有七种不同的的命名方式（见表2）。其中，使用频次最高的是*child policy，用来专指不同时期特定的人口政策，譬如one-child policy指自20世纪70、80年代提倡的“一对夫妇只生一个孩子”的优生优育政策，three-child policy指2021年5月31日出台的“三孩”政策；family-planning policy、family planning policy的使用频次位居第二，在2015年前用来专指“一孩政策”，后期与population policy一同泛指所有试行实施的生育政策；birth control policy、birth-control policy与birth restriction policy、family size restrictions在语料中使用频次偏低，特指“一孩政策”，其中control、restrictions这类将抽象过程名词化的用法，既反映了美国媒体对我国特定历史阶段人口生育政策持有明显的消极负面态度，同时将信息打包（packing）成名词，以名词化的方式将政策“实物化”，也表明美国媒体向其目标受众传递这种负面评价的“事实性”和“客观性”。pro-fertility policy只出现在2021年的两篇报道中，专指放开“三孩”政策以后的“生

育友好型”政策。由此可见，美国媒体在政策的命名上既表现出多样性，同时也带有明显的态度和舆论导向，从2013年的“生育控制、生育限制”逐渐过渡到2021年的“生育友好”，暴露了美方媒体蓄意操纵目标受众认知的企图。

表2 命名频次及分布情况

命名名称	绝对频次（次）	相对频次（次/万词）	语料库中语篇占比
* child policy	456	33.82	118/128
family-planning/ family planning policy	31	2.30	19/128
population policy	12	0.89	12/128
birth-control/birth control policy	3	0.22	3/128
birth restricition policy	3	0.22	3/128
family size restrictions	3	0.22	3/128
pro-fertility policy	2	0.15	2/128

如图2所示，“policy”的搭配词分布在以节点词“policy”为“圆心”、以搭配强度为半径的圆周上，像磁场一样向四周呈发散状分布。搭配词距离圆心的距离越近，表示该词与节点词的搭配强度越强，则“磁场”效应越明显。搭配词所代表的黑点颜色越深，代表它与节点词共现搭配的频率越高，节点词对它的“引力”越强。

我们从搭配词靠近圆心的圆环上搜索到以下与政策制定和调整相关的行为者名词，它们与policy的搭配频率由高到低排列依次是China、China’s、government、party、policy-makers，见表3。

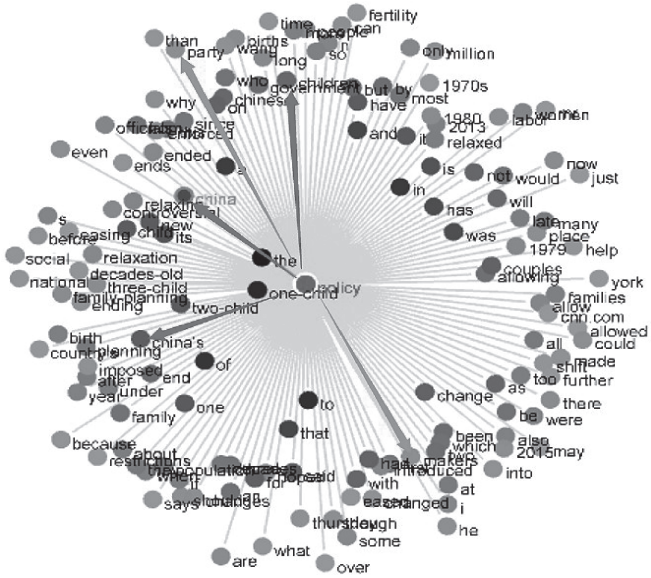


图2　policy搭配的“磁场效应”

表3　与policy搭配的社会行为者名词

序号	名词	绝对频次	相对频次（/万词）	语料库中语篇占比
1	China	108	8.01	57/128
2	China’s	62	4.60	40/128
3	government	39	2.89	29/128
4	party	13	0.96	9/128
5	policy-makers	13	0.96	9/128

我们借鉴系统功能语法的及物性过程的分类框架讨论这些行为参与者名词在小句中的功能成分，即参与者角色。China充当物质过程小句的施事行为者（Actor），China’s通常充当政策的所有者（Possessor），强调人口政策的地域归属。government虽然也是物质过程小句的施事行为者，但是使用频次明显低于China、government、policy-makers在表达愿望和期待的心理过程中充当感知者（Senser），也偶尔用作物质过程的行为者。

总之，美国媒体将中国政府和政党组织建构为经验意义及物性过程的行为者、言说者和感知者，强调其主动施事性，借人口问题干涉中国内政，企图将中国人口生育率增速缓慢、人口老龄化进程加快的现实全部归因于中国政府和制定决策者的行为、言论和想法，揭示了美国媒体恶意煽动美国民众对中国政府的敌视情

绪，以达到诋毁和破坏中国国家形象的政治阴谋。

如例（1）和例（2）所示，美国媒体将以美国、德国等国家为代表的西方发达国家建构为有能力成功解决人口问题的群内“自我”，而中国和其他东亚国家及地区则被建构为未能有效解决人口挑战的群外“他者”，进而挑起“群内”积极的“自我”与“群外”消极的“他者”两大阵营之间的对立和对抗。

（1）Now, the population (of **China**) is aging more rapidly than **those of many developed countries, including the United States**, and some argue that the government cannot afford to keep any restrictions on procreation.（NYT, May 13, 2021）

（2）**China**’s aging-related challenges are similar to those of **developed countries like the United States**. But its households live on much lower incomes on average than **in the United States and elsewhere**.（NYT, May 10, 2021）

媒体报道中the United States出现的相对频次为每万词4.6次，与China每万词12.61次的使用频次相比，差异较为显著。但值得注意的是，the United States这一专有名词在文本中除了表示经验意义所处的地域之外，大部分索引行都出现在与中国的对比语境中，如例（3）和例（4）。

（3）China’s total fertility rate, at an estimated 1.6 children per woman in 2017, is similar to Canada’s **but below those of the United States and UK**. This rate is below the 2.1 rate needed to keep the population steady, according to the World Factbook.（CNN, April 12, 2018）

（4）“People say we can be **two to three times the size of America’s economy**,” Dr. Yi said. “I say it’s totally impossible. It will **never overtake America’s**, because of the decrease in the labor force and the aging of the population.” **The United States has a much healthier age distribution**, he said.（NYT, March 24, 2016）

以往研究表明，外来移民和美国少数族裔出生率较高是1987—1990年美国人口数量保持较快增长的重要原因（楚树龙、方力维 2009）。例（3）中，美国《事实概况》竟然在年报中作出“美英人口出生率高于中国”的荒谬论断。由此不难发现，美国官方文件和媒体报道都将中美两国置于竞争对立的两端，并试图在包

括经济、军事、外交、人口在内的各方面扼制中国发展。

4.2.2 谓述策略

语言历史分析认为，谓述策略是指话语中赋予社会行为者、物体、现象、事件及过程积极或者消极评价的属性和特征的语言手段，在词汇语法层主要体现为谓词性名词、谓词和谓词性形容词（Reisigl & Wodak 2009：94）。

我们在KWIC功能中检索所有形容词索引行，将“节点词为policy、左右跨距为4”设置为筛选条件，分别得到218条和263条含policy和评价类形容词的索引行，手动删除含名词性修饰语或其他非评价类修饰语的索引行，并加以人工标注，最后筛选得到以下评价类形容词修饰语，如图3和图4所示。

图3中，policy一词左侧的形容词前置修饰语中，负面消极语义形容词占比为84.7%，如senseless、coercive、restrictive、draconian、barbaric、notorious等。中性语义的形容词不表达明确的评价态度，诸如controversial、universal、decades-long、original、major、national、current、basic、public等，占比为12.6%。表达正面积极语义的形容词修饰语如new、effective、famous等的出现频次最低，仅占2.7%。

Corpus: AM Corpus| Search Term: _/JJ.*/| Occurrences: 256/11607 (19.50)| Texts: 98/128| Filter applied|

Index	File	Left	Node	Right
113	2013_02_Nov.txt	End to a	Senseless	Policy The New
115	2013_02_Nov.txt	would loosen its	famous	''one-child'' policy, enforced
116	2013_02_Nov.txt	The world's most	controversial	birth-control policy, initially
131	2013_02_Nov.txt	that without a	coercive	and costly policy,
132	2013_02_Nov.txt	a coercive and	costly	policy, China's birthrate
190	2013_03_Nov.txt	would relax China's	draconian	one-child policy and
213	2013_03_Nov.txt	domestic security	and	foreign policy. Since they
215	2013_03_Nov.txt	for reforms in	economic	policy, while also
248	2013_03_Nov.txt	force. Under the	new	policy, couples will
257	2013_03_Nov.txt	officials to outline	major	policy changes that
283	2013_04_Nov.txt	carry out the	new	policy, and he
306	2013_04_Nov.txt	''Adjusting and improving	family-planning	policy is not
342	2013_04_Nov.txt	child under the	new	policy, in addition
370	2013_05_Nov.txt	carry out the	new	policy, and he
446	2013_06_Sep.txt	on how the	current	family planning policy
540	2013_07_June.txt	who break China's	strict	family planning policy,
631	2013_08_May.txt	government' s	barbaric	policy, the people
638	2013_09_Dec.txt	to relax its	controversial	decades-long, one-child policy
639	2013_09_Dec.txt	relax its controversial	decades-long,	one-child policy has
1,736	2015_04_Dec.txt	recently embarked on	important	policy changes, leaving
764	2013_11_Nov.txt	changes: "It's a	great	new policy. Raising
774	2013_12_Nov.txt	of its 3-decade-old	restrictive	birth policy. First
792	2013_13_Dec.txt	officially amend its	controversial	one-child policy and
3,284	2014_04_Apr.txt	instrument of	national	economic policy) has its

图3　policy左侧跨距为4的搭配形容词

Corpus: AM Corpus| Search Term: _/JJ.*/| Occurrences: 218/11607 (16.17)| Texts: 94/128|

Index	File	Left	Node	Right
2,387	2015_11_Nov.txt	policy change as a	technical	adjustment akin to how
349	2013_04_Nov.txt	policy opening will grow	bigger	and bigger.''
5,470	2017_02_May.txt	child policy was too	late	and piecemeal to arrest
3,634	2015_27_Dec.txt	the one-child policy was	ineffective	and unnecessary, since China
7,793	2019_09_Nov.txt	policy. ''It was a harsh	and	very strict policy
1,510	2015_01_Nov.txt	one-child policy had become	redundant:	As China developed and
6,638	2018_07_Aug.txt	policy has not been	effective,	as China's urban middle
2,603	2015_14_Jan.txt	the one-child policy as	early	as the 1980s, now
1,704	2015_03_Nov.txt	policy; its serious consequences	such	as the pension crisis,
4,894	2016_08_Oct.txt	policy protection create an	impossible	balancing act. ''Raising two
8,745	2021_02_June.txt	with maternity leave and	other	benefits, as well as
3,894	2015_31_Oct.txt	policy update is no silver	bullet for	economy After
3,598	2015_27_Dec.txt	policy update is no silver	bullet for	economy When
7,260	2019_04_Mar.txt	social policy-- nor the	vast	bureaucracy that still enforces
2,240	2015_09_Oct.txt	one-child policy came in	dull,	bureaucratic language. ''Comprehensiv
6,763	2018_08_Aug.txt	viewed the policy as	"painful	but necessary," it had
5,688	2017_04_ Jan.txt	a national policy discouraging		elective C-sections. The C-section rat
2,380	2015_11_Nov.txt	one-child policy, lack of	medical	care, neglect and other
3,650	2015_28_Mar.txt	draconian one-child policy on		most Chinese to curb population
6,758	2018_08_Aug.txt	baby. The policy had	worrying	consequences for the gender
2,265	2015_09_Oct.txt	policy decades ago. The	bitter	consequences of the policy
1,703	2015_03_Nov.txt	China's family-planning policy; its	serious	consequences such as the
3,608	2015_27_Dec.txt	in policy was "not enough."	"Couples	that have two
3,772	2015_29_Oct.txt	one-child policy is not	enough.	Couples that have two

图4 “policy”右侧跨距为4的搭配形容词

如图4所示，policy一词右侧的形容词常出现在归属类关系过程小句中，其中表达消极语义的形容词占比高达92.4%，充当关系小句中的属性（Attribute）成分，policy充当关系过程的载体（Carrier），不仅表达不同“声音”对我国人口政策实施的效果、影响及价值的主观判断，更暴露出美国媒体“把关人”所代表的政党群体和西方政治集团的“反华”价值观念和意识形态。他们蓄意以我国人口问题为借口，趁机干涉我国内政，企图煽动受众的负面情绪，制造不利于中国的国际舆论，最终使其诋毁中国国际声誉的政治阴谋得逞。

从评价性形容词在九个次语料库的分布来看，在2015年“全面二孩”政策实施以前，关于政策的评价中消极负面语义韵占主导，如例（5）和例（6）。

（5）This suggests that without a **coercive and costly** policy, China’s birthrate would have declined as well, **though maybe not as much**.（NYT, November 20, 2013）

（6）Perhaps surprisingly, Wang says that the one-child policy was **both ineffective and unnecessary**, since China’s fertility rates were already slowing by the 1980s.（CNN, March 24, 2015）

2015年以后，消极评价语义的占比逐渐降低，正面和中立性语义占比有所上升，如例（7）中的controversial表达生育政策调整具有争议，属于中性语义韵，例（8）中的new肯定了“三孩”政策的积极作用，即家庭在生育决策中的独立主

导性增强。

（7）Faced with a surging population, 35 years ago China attempted to put the brakes on procreation by implementing a **controversial** policy limiting most couples to just one child.（CNN, October 29, 2015）

（8）Under the **new** policy, called "independent fertility," families would be able to decide for themselves how many children to have.（USNEWS.com, May 21, 2018）

尤其在"三孩政策"正式公布前后的一段时间，随着媒体报道数量骤然增多，美国媒体对中国人口生育政策的批评声音却在减弱，如例（9）中的inclusive暗示政策松绑以后政策更具"包容性"，例（10）中的swiftly，new表明媒体对"三孩"新政策出台速度之快表示惊讶，timid and unimaginative、irrelevant透露出媒体对中国政府及政策制定者挖苦和质疑的态度；例（11）中媒体将中国政府采取的鼓励生育及支持配套措施建构为"哄诱"行为（coax），他们有意扭曲中国政府主动担当、主动作为的正面形象，对中国政府积极应对人口老龄化、建构"生育友好型"社会的配套福利和支持举措闪烁其辞，也表明美国不愿正视中国"迅速崛起"的现实，而是选择回避的态度。

（9）Party leaders have pledged to make population policies more **inclusive**, a signal that some have taken to mean the rules will be eased further.（NYT, May 12, 2021）

（10）However **swiftly** the **new** three-child policy followed those results, it is **timid and unimaginative**, and it will be **largely irrelevant**.（NYT, June 7, 2021）

（11）And so the Chinese government isn't just encouraging women to have more children- and hoping to **coax them with maternity leave and other benefits, as well as promises to mobilize resources at all levels of the state**.（NYT, June 9, 2021）

4.2.3 视角化策略

视角化策略指那些用来定位说话人或作者观点，表达作者介入或是疏离的语言手段及话语实践（Reisigl & Wodak 2009：94）。

Bakhtin（1981）指出，任何语篇都具有"对话性"和"多声性"的特质。本

节将聚焦语料库文本中的“投射”现象，类似于传统语法所提及的直接引语和间接引语。我们将在系统功能语言学派的语义评价系统框架下讨论“投射”现象。“介入”是评价系统的三个子系统之一，它关注投射、情态、极性等语法资源如何定位读者、作者与他们所赞同的立场观点之间的关系。Martin & White（2005）将引述、转述言语行为称为“投射”，是定位作者立场观点的语法结构，一般由“投射小句+被投射小句”组成，包括投射源、投射方式和投射内容三个维度。Martin & White（2005：135）明确指出：“相比投射命题的来源，作者如何定位自己的‘声音’显得更为重要。”换言之，就“投射”现象的研究而言，考察作者本身的立场、观点和态度，即厘清作者对观众或读者反应的预期持赞成还是反对态度，比研究投射来源更可靠、更有意义。

通过人工逐条检索观察语料库文本，并提取语料中的动词及相应的搭配词项后，笔者以say、tell、add、show、argue、claim这六个转述动词及对应的屈折变化词形为检索词，人工手动筛除检索词为非投射动词用法的索引行。此外，引述短语according to同以上投射动词一并归为“投射标记”，在语料库中逐一检索，共获得1,661条含“投射标记”的索引行（KWIC），各类投射标记的出现频次及所占比例统计如下（见表4）。

表4　投射标记在语料中的出现频次及比例

投射标记（转述动词或短语）	语义类别	介入类型	绝对频次（次）	相对频次（次/万词）	所占比例
shows/-ed/-ing	收缩	公告：背书	36	2.67	2.17%
say/-s/said/-ing	扩展	归属：承认	1,385	102.7	
argue/-s/-d	扩展	归属：承认	37	2.74	
add/s/-ed	扩展	归属：承认	28	2.08	97.47%
tell/-s/-ed	扩展	归属：承认	16	1.19	
according to	扩展	归属：承认	153	11.35	
claim/-s/-ed	扩展	归属：疏离	6	0.45	0.36%
		总计	1,661	123.18	100%

由表4可见，表达扩展对话空间语义的投射标记占比达97.83%，媒体记者使用转述动词或短语来投射言语（Locution）或者思想（Idea），使用对不同来源的“声音”和观点表达“承认”（97.47%）或者“疏离”（0.36%）态度的语言手段，间接表明媒体从业者及其利益集团的观点和立场，达到定位作者自身态度立场、为图达到影响受众的认知情感取向和结盟读者的目的。相反，表达收缩对话空间语义的投射标记的使用频次仅占2.17%，这类投射动词的投射来源均为无生命的

事物，即官方披露的具有权威性的政府报告、文件、统计数字和数据等，如the policy、the report、data、census、figures、statistics。

笔者选取使用频次最高的投射动词said（1,141次）作为检索词，在语料中搜索它的搭配共现高频词，可视化分析（见图5）表明，“投射源”为名词的搭配高频词的搭配强度由高到低排列依次为：professor、China、NHC（National Health Commission）、demographer、officials、experts。

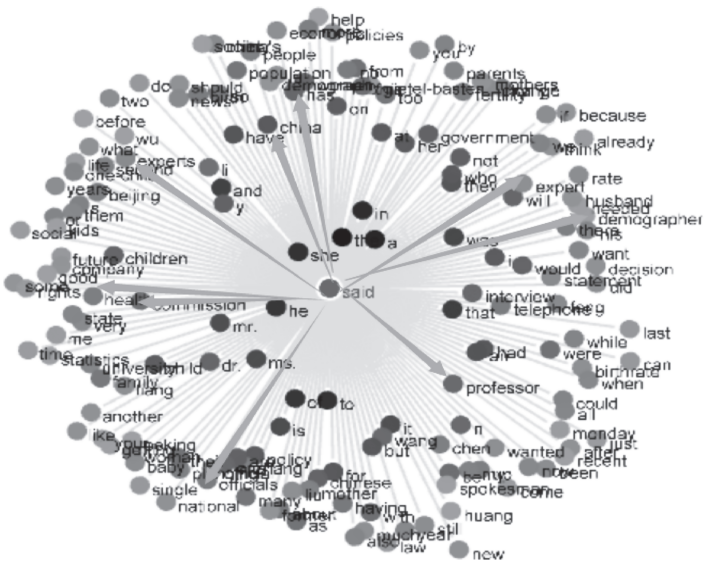


图5　said搭配“磁场效应”

媒体以政府及权威健康机构、官方颁布的文件、统计数据等无生命事物和学术权威、政府官员等作为主要投射源，目的在于将被引述话语和观点“合法化”，进而达到说服受众认同自己观点的目的。

4.2.4　辩论策略

话语历史分析的辩论策略是指运用相应的论题（Topoi），对正面或负面的观点进行论证（Reisigl & Wodak 2009：94）。下面将重点从“数字”“危险”“威胁”和“负担”四个论题分析美国媒体如何建构积极的“自我”和消极的“他者”形象。

例（12）中，说话人引述国家卫健委主任推测的言语，引用确切的时间和人口相关数字，有助于增强数据的“客观性”和“真实性”，提高论证的说服力，为“人口政策调整不会加速中国人口增长，与新华社官方报道的人口增长预期有一定差距”这一论断提供“证言”。

（12）According to Xinhua, China's food safety and public service schemes are designed to meet the needs of **1.43 billion** people by **2020** and **1.5 billion** in **2033**. Even with the policy change, the total population will not exceed **1.38** billion in **2015**, Li predicted.（CNN, December 23, 2013）

crisis一词在语料库中的相对频次达到6.23次/万词，出现在超过三分之一的语料文本中，尤其是“人口危机”“老龄化危机”等搭配强度相对较高。例（13）中demographic crisis looming意在说明80年代实施的“一对夫妻只生一个孩子”的生育政策带来的影响和后果，即人口增长缓慢和老龄化加剧等，潜在的人口危机风险步步逼近，说话人使用“危险”论题的辩论策略，有可能引起读者情绪上的恐慌和不安。

（13）For decades, couples were generally limited to one child to slow population growth. With a potential **demographic crisis looming**, the government now wants them to have more.（NYT, June 1, 2021）

如例（14）中所示，劳动力短缺和人口老龄化，直接对中国实现“建设工业现代化经济强国”的发展战略构成威胁和阻碍，threat一词在语料中的使用频次仅为2.67次/万词，但这一辩论策略也成为美方媒体话语的合法化策略之一。

（14）The labor pool is shrinking and the population is graying, **threatening** the industrial strategy that China has used for decades to emerge from poverty to become an economic powerhouse.（NYT, May 31, 2021）

例（15）中，美方媒体转述模糊来源的working-class parents的观点，试图将自己毫无根据的妄加揣测强加给受众，即“三孩”新政策给育龄夫妻带来的教育、养育成本负担，让这些家庭“不堪重负”。使用“负担”论题的辩论策略旨在影响受众的舆论导向，恶意放大国内民众对政策调整的不友好态度，引发受众的负面情绪和反应。

（15）**Working-class parents** said the financial burden of more children would be unbearable.（NYT, June 1, 2021）

4.3 历史和社会语境分析

党中央国务院历来高度重视我国人口问题，探索出了一条具有中国特色的解决人口问题的道路。2021年8月20日全国人大常委会会议表决通过了《关于修改人口与计划生育法的决定》。修改后的人口计生法规定，国家提倡适龄婚育、优生优育，一对夫妻可以生育三个子女。国家采取财政、税收、保险、教育、住房、就业等一系列优惠条件和支持措施，例如：取消社会抚养费、设立父母育儿假、推动建立普惠托育服务体系等，切实减轻家庭生育、养育、教育负担。2022年3月，李克强总理在《政府工作报告》中也明确指出，完善三孩生育政策配套措施，将三岁以下婴幼儿照护费用纳入个人所得税专项附加扣除，发展普惠托育服务，减轻家庭养育负担。

这些不争的事实足以让美方媒体报道中很多恶意诋毁中国政府形象的言论不攻自破。无论是“二孩”还是“三孩”政策，都是计划生育政策优化的具体体现，体现了政策的包容性、开放性、鼓励性和融合性。它们都是过渡性政策，渐进性路径，增量性改革，是阶段性的优化改良（穆光宗 2021：68-72）。美方无视中国改革和发展的阶段性，无视中国国情，忽视政策的时代性和动态发展规律，在媒体报道中不仅有意回避生育政策调整带来的经济高速发展的利好，还恶意放大政策的短板，以人口政策为幌子，将人口问题与少数民族、人权、女性权益等问题捆绑在一起，借机对中国的国事和家事指手画脚、评头论足，将媒体所维护的政治集团的意识形态和政治立场强加给国际受众，妄图达到煽动少数民族分裂情绪、阻碍中国经济发展、构建消极负面的国家形象的目的。

5 结论

本文运用话语历史定性分析与语料库定量分析相结合的方法，自建专用语料库，考察美国四家媒体近九年有关中国人口政策报道的话语动态演变轨迹。结合社会历史语境，研究发现美国媒体运用了命名策略、谓述策略、视角化策略和辩论策略，对我国人口政策调整的态度经历了从“全盘否定和批评”，过渡到“观望和质疑”，最后发展为“勉强认可”这一复杂的历史演变过程。美国媒体将包括中国在内的东亚国家及地区构建为“群外”消极、负面、无助的“他者”，而把自己和英国、德国等西方发达国家建构为有能力成功解决人口问题的“群内”成功的、有能力的“自我”，人为制造“群内”与“群外”两大阵营之间的对立和对抗，暴露了美方对我国人口政策的恶意诽谤和攻击，以及妄图遏制中国经济发展、抹黑中国政府的国际形象、影响媒体舆论导向的险恶用心和实施“霸权政治”的政治图谋。

作为外化为符号表达的观念和信息，政策具有很强的时效性和阶段性。政治

话语研究需要充分考虑话语产生的历史和社会语境，描绘话语的动态演变轨迹，这也是本研究采用话语历史分析研究方法的原因。

本研究尝试融合话语历史分析的理论概念、方法和语料库分析方法，考察美国媒体关于我国人口政策报道的话语建构和动态演变，为今后相关研究提供借鉴和参考。诚然，语料库规模有待扩大，语料来源也有待丰富，期待今后相关学者能从更广阔的学科前沿理论视角开展研究，以丰富政治话语的研究成果。

参考文献

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

BAKHTIN M. The dialogic imagination [M]. Austin: University of Texas Press, 1981.

CHILTON P. Analysing political discourse [M]. London: Routledge, 2004.

ENGSTROM 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.

FAIRCLOUGH N. Language and power [M]. London: Longman, 1989.

FAIRCLOUGH I, FAIRCLOUGH N. Political discourse analysis [M]. London: Routledge, 2012.

GOMEZ-JIMENEZ E. “An insufferable burden on businesses?” On changing attitudes to maternity leave and economic-related issues in the Times and Daily Mail [J]. Discourse, Context & Media, 2018, 26: 100-107.

LIU M, LIN L. “One Country, Two Systems”: a corpus-assisted discourse analysis of the politics of recontextualization in British, American and Chinese newspapers [J]. Critical Arts, 2021, 35(3): 17-34.

MARTIN J, WHITE P. The language of evaluation [M]. New York: Palgrave Macmillan, 2005.

MAUTNER G. Checks and balances: how corpus linguistics can contribute to CDA [C]//WODAK R, MEYER M. Methods of critical discourse analysis (2nd edition). London: Sage, 2009: 128-150.

O’HALLORAN K, COFFIN C. Checking overinterpretation and underinterpretation: help from corpora in critical linguistics [C]//COFFIN C, HEWINGS A, O’HALLORAN K. Applying English grammar: corpus and functional approaches. London: Arnold, 2004: 275-297.

PARTINGTON A. The linguistics of political argumentation: the spin-doctor and the

wolf-pack at the White House [M]. London: Routledge, 2003.

REISIGL M, WODAK R. The discourse-historical approach [C]//WODAK R, MEYER M. Methods of critical discourse analysis (2nd edition). London: Sage, 2009: 87-121.

THOMPSON J. Ideology and modern culture [M]. Cambridge: Polity Press, 1990.

楚树龙，方力维. 美国人口状况的发展变化及其影响 [J]. 美国研究，2009（4）：75-89.

穆光宗. 三孩政策与中国人口生育的优化：背景、前景和愿景 [J]. 扬州大学学报（人文社会科学版），2021（4）：65-77.

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.

杨敏，侍怡君. 中美贸易战中美方“合法化”话语建构——基于语料库的话语–历史分析 [J]. 外语研究，2021（3）：7-13.

通信地址：528000　广东省佛山市　佛山开放大学外语系

DiSCUSS现代汉语平衡口语语料库的创建*

北京外国语大学 孙铭辰

提要：本文主要介绍“DiSCUSS现代汉语平衡口语语料库”（简称DiSCUSS库）的建设过程。作为国内首个开源的百万词级现代汉语平衡口语语料库，DiSCUSS库采用与“国际英语语料库”相同的取样模式创建，库容为100万词。该语料库具有较好的平衡性和代表性，使其可广泛应用于汉语口语研究、汉外口语对比等领域。此外，DiSCUSS库提供的社会语言学变量、说话人标记和词性标注等也可作为开展语言变异、话轮转换机制、叙事结构等研究议题的重要数据基础。

关键词：DiSCUSS库、现代汉语平衡口语语料库、国际英语语料库、国际可比语料库

1 引言

DiSCUSS库是按照国际英语语料库（简称ICE）的取样模式（Greenbaum & Nelson 1996）创建的百万词级现代汉语平衡口语语料库。该语料库由北京外国语大学中国外语与教育研究中心许家金教授主持建设。作为国内首个开源的百万词级现代汉语平衡口语语料库，DiSCUSS库在汉语口语话语研究、汉外口语对比、汉语二语习得等领域具有广泛的应用前景。

2 研制背景

2.1 口语语料库建设的三个世代

口语语料库的建设一直以来都是困扰学界的难题。口语语料库的建设难度数倍于书面语语料库，因为前者收集转写语音的难度更大、成本更高、速度更慢（Love *et al.* 2017）。Burnard（2002）更是认为建设100万词口语语料库的工作量是建设同样规模新闻报刊类语料库的10倍。以布朗语料库和伦敦–隆德口语语料库

* 本文系国家社科基金一般项目“概率语境共选视角下的多语外汉词典数据库建设与研究”（21BYY021）的阶段性成果。感谢许家金、董通、陈哲、康卉、苏杭、李银美、刘芳芳、王波、王义娜、王彦，马博森、权立宏、陆军、朱周晔、钱一华、刘朝霞等老师和同学对DiSCUSS库建设的支持和所付出的巨大努力。感谢许家金教授对本文写作提出的宝贵意见。

（简称LLC）为例，第一个电子化英语平衡书面语语料库——Brown语料库——于1962年立项，历经三年建设，于1964年问世（Francis & Kučera 1964）；第一个电子化通用英语口语语料库——LLC语料库——自Qurik于1959年发起“英语用法调查”项目，陆陆续续历经二十年，才在1980年初步问世，但库容也仅为50万词左右（Svartvik 1990）。

随着计算机运算与存储技术的不断革新，口语语料库的建设与语料库研究齐头并进，经历了1.0、2.0、3.0三个世代（许家金 2017）。英语口语语料库的建设在三个世代都有其代表性成果。在1.0世代即前电子化时代，Fries（1952）曾录制并转写了25万词美国中北部居民的标准英语对话，用以编写英语语法（许家金 2019）。1959年，Quirk发起“英语用法调查”项目，通过数以千计的纸条记录英语口语用例。进入口语语料库建设的2.0世代后，1975年，Svartvik在“英语用法调查”项目的基础上增补语料，并将纸介转为电子文档，于1980年初步建成第一个电子化通用英语口语语料库LLC（Svartvik 1990）。20世纪80年代，Sinclair（1989）领导“英语文库”（Bank of English）项目，库中部分口语语料取自广播电视、非正式对话等。1993年，青少年英语口语语料库COLT问世（Stenström & Breivik 1993）。1994年，兰卡斯特大学发布英国国家语料库（简称BNC），其中包含1,000万词的英国英语口语语料（Crowdy 1995）。二十多年后兰卡斯特大学又推出BNC2014口语语料库（Love *et al.* 2017）。1998年，“国际英语语料库”ICE项目推出第一个英国英语分库ICE-GB，其中60%的部分由口语语料构成，共计60万词（Greenbaum & Nelson 1996）。在2.0世代后期，通用口语语料库的建设逐渐转向专用口语语料库，如学术口语语料库MICASE语料库（Simpson *et al.* 2000）、T2K-SWAL语料库（Biber *et al.* 2001）、BASE语料库（Thompson & Nesi 2001）等。在大数据技术推动下，21世纪口语语料库的建设进入3.0世代。超大规模性是3.0世代口语语料库的重要特点。1.3亿词（截至2020年3月）的COCA语料库口语子库是目前规模最大的口语语料库之一（Davies 2010，2020）。Davies所领导建设的其他系列语料库，如电视语料库、电影语料库、美国电视剧语料库等也均在亿词规模，为英语口语研究提供了丰富的研究语料。

相较于英语口语语料库的建设，现代汉语口语语料库的研制兴起于20世纪80年代前后，虽起步较晚，但发展势头迅猛。在1.0世代即前电子化时代，Chao（1968）曾基于真实口语语料完成《中国话的文法》一书。进入电子化的2.0世代后，北京语言学院于1981年开展“北京口语调查”项目（宋孝才 1987），基于社会语言学人口抽样原则，调查近378名北京人非正式场合的连贯话语，收集了150多个小时录音，共计230万词左右的口语语料。顾曰国（2002）于1999年前后主持开展北京地区现场即席话语语料库的研制工作，根据北京地区电话黄页设计取样方案，最终完成650小时录音的语料收集。中国传媒大学也于2005年

开始搭建有声媒体文本语料库（http://ling.cuc.edu.cn/RawPub/）。此外，为服务特定研究目的，学者们会创建中小型汉语口语语料库。例如，Biq（1995）曾转写了5段总时长120分钟的对话建立汉语口语语料库，探究汉语口语中的因果关系。CALLHOME语料库建立于1996年，内含120个在美中国留学生同国内的电话录音转写（Alexandra *et al.* 1996）。方梅（2000）转写了6段总时长4小时的自由对话录音材料并由此研究汉语口语中的弱化连词。在青少年汉语口语语料库的建设方面。许家金（2008，2009）转写了14万字的城市青少年汉语口语语料库，并开展了话语标记系列研究。在汉语口语语料库建设的3.0世代，北京语言大学的BCC语料库（荀恩东等 2016）和北京大学的CCL语料库现代汉语部分（詹卫东等2019）最具代表性。BCC语料库共有95亿字，其中口语部分有6亿字，均由爬取微博和影视字幕获得。6亿字的CCL语料库含有约150万字的口语（对话）部分、1,000万字的电视电影部分和160万字的相声小品部分。

2.2　国际英语语料库和国际可比语料库

“国际英语语料库”项目始于20世纪80年代末，由伦敦大学学院英语用法调查研究所Randolph Quirk的继任者Sidney Greenbaum（1988）发起，下辖澳大利亚、巴哈马、加拿大等27支语料库建设队伍，旨在基于统一的采样框架，在英语作为官方语言的国家或地区，建设代表世界各国家、地区英语变体的多个可比语料库，并开展英语变体的共时比较研究。截至2016年，共建成英国、加拿大、爱尔兰、印度、菲律宾、牙买加等11个分库（The ICE Project 2016）共计500个2,000词的文本，库容为100万词，其中60%为口语，40%为书面语，体现了“国际英语语料库”项目对口语研究的重视。“国际英语语料库”项目的采样框架继承了“英语用法调查”项目（Quirk *et al.* 1972）的宝贵经验，并进一步改进创新（见图1）。DiSCUSS库按照此框架进行采样（见表1），并在此基础上扩容1.67倍，使其成为百万词级的现代汉语口语语料库。

“国际可比语料库”项目是采用与ICE大致相同的采样框架建设的多语种百万词级可比语料库，其中60%为口语，40%为书面语。项目下辖捷克语、芬兰语、法语、德语、爱尔兰语、意大利语等12支语料库建设队伍（Čermáková *et al.* 2021）。ICC的建立旨在同ICE的各英语变体及在ICC内各语言之间开展多语种、多语体的对比研究（Kirk & Čermáková 2017）。其汉语分库ICC-CN部分由许家金教授带领的北京外国语大学语料库语言学团队负责建设。以ICE模式创建的DiSCUSS库将抽取部分语料构成ICC-CN的60%口语部分；ToRCH2019语料库（李佳蕾等 2022）的部分语料及爬取的10万词电子博客将构成ICC-CN的40%书面语部分。

综上，英语口语语料库的建设总体上先于汉语口语语料库的建设。汉语口语语料库建设方面比较突出的问题是平衡语料库的缺乏。DiSCUSS库基于具有代表

性和平衡性的ICE采样框架进行建设，能够在一定程度上推进现代汉语口语语料库中平衡语料库的建设步伐。

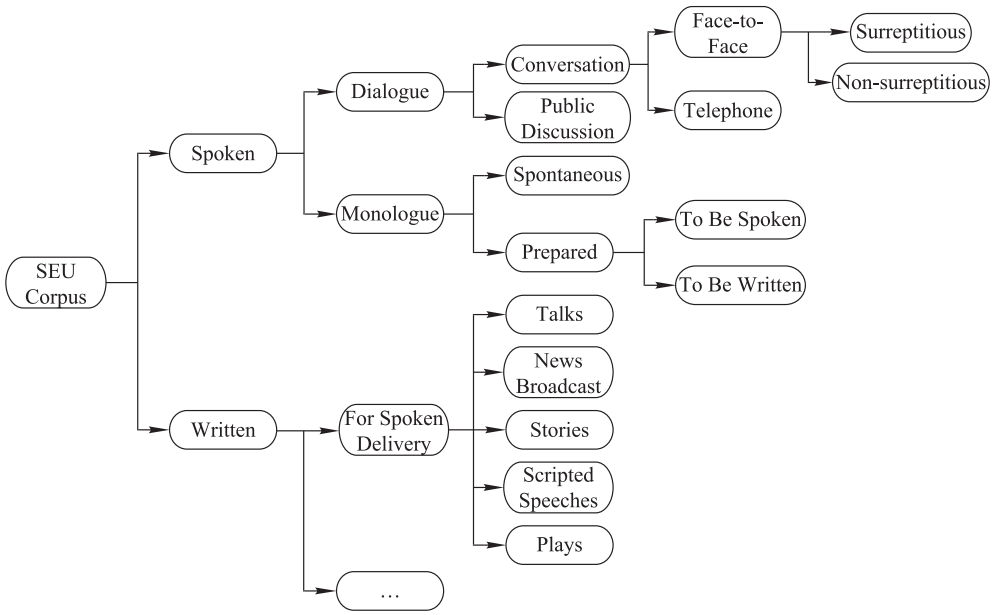


图1 “英语用法调查”项目语料库口语相关部分架构

3 DiSCUSS库的建设概况

DiSCUSS库按照ICE口语部分的构成采集样本，并将其60万词既定规模扩大到百万词级。如表1所示，DiSCUSS库共包含300个文本，1,002,538词（正则表达式：[\u4e00-\u9fa5a-za-zA-ZA-Z0-90-9\.%％]+），分为对话和独白两大类，以及私人、公开、脱稿、念稿四小类，并可细分为15个子类。各子类代表不同口语场景或语境，与DiSCUSS库名称中的Social Settings相呼应。

表1 DiSCUSS现代汉语平衡口语语料库构成情况

文本类别			文本名称	文本数量	文本词数
对话	私人	当面交谈	S1A-001至SlA-090	90	296,694
		电话交谈	S1A-091至SlA-100	10	33,106

（待续）

（续表）

文本类别			文本名称	文本数量	文本词数
对话	公开	课堂教学	S1B-001至SlB-020	20	67,108
		媒体讨论	S1B-021至SlB-040	20	64,915
		媒体采访	S1B-041至SlB-050	10	33,320
		赛场辩论	S1B-051至SlB-060	10	31,944
		法庭质证	S1B-061至SlB-070	10	35,153
		商业交易	S1B-071至SlB-080	10	34,618
独白	脱稿	自发评论	S2A-001至S2A-020	20	67,788
		无稿演讲	S2A-021至S2A-050	30	106,089
		演示介绍	S2A-051至S2A-060	10	33,703
		法庭陈述	S2A-061至S2A-070	10	35,178
	念稿	媒体新闻	S2B-001至S2B-020	20	66,634
		媒体讲话	S2B-021至S2B-040	20	62,887
		有稿演讲	S2B-041至S2B-050	10	33,401
				300	1,002,538

DiSCUSS库中的转写文本，经过了一定的人工标注与清洗，包括标注说话人及话轮、标注部分说话人动作状态、修改错误转写内容、隐私信息处理等。DiSCUSS库中全部300个文本经过了系统清洗，并使用Jieba（https://github.com/fxsjy/jieba）组件进行了分词与词性标注处理。

DiSCUSS库分为三个版本：纯文本版、分词版和分词标注版，以UTF-8编码的txt纯文本格式储存。DiSCUSS库的元信息Excel表存储着一一对应的文本编号、说话人、说话人性别（S1A001至SlA-100）、场景、时间、来源、词数统计等元信息。

DiSCUSS库的语料产生于1999—2022年，其中2017—2022年的语料比例占全库的50%左右。

4 DiSCUSS库的主要特点

相较于以往汉语口语语料库的研制，DiSCUSS库具有多个显著特点，可广泛应用于汉语口语话语研究、汉外口语对比、汉语二语习得等领域。

（1）代表性与平衡性。DiSCUSS库的建设参照国际标准，采用“国际英语语料库”口语部分的采样框架。“国际英语语料库”模式能够使DiSCUSS库在较好保证代表性与平衡性的同时，收集多种语境下的口语语料，充分体现DiSCUSS库名称中Diversified的特点，可为汉语口语话语研究，特别是汉语口语的语域变异研究提供支持，也可为汉语二语习得领域的对外汉语教学、教材设计等提供充分、真实的口语语料。

（2）可比性。DiSCUSS库的精简版（56万词）将构成“国际可比语料库”汉语分库ICC-CN的口语部分，并在未来与“国际英语语料库”家族内27个英语变体语料库、家族内12个多语种语料库构成可比语料库，从而开展汉外口语对比研究。

（3）丰富的元信息及标注。DiSCUSS库提供了丰富的社会文化元信息语境光谱，包括说话人性别、说话人关系、场景、主题、时间、来源等（见表2）；DiSCUSS库也在文内进行了说话人及话轮标注、词性标注（见表3）。其丰富的社会文化语境元信息、说话人标记和词性标注将会在多因素分析、语域变异、话轮结构、叙事结构等基于语料库的话语研究子领域中发挥重要作用。

表2　DiSCUSS现代汉语平衡口语语料库元信息表（部分）

文本编号	二级编号	说话人（场景、主题等）	时间	词数	来源	性别
S1A-050	1	朱（女）、陈（男）（同学）	2020年	3,483	康卉	1男1女
	2	采访者A、被采访者B	2019年		刘朝霞	2女
S1A-067	1	大哥、二哥、三哥、四哥（室友）	2021年	3,313	康卉	4男
	2	同学A、同学B	2021年		康卉	2女
S1B-064		江歌被害案之江秋莲诉刘暖曦生命权纠纷案庭审记录：审判长、审判员、原告、被告、人民陪审员	2022年01年10日	3,490	青岛市城阳区人民法院	

表3　摘自DiSCUSS现代汉语平衡口语语料库的S1A–029文本

说话人及话轮标注	内容（分词及词性标注）
刘先生：	你_r 吃_v 不_d 惯_v，_x 那个_r 饭_n，_x 煮_v 来_v 你_r 也_d 吃_v 不_d 惯_v 。_x
刘妻：	啊_zg，_x 对_p 。_x
赵先生：	可以_c，_x 可以_c，_x 可以_c 。_x

（4）开源性。DiSCUSS库的纯文本版、分词版和分词标注版三个版本将全文发布于“北外语料库语言学”网站官网。“共建共享，取之于用，返之为用”，是北外语料库团队一直奉行的“语料为公”理念。

5 结语

本文介绍了DiSCUSS现代汉语平衡口语语料库的研制背景、建设方案、主要特点及应用方向。对于DiSCUSS库的获取方式，感兴趣的读者可访问“北外语料库语言学”网站（http://corpus.bfsu.edu.cn），全文免费下载DiSCUSS现代汉语平衡口语语料库，或登陆北外CQPweb多语种语料库平台（http://114.251.154.212/cqp/；账号：test，密码：test）进行检索分析。

参考文献

ALEXANDRA C, ZIPPERLEN G. CALLHOME. Mandarin Chinese speech LDC96S34 [DB/OL]. Philadelphia: Linguistic Data Consortium, University of Pennsylvania, 1996. https://catalog. ldc. upenn. edu/LDC 96S34.

BIBER D, REPPEN R, CLARK V, et al. Representing spoken language in university settings: the design and construction of the spoken component of the T2K-SWAL Corpus [C]//SIMPSON R, SWALES J. Corpus linguistics in North America. Ann Arbor: University of Michigan Press, 2001: 48-57.

BIQ Y. Chinese causal sequencing and yinwei in conversation and press reportage [C]// BILMES L, LIANG C, OSTAPIRAT W. The proceedings of the 21st annual meeting of the Berkeley Linguistics Society. Berkeley: Berkeley Linguistics Society, 1995: 47-60.

BURNARD L. Where did we go wrong? A retrospective look at the British National Corpus [C]//KETTEMANN B, MARKUS G. Teaching and learning by doing corpus analysis. Amsterdam: Rodopi, 2002: 51-71.

ČERMÁKOVÁ A, JANTUNEN J, JAUHIAINEN T, et al. The International Comparable Corpus: challenges in building multilingual spoken and written comparable corpora [J]. Research in Corpus Linguistics, 2021, 9(1): 89-103.

CHAO Y R. A grammar of spoken Chinese [M]. Berkeley: University of California Press, 1968.

CROWDY S. The BNC spoken corpus [C]//LEECH G, MYERS G, THOMAS J. Spoken English on computer: transcription, mark-up and annotation. Harlow: Longman, 1995: 224-235.

DAVIES M. The Corpus of Contemporary American English as the first reliable monitor

corpus of English [J]. Literary and Linguistic Computing, 2010, 25(4): 447-464.

DAVIES M. The COCA corpus [R/OL]. (2020-03-01) [2022-06-25]. https://www.english-corpora.org/coca/help/coca2020_overview.pdf.

FRANCIS W N, KUČERA H. Brown Corpus [DB/OL]. Providence: Department of Linguistics, Brown University, 1964.

FRIES C. The structure of English: an introduction to the construction of English sentences [M]. New York: Harcourt, 1952.

GREENBAUM S. Proposal for an International Corpus of English [J]. World Englishes, 1988, 7: 315.

GREENBAUM S, NELSON G. The International Corpus of English (ICE) project [J]. World Englishes, 1996, 15: 3-15.

KIRK J, ČERMÁKOVÁ A. From ICE to ICC: the new International Comparable Corpus [C]//BAŃSKI P, KUPIETZ M, LÜNGEN H, et al. Proceedings of the workshop on challenges in the management of large corpora and big data and natural language processing (CMLC-5+ BigNLP) 2017 including the papers from the Web-as-Corpus (WAC-XI) guest section. Mannheim: Institut für Deutsche Sprache, 2017: 7-12.

LOVE R, DEMBRY C, HARDIE A, et al. The Spoken BNC2014: designing and building a spoken corpus of everyday conversations [J]. International Journal of Corpus Linguistics, 2017, (3): 319-344.

QUIRK R, GREENBAUM S, LEECH G, et al. The Grammar of Contemporary English [M]. London: Longman, 1972.

SIMPSON R, LUCKA B, OVENS J. Methodological challenges of planning a spoken corpus with pedagogic outcomes [C]//BURNARD L, MCENERY T. Rethinking language pedagogy from a corpus perspective: papers from the third international conference on teaching and language corpora. Frankfurt: Peter Lang, 2000: 43-49.

SINCLAIR, J. Collins COBUILD English language dictionary [M]. Glasgow: Collins Publishers, 1989.

STENSTRÖM A, BREIVIK L. The Bergen Corpus of London Teenager Language (COLT) [J]. ICAME Journal, 1993, 17: 128.

SVARTVIK J. The London-Lund Corpus of Spoken English: description and research [M]. Lund: Lund University Press, 1990.

THE ICE PROJECT. International Corpus of English [EB/OL]. (2016-02-09) [2022-09-24]. http://ice-corpora.net/ice/index.html.

THOMPSON P, NESI H. The British Academic Spoken English (BASE) corpus project [J]. Language Teaching Research, 2001, (3): 263-264.

方梅. 自然口语中弱化连词的话语标记功能[J]. 中国语文，2000（5）：459-470.

顾曰国. 北京地区现场即席话语语料库的取样与代表性问题[C]// 中国社会科学院，法国国家科研中心. 全球化与21世纪首届“中法学术论坛”论文集. 北京：社会科学文献出版社，2002：484-500.

李佳蕾，孙铭辰，许家金. ToRCH2019现代汉语平衡语料库[DB]. 北京：北京外国语大学中国外语与教育研究中心/人工智能与人类语言重点实验室，2022.

宋孝才. 谈“北京口语调查”[J]. 世界汉语教学，1987（2）：25-29.

许家金. 汉语自然会话中话语标记“那（个）”的功能分析[J]. 语言科学，2008（1）：49-57.

许家金. 青少年汉语口语中话语标记的话语功能研究[M]. 北京：外语教学与研究出版社，2009.

许家金. 语料库研究学术源流考[J]. 外语教学与研究，2017（1）：51-63.

许家金. 美国语料库语言学百年[J]. 外语研究，2019（4）：1-6.

荀恩东，饶高琦，肖晓悦，等. 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，等. 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：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

deGLOBE当代德语书面语平衡语料库的创建*

北京外国语大学 周顾盈 宋瑛明 舒 哲 孙 昱 徐 亮

提要：deGLOBE当代德语书面语平衡语料库是“北外全球语料库集群”项目（又称“GLOBE语料库”项目）下的一个子课题，旨在收集近十年的德语书面语文本，创建百万词级的平衡语料库。本文首先简述当前面向德语的语料库建设情况，在此基础上对deGLOBE的建库理念与建库过程进行较为全面的论述，并对基于该语料库的语言研究与教学，以及后续建设规划作出展望。

关键词：deGLOBE语料库、当代德语书面语平衡语料库、德语教学与研究

2021年12月29日，北京外国语大学启动了“北外全球语料库集群”项目，又称“GLOBE语料库”项目。GLOBE为Global Languages Out of BFSU Expertise首字母缩略词。该语料库集群项目致力于建设北外开设的101个语种的当代书面语语料库。其下所有单语平衡库均借鉴布朗语料库（The Brown Corpus）的采样方案，使之可与现有英汉语布朗家族语料库进行对比，从而开展外汉、外英或多语对比研究。“deGLOBE当代德语书面语平衡语料库”（简称deGLOBE语料库）为GLOBE语料库集群下的德语子库，旨在收集2012—2022年首次出版或发表的原创德语文本，其设计规模为100万词。

1 面向德语的语料库建设简述

德语语料库建设可追溯到前电子化时代。1897年，德国速记员Friedrich Wilhelm

* 本文系北京外国语大学2022年度“双一流”重大标志性项目“多语种词典编纂理论与实践研究”（2022SYLZD015）及北京外国语大学2022年度“双一流”重大标志性（培育）项目“全球语料库集群建设与研究”（2022SYLPY004）的阶段性成果。周顾盈是本文通讯作者。
作者贡献：
周顾盈：选题构思、研究方法、数据收集（语料贡献占比28%）、数据分析、讨论结论、初稿撰写；
宋瑛明：数据收集（语料贡献占比21%）、修改润色；
舒 哲：数据收集（语料贡献占比17%）、修改润色；
孙 昱：数据收集（语料贡献占比17%）、修改润色；
徐 亮：数据收集（语料贡献占比17%）、修改润色。

Kaeding出版了《德语词频词典》（Häufigkeitswörterbuch der deutschen Sprache），旨在基于有代表性的德语语料，通过词频统计的方式获得常用词表，用于改进德语速记法。该项目共计收集近1,100万词各类体裁的德语文本，并统计其中超过25万个单词的频数（Kaeding 1897/1898）。除了不可机读外，Kaeding在早期项目中所建立的德语文本数据库与如今我们所熟知的电子语料库别无二致，因此也称为Kaeding-Korpus（Kübler & Zinsmeister 2015：5）。

进入电子化时代，得益于计算机技术的发展，越来越多类型丰富、用途广泛的可机读语料库如雨后春笋般涌现，各种规模的语料库层出不迭。当前，面向德语的电子化语料库主要包括但不限于以下几种类型。（1）大规模参照语料库。以德国语言研究院主持的德语参照语料库DeReKo、柏林–勃兰登堡科学院资金支持下的DWDS词典项目为代表。前者规模达百亿词，称得上是当前世界上最大的德语文本库（Lüngen 2017：161）；后者致力于建设涵盖20世纪和21世纪文本的大型平衡语料库，进而编写当代德语电子词典（Geyken 2007：23）。（2）口语语料库。包括覆盖多个场景的口语库FOLK（Schmidt 2018：216），以及其他类型的德语口语库，如包含本族语者与学习者在内的BeMaTaC口语库、学术口语库GeWiss等。（3）历时语料库。包括覆盖时间段较长的DTA语料库（1465—1969年）和RIDGES语料库（1450—1900年），以及专门针对古德语（750—1050年）、中古高地德语（1050—1350年）、早期新高地德语（1350—1650年）的参照语料库DDD、ReM、ReF等。（4）学习者语料库。如洪堡大学开发的德语学习者错误标注语料库Falko，以及包含德语在内的欧盟框架下多语种学习者语料库MERLIN。（5）专用语料库，例如新闻语料库（如TIGER、TüBa-D/Z）、网络语料库（如DeWaC、DECOW）、德国议会演讲语料库（Parlamentsreden Deutscher Bundestag）等。

以上着重列举了公开可访问的语料库。可以说，面向德语的语料库建设总体较为成熟。尽管如此，当代德语书面语，尤其是2010年后德语书面语的平衡语料库建设仍然值得继续推进。从前电子化时代的Kaeding-Korpus，到当前最大的德语参照语料库DeReKo，再到当前最大的德语平衡语料库DWDS，“平衡性”始终贯穿于建库理念中。DeReKo以新闻语料为多数，未严格采用平衡语料库的建库模式，而是由多个子语料库组成，其中包括平衡语料库LIMAS-Korpus（1964）。DWDS词典项目采用平衡采样原则收集20世纪和21世纪的德语语料，目前可供检索的最新语料仅至2010年。本文介绍的deGLOBE语料库项目，以2010年之后十年左右时间内首次出版和发表的德语书面语为目标语料，可作为现有德语平衡语料库的有益补充。

2　deGLOBE语料库的创建

deGLOBE语料库是按照布朗语料库模式创建的百万词级平衡语料库，主要收集

2010年之后出版和发表的德语书面语文本。该库包括生语料、词性赋码和词形还原三个版本，其中词性赋码及词形还原皆采用TreeTagger标注工具。在此基础上，提供德语词频表（Word List）和短语列表（Phrase List），可供教学与研究之用。目前，deGLOBE语料库已上传至“北外CQPweb多语种语料库平台”（http://114.251.154.212/cqp/）。该在线平台可提供索引分析、搭配计算、词表生成和主题词分析等功能。

2.1 采样方案

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

表1 deGLOBE语料库文本类型及文本数量

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

布朗语料库全称The Standard Corpus of Present-Day Edited American English，

于20世纪60年代由美国布朗大学研制。该语料库是世界上最早的电子化英语平衡语料库。布朗语料库依据均衡采样原则，收集1961年间出版和发表的美国英语书面语文本，确定所收语料的体裁、子体裁类型及各类别的文本数量，这在一定程度上保证了语料库的相对平衡性与代表性，使得所收样本更好地反映语言整体面貌。建成后，布朗语料库产生广泛影响，诸多语料库依据相同或相似采样原则陆续建成，如代表美国英语的FROWN、CROWN和CROWN2021语料库，代表英国英语的LOB、FLOB、B-BLOB、CLOB语料库等。此外，也不乏代表其他英语变体（如印度英语、澳大利亚英语、新西兰英语等）和其他语种（如汉语、保加利亚语、尼泊尔语等）的同类语料库（McEnery & Hardie 2012：98-99）。这类具有相同规模、依据相同采样原则建成的语料库习惯上被称为布朗家族语料库（the Brown family corpora），彼此之间具有较高的可比性。在此基础上可开展多种研究，如针对某一英语变体的历时研究，针对两种或多种英语变体的变异研究，针对两种或多种不同语言的跨语言对比等。deGLOBE语料库及其所属的北外全球语料库集群均借鉴布朗语料库的采样方案，使之与现有布朗家族语料库具有可比性，可开展相关外英、外汉或多语对比研究，从而有效拓宽语言研究的广度，为多语种、类型学研究提供便利。

大体上，deGLOBE语料库沿用布朗语料库的采样原则，涉及新闻、通用、学术、小说4种体裁类型，并可进一步细分为15个子类。在个别子体裁类型的采样过程中，课题组作了适当调整。一方面，布朗语料库的采样原则适用于美国英语特征，其中的子体裁类型N为“冒险小说与西部小说”。由于语言文化与国情存在差异，“西部小说”这一类别不适用于德语语料，故在deGLOBE语料库中未有涉及，该库中子体裁类型N主要收集历险悬疑小说。另一方面，体裁类型随时代动态也相应有所调整，尤其是近些年来随着互联网的发展，涌现出越来越多基于互联网的新型文本类型。考虑到与现有布朗家族语料库之间的可比性，本次建成的deGLOBE 1.0版中并未大规模纳入新型体裁，仍以传统体裁为主，仅在“新闻”这一体裁类型下（包括子体裁A、B、C）适当收入了部分网络新闻（占比约20%），从而更好地维护语料的平衡性与代表性。

2.2　语料采集

确定适用于德语的采样方案后，方可进行语料的采集。deGLOBE语料库将按照上述采样方案采集500个2,000词左右的德语文本并进行初步加工。语料采集过程主要包括文本收集与取样、文本录入以及语料库元信息标注三个主要环节。

2.2.1　文本收集与取样

deGLOBE语料库在第一版的规划中主要面向德国本土的德语文本，暂不涉及其他德语变体（如奥地利德语、瑞士德语等），因此该库中所收文本的第一作者国

籍原则上均为德国，且所收文本均为原创德语，由其他语言译入德语的文本不在本库的收集范围之内。此外，与英语文本相比，德语文本在流通中的总量整体上规模较小。考虑到语料收集的可操作性，deGLOBE语料库与布朗家族语料库相比扩大了语料的时间范围：新闻类语料发布时间为近三年内、其他三类体裁（通用、学术、小说）所收文本的首次出版时间为近十年左右。所收语料的时间与数量[2]分布可见图1。

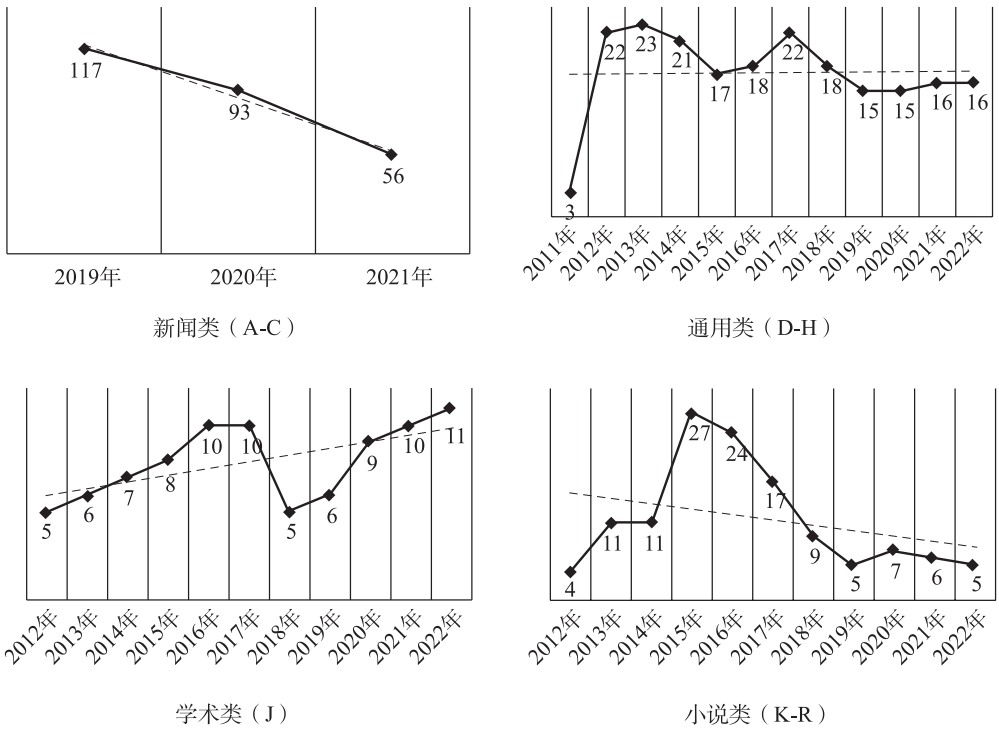


图1　四类体裁所收语料时间数量分布

为尽可能确保语料的平衡性，除考虑语料的时间数量分布外，deGLOBE在建库过程中还兼顾了语料来源及主题的多样性。针对新闻类语料，兼顾全国性的大报、地区小报、新闻周刊以及线上新闻，同时涵盖政治、经济、社会、文化、运动、旅游等多种话题[3]。对于学术类语料，综合考虑核心期刊文章、普通期刊文章、学术专著或文集中的文章，涉及不同学科门类及研究领域。至于通用类、小说类中的书籍，则取自多家出版社，体现多种主题。通用类下的子体裁H“政府或机构公文及文宣”则摘自德国多个政府部门、多家企业及机构。此外，针对篇幅较长的书籍，课题组截取其前、中、后三个部分拼接成一个2,000词的文本，以体现取材在篇章内部的代表性。

2.2.2　文本录入

deGLOBE语料库所收文本长度为2,000词左右，通过正则表达式[A-Za-zÄäÖöÜüß0-9-]+进行计数。当涉及篇幅较短的语料体裁类型，如部分新闻、学术出版物等，则用多篇同类型语料拼合——用于拼合的语料单独保存，在文件名末尾添加A/B/C等以示区分。如deGLOBE库所收的第一号文本包含两篇新闻，因此存储为两个文件：DEA01A、DEA01B，两者共同构成表1子体裁A下的第一个语料文本。当涉及的语料篇幅较长，如不同体裁的书籍，则分别截取书中前、中、后三个部分，存储为A/B/C三个文件，共同合成一个语料文本，如本库中通用类体裁下所收的第一个语料文本：DED01A、DED01B、DED01C。所有语料均经过人工校对，并以UTF-8编码的txt文本文件格式储存。

2.2.3　元信息标注

整体来看，deGLOBE语料库包含500个2,000词长度的文本，共存储为1,348个语料子文件，其命名格式遵循“两位字母语种代码-一位字母体裁编码-数字编号-字母编号”。通过这种方式，仅从文件名就能了解语料所属体裁信息。除文件名本身承载的分类信息外，deGLOBE语料库还以Excel表格记录元信息标注，为后续研究提供更多的语料信息。

针对所有语料文件，deGLOBE语料库提供文件编码、书名或文章名、作者、出版商、出版年份、字数统计六项基本信息。在此基础上，针对不同体裁的语料，进一步提供更多信息。针对新闻类语料，语料库将出版时间精确到新闻发布的日期，并标注了新闻所属栏目及报纸类型。对于通用类、小说类语料，额外提供了所收书籍的ISBN或DOI编码。至于学术类语料，则标注了学术出版物的来源名称、来源类别、所属领域、ISBN或DOI编码。

2.3　语料版本及应用

建成后的deGLOBE语料库共包含生语料、词性赋码和词形还原三个版本。生语料指未添加元信息和语言学标注的文本。后两个版本是借助自动词性赋码软件TreeTagger对语料进行标注后所生成的两个不同版本语料。

表2以DEA03C语料文件中的第一个句子为例，展示三个版本的语料区别。其中，词性赋码版语料以“单词_词性码”（Word_POS）的形式呈现，如非反身人称代词赋码为PPER、限定动词为VAFIN、冠词为ART、作状语或表语的形容词为ADJD、作定语的形容词为ADJA等[4]。词形还原版语料则将原先文本中的所有单词逐一替换为该单词的原形，如将动词过去式war替换为原形sein，将中性不定冠词ein替换为阴性一格eine，将修饰阳性一格名词的形容词kleiner替换为原形klein等。不同版本的语料有助于灵活展开不同类型的研究，例如，基于词性赋码语料检索特定德语句式，以及根据单词原形计算某一词目的词频及其搭配等。

表2　三种语料版本示例

语料版本	语料标注示例	来源文件
生语料	Es war ein vermeintlich kleiner Lapsus: ein einziges falsches Wort.	DEA03C
词性赋码	Es_PPER war_VAFIN ein_ART vermeintlich_ADJD kleiner_ADJA Lapsus_NN:_$. ein_ART einziges_ADJA falsches_ADJA Wort_NN ._$.	DEA03C_POS
词形还原	es sein eine vermeintlich klein Lapsus: eine einzig falsch Wort.	DEA03C_LEM

上述三个版本语料均可以在“北外CQPweb多语种语料库平台”在线使用。平台提供的简单查询（simple query）和CQP专属检索语法（CQP syntax）模式，既能满足普通语料库用户的一般需求，又能实现高级检索和分析。另外，将语料库部署在CQPweb平台上，是把语料库文本按特定格式建成索引后存储于服务器，并不把语料库整体复制到第三方计算机，用户搜索结果只显示有限的上下文语境，通过这种方式，有效规避了deGLOBE所涉及语料的版权问题（许家金、吴良平2014：12）。

3　研究展望

3.1　基于语料库的德语语言学及教学研究

基于deGLOBE语料库可进行多种类型的德语语言学及教学研究。借助deGLOBE全库首先可以进行词汇、短语、句法、篇章语用层面的共时研究。由deGLOBE析出的词频表、短语表，可直接或间接应用于德语教学。此外，deGLOBE语料库中的四种体裁文本也可各自单独成库，为篇章语言学研究、语域变异研究等提供实证基础。

deGLOBE语料库遵循布朗语料库采样方案，追求语料的平衡性、代表性，可与此前建成的系列布朗家族语料库组合使用。首先，deGLOBE可与CROWN、CLOB组合使用，开展德语与英语的对比研究，也可与ToRCH系列汉语语料库（2009，2014，2019）组合使用，开展德汉对比研究，抑或是与前述英语、汉语语料库以及北外全球语料库集群组合使用，开展类型学研究。其次，deGLOBE还可与当前德国规模最大的平衡语料库DWDS进行组合——后者主要收集20世纪的德语语料，尽管在规模上远大于deGLOBE，但在采样上遵循类似的文本分类原则，同样与deGLOBE具有较好的对应性，可在两者的基础上开展针对德语的历时变化研究。

3.2　后续语料库建设

目前，deGLOBE语料库1.0版本已上线。今后还可从多个方面对该库进行扩容，以不断满足教学与语言研究所需。一方面，可以仿照当前德国最大的20世纪平衡语料库DWDS的做法，在核心语料库外建立规模更为庞大的新闻库作为补充语料库。另一方面，也可对学术体裁进行语料扩充，进一步建立学术德语语料库，从而开展学术话语研究。除了当前涉及的四类体裁，在后续版本的deGLOBE语料库建设中，还可以考虑纳入新兴的体裁类型，如网络文本、新媒体文本等，从而促成传统体裁与新兴体裁的对比研究。在条件允许的情况下，也可以考虑收集一定规模的口语语料库，开展德语口笔语语言对比研究。此外，当前deGLOBE语料库仅涉及德国本土的原创德语文本，后续可进一步考虑收集德语的其他变体作为语料。除德国本土语料外，还可收集来自其他以德语作为官方语言的国家的语料，如建立针对奥地利德语、瑞士德语、比利时德语等的平衡语料库；除原创德语外，也可以进一步收集翻译德语语料。

4　结语

尽管面向德语的语料库建设已经较为成熟，现有的德语语料库类型丰富、用途广泛、规模各异，但面向当代德语书面语，尤其是2010年后德语书面语的平衡语料库建设仍然具有现实价值。在“北外全球语料库集群”项目的推动下，deGLOBE语料库按照布朗语料库模式创建，依托第四代语料库分析工具“北外CQPweb多语种语料库平台”提供数据检索与分析功能。研究者可根据实际教学或科研所需，基于deGLOBE自身或者通过deGLOBE与其他可比语料库开展相关研究。此外，该语料库遵循“共建共享”理念，对后续语料库的建设持开放态度，希望并倡导更多同行加入语料库建设，进一步促进德语教学与研究。

注释

1　参照并改编自https://varieng.helsinki.fi/CoRD/corpora/BROWN/basic.html。

2　此处根据实际所收语料篇数计算，因此与表1中的数量略有出入——deGLOBE所收的四类体裁语料中，新闻类、学术类均存在单篇语料词数少于2,000的情况，将由多篇语料拼接成一个文本。

3　新冠肺炎疫情暴发以来的新闻报道较多涉及疫情相关话题。为避免同一话题过多影响语料的平衡性，deGLOBE所收集的新闻语料以2019年为最多来源年份。

4　德语词性赋码集可从https://cis.uni-muenchen.de/~schmid/tools/TreeTagger/data/STTS-Tagset.pdf读取。

参考文献

GEYKEN A. The DWDS corpus: a reference corpus for the German language of the 20th century [C]//FELLBAUM C. Collocations and idioms. London: Continuum, 2007: 23-40.

KAEDING, F. Häufigkeitswörterbuch der Deutschen Sprache [M]. Berlin: Self-published, 1897/1898.

KÜBLER S, ZINSMEISTER H. Corpus linguistics and linguistically annotated corpora [M]. London: Bloomsbury, 2015.

LÜNGEN H. DeReKo – das deutsche referenzkorpus [J]. Zeitschrift für germanistische Linguistik, 2017, 45(1): 161-170.

MCENERY T, HARDIE A. Corpus linguistics: method, theory and practice [M]. Cambridge: Cambridge University Press, 2012.

SCHMIDT T. Gesprächskorpora [C]//KUPIETZ M, SCHMIDT T. Korpuslinguistik. Berlin: Meuton de Gruyter, S, 2018: 209-230.

许家金，吴良平．基于网络的第四代语料库分析工具CQPweb及应用实例[J]．外语电化教学，2014（5）：10-15.

通信地址：100089　北京市　北京外国语大学中国外语与教育研究中心/国家语言能力发展研究中心（周顾盈、宋瑛明）

100089　北京市　北京外国语大学德语学院（舒哲、孙昱、徐亮）

MgmtDEAP管理科学与工程学术英语语料库的创建*

国防科技大学 邓静子 韩正猛 张宇轩 李雨龙 吴禹成 陈 荣 梁 芸

提要：MgmtDEAP管理科学与工程学术英语语料库是“DEAP学术英语语料库”的重要组成部分。本文主要介绍该语料库的建库目标、建库原则、建库过程和应用前景。在详细阐述该语料库的学科分布、期刊选取、语料下载与命名、格式转换、文本清理、文本标注和文本校对等建库步骤后，进而探讨其在语言研究、学科建设、教学、教材和词典建设与经济建设等方面的应用意义。

关键词：管理科学与工程学、学术英语、语料库

1 引言

语料库是一个由大量在真实情况下使用的语言信息经过科学收集和组织而集成的专供研究使用的资料库。语料库语言学是以语篇语料为基础对语言进行研究的一门学科（郭曙纶 2011）。20世纪60年代，语料库语言学的研究开始兴起，当时大型通用书面语与口语语料库的开发才刚刚起步。20世纪80年代，语料库语言学研究达到兴盛时期，但当时的语料库研究方法只着重挖掘语言数据的重要性和重复出现词语的语言规律（布占廷等 2018）。进入21世纪，随着计算机研究方法的深入发展，研究者可以借此进行超大规模的数据采集和加工，因而计算机化的超大型语料库逐渐问世，专业性、针对性较强的中小型语料库的构建也成为可能。

英语语料库有很多不同的类型。从应用层面来看，可以将其划分为通用英语

* 本文系第十批中国外语教育基金项目“管理科学与工程学术英语语料库的研制”（ZGWYJYJJ10B001）的结项成果。邓静子为本文通讯作者。
作者贡献：
邓静子：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；
韩正猛：数据收集；
张宇轩：数据收集；
李雨龙：数据收集；
吴禹成：数据收集；
陈 荣：数据收集；
梁 芸：数据收集。

语料库和专门用途英语语料库。相比客观追求语言原貌的通用英语语料库，专门用途英语语料库有着独特优势，更加符合特定的研究需要，能够给相关领域的研究人员提供大量专业、鲜活、真实的语言材料，从而服务于特定的研究目的，助力特定的研究领域。尤其是在当今计算机技术深度发展、网络高度融合、语料库发展势头正猛的背景下，专门用途英语语料库有着更为广阔的发展天地和更加重要的用途。目前，许多学科和研究领域都相继搭建了专门用途英语语料库，如解放军外国语学院的“军事英语语料库”、北京外国语大学的“学术英语语料库”等（董爱华 2013）。其中由北京外国语大学许家金教授于2016年牵头建设的DEAP学术英语语料库是目前国内最大的学术英语语料库。该库拟建成总容量超过一亿词次，包含人文社会科学、自然科学各主要学科领域的专门用途英语平衡语料库。目前，该语料库下的农学、艺术学、生命科学、化学、土木工程、经济学、教育学、环境工程、地理学、信息科学（计算机科学）、语言学、文学、材料科学、数学、医学（临床）、军事学、新闻学、哲学、物理学、政治学、心理学、船舶与海洋工程学、社会学、统计学等24个子语料库已经建成。整个DEAP语料库的总容量已经达到1.227亿词，包含27,128篇各学科领域论文，已经超过建库的初始目标，成为国内迄今最大的学术英语语料库。本项目作为DEAP学术英语语料库建设项目的子课题之一，主要研究管理科学与工程学的学术英语语料库的构建，目的是建成与上述24个子语料库相似规模的学术英语语料库。该库的建成对于管理科学与工程学的学术研究将具有重要推进作用。

与其他学科领域的语料库建设现状相比，目前国内管理科学与工程领域的学术英语语料库由于已建成数目较少，涵盖的学科领域较狭窄，还处于初级阶段。梁波、黄琨桢（2013）自行创建了小型的药事管理学英语笔语语料库。万雯婷（2014）将30篇管理科学论文的结论部分建成了一个小型学术英语语料库，等等。这些文本语料库一般容量较小，无法全面地反映管理科学与工程学科领域的整体面貌。同时，因为没有确立统一的规范与标准，所以对建库原则和制定语料标注方法的阐述不够全面，所建成的语料库也缺少代表性。鉴于语料库的建设对于管理科学与工程学的学科建设和教学研究具有重要意义，本课题组集体攻关，在DEAP学术英语语料库项目的建库思想和许家金教授的指导下，建立了MgmtDEAP管理科学与工程学术英语语料库。该语料库不仅可以对管理科学与工程学科的学术英语文献信息进行整合，还可以通过语料库软件实现对研究文献的检索、内容关联、文本分析等一系列人工难以完成的工作，进而帮助研究人员了解该学科的国际研究现状、总结国际化研究成果，进而促进国内管理科学与工程学科学术研究与国际学术研究接轨，提高管理科学与工程学科的学术研究水平。本文主要包括两方面内容：说明管理科学与工程学术英语语料库建库的基本步骤（建库目标、语料收集、语料整理等）；阐述该语料库的应用意义。

2　建库目标

MgmtDEAP管理科学与工程学术英语语料库的建设严格按照中国外语教育基金“专用英语语料库建设项目——DEAP学术英语语料库总库”的设计方案和基本要求进行。在建设过程中，我们根据《学位授予和人才培养一级学科简介》（国务院学位委员会第六届学科评议组 2013）确定了管理科学与工程学科下属的10个二级学科，并以此为依据选择相关期刊。在充分考虑该语料库的学科代表性和体现核心期刊论文语言特征的基础上，选取29种高质量英文学术期刊作为目标期刊，从中下载收集了777篇论文，建成库容为780万词次的“MgmtDEAP管理科学与工程学术英语语料库”，服务于该领域的学术研究和教学实践。以下我们从语料收集、语料整理等方面来介绍该语料库的构建过程。

3　语料收集

语料库的构建原则是在进行大规模文献的下载收集前，先确立文献的下载收集标准（分层取样），界定语料库基本特征，然后展开语料收集工作。

语料收集工作将语料库建设分层取样原则和语料库基本特点相结合，在综合考量所选取语料的代表性、平衡性和时效性的基础上，我们按照学科分布、期刊选取、语料下载和命名的工作顺序完成该语料库的语料收集工作。

3.1　学科分布

根据国务院学位委员会、教育部《学位授予和人才培养学科目录（2013年）》，在管理学（代码12）下，所设定的管理科学与工程（代码1201）的学科范围主要包括管理科学、管理系统工程、工业工程、信息管理与信息系统、工程管理、社会管理工程、管理心理与行为科学、电子商务技术、科技与创新管理、服务科学与工程（国务院学位委员会第六届学科评议组 2013）。

管理科学与工程学术英语语料库的语料文本覆盖以上全部学科方向，具有完整性、代表性和平衡性，能够全面展示出该学科的国际化研究现状和基本发展趋势，并能体现出该领域国际学术论文的英语语言特征。

3.2　期刊选取

本课题组以教育部确定的管理科学与工程学的学科领域为基准，参考中科院2019年JCR（Journal Citation Report）收录的关于管理科学与工程的学科代表期刊和Web of Science中管理学专门类别列举的233种期刊，并将其包含的所有核心期刊与管理科学与工程学科下设的各个方向进行逐项比对，按照影响因子降序排列，选取了33种核心期刊。在选取过程中，为了平衡各个学科方向选取的期刊数量，

我们在各个学科方向中至少选取1种期刊（但至多不超过4种）。通过咨询管理科学与工程学科专家学者的建议后，本课题组确立了29种国际核心期刊，具体信息如表1所示。

表1　29种管理科学与工程学科国际核心期刊信息

序号	类别	期刊名称
1	管理科学	*Omega-International Journal of Management Science*
2	管理科学	*Management Science*
3	管理科学	*European Journal of Operational Research*
4	运筹学与管理科学	*Journal of Operations Management*
5	运筹学与管理科学	*M&SOM – Manufacturing & Service Operations Management*
6	管理系统工程	*Socio-Economic Planning Science*
7	工业工程管理	*Reliability Engineering & System Safety*
8	工业工程管理	*Journal of Product Innovation Management*
9	工业工程管理	*International Journal of Production Research*
10	信息管理与信息系统	*Journal of Strategic Information Systems*
11	信息管理与信息系统	*Information & Management*
12	信息管理与信息系统	*Journal of the American Medical Informatics Association*
13	工程管理	*International Journal of Technology Management*
14	社会管理工程	*Strategic Management Journal*
15	社会管理工程	*Corporate Social Responsibility and Environmental Management*
16	管理心理与行为科学	*Annual Review of Orgabizational Psychology and Organizational Behavior*
17	管理心理与行为科学	*Leadership Quarterly*
18	管理心理与行为科学	*Personnel Psychology*
19	管理心理与行为科学	*Organizational Research Methods*
20	电子商务技术	*Journal of Computer-Mediated Communication*
21	电子商务技术	*Decision Support Systems*
22	电子商务技术	*Electronic Commerce Research and Applications*

（待续）

（续表）

序号	类别	期刊名称
23	电子商务技术	*Telecommunications Policy*
24	科技与创新管理	*Journal of Innovation & Knowledge*
25	科技与创新管理	*Technovation*
26	科技与创新管理	*Journal of Knowledge Management*
27	服务科学与工程	*Journal of Supply Chain Management*
28	服务科学与工程	*Supply Chain Management – An International Journal*
29	服务科学与工程	*Journal of Service Management*

3.3　语料下载与命名

如前所述，本课题组选定了29种核心期刊，并下载其中论文。在下载过程中，为使抽样论文尽可能具有代表性，我们首先选用被引频次作为下载标准。但在下载过程中发现，由于受时间存续性的影响，所选论文发表时间越早，该论文被引用的概率就越大，不能简单以“被引频次”来判定所选论文的代表性。据此，为确保所选论文的代表性，我们重新议定期刊论文的下载标准，即每一年的期刊论文收集数量按照权重进行分配（权重=本年内该期刊文献数量/三年内该期刊文献数量），计算得出每种期刊从2018—2020年共收集27篇论文。这27篇论文根据该期刊年度被引频次由高至低来确立，收集格式为PDF或HTML文件。以下为论文语料下载方法的基本描述。

首先，在Web of Science数据库里检索选定的期刊名称，得到该期刊的全部文献列表。其次，将“出版年份”设定为2018，按“被引频次”排序，从高到低选取文本进行下载，并将其中的通信类、会议类等其他类型文体舍弃（章柏成、杨玲 2020）。之后，依次将出版年份调整至2019年、2020年，重复上述文献检索下载过程。以平均分配为基本原则，每种期刊每年下载9篇文档，但也根据被引频次对年度期刊数量进行弹性调整，优先考虑被引频次高的论文。以期刊*Journal of Computer-Mediated Communication*为例，根据以上步骤，该期刊2018年下载论文13篇，2019年下载论文12篇，2020年下载论文3篇。最终课题组共收集777篇HTML或PDF格式的论文，语料库规模达到780万词次以上，基本满足该语料库的建设要求。

为方便该语料库与DEAP语料库总库汇总，MgmtDEAP管理科学与工程学术英语语料库采用“学科方向–期刊（序号）–年份–文献类型–序号”的顺序，对所采集的文件进行分类命名。其中，学科方向采用汉语拼音首字母表示，如运

筹学与管理学科的代号为YCXYGLKX；期刊采用Journal的首字母J，并附上该期刊在课题组确定的期刊列表中的序号；文献类型主要分为Article、Correction、Editorial Material、Proceeding Paper、Review五种类型，均采用英文单词的前两个字母表示（见表2）。例如，文件名YCXYGLKX–J4–2018–AR1表明该文本是运筹学与管理科学方向下第4种期刊2018年的第1篇文章，文献类型为论文。

表2　文献类型的标注符号

文献类型	缩写
Article	AR
Correction	CO
Editorial Material	EM
Proceeding Paper	PP
Review	RE

4　语料整理

语料整理环节主要按照格式转换、文本清理、文本标注、格式转换和文本校对等步骤进行，见图1。

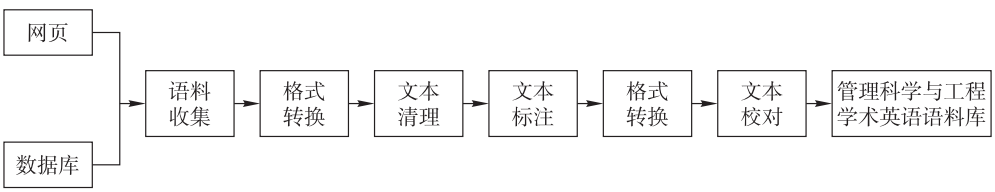


图1　管理科学与工程学术英语语料库语料整理步骤

从文献数据库中下载的原始论文语料，经过格式转换后会出现内容缺失、排版混乱、数据噪声等问题，因此需要对文本进行深度清理，再对清理后的语料进行标注和文本校对。上述工作步骤环环相扣，依次递进，逐步推进，形成了语料库语料整理的技术闭环系统。

4.1　格式转换

从 Web of Science 数据库中下载的2018—2020年的代表性论文共777篇，多为PDF文件格式，而语料库要求的文本格式为UTF-8编码的TXT格式，因此PDF

文件格式文献无法直接用于语料库研究，需要对下载的PDF格式文献进行文献格式的转换。本课题组经过研究决定，先将PDF格式文献转换为WORD格式文献，再将WORD格式文献转换为TXT格式文献。由于所选文献数量较多，我们首先使用Abbyy Finereader软件将下载的PDF格式文献批量转换为可编辑的WORD格式文献，转换后的WORD格式文献仍以原文献名称进行命名。但转换后的WORD格式文献与原PDF格式文献相比，内容和格式都存在较大出入，如图2和图3所示。

3.2 | Methodology

3.2.1 | Difference-in-differences

To examine whether firms increase their CSR following the rejection of the inevitable disclosure doctrine, we use a difference-in-differences methodology based on the 14 treatments listed in Table A2 (Online Appendix S1). Our methodology follows Bertrand and Mullainathan's (2003) application of the difference-in-differences methodology in the presence of staggered treatments at the state level. Specifically, we estimate the following regression:

$$KLD_{it} = a_i + a_j \times a_t + a_r \times a_r + \beta \times IDD_{st} + \mathrm{g'X}_{it} + \mathrm{e}_{it}, \quad (1)$$

图2　原PDF格式文献摘取部分

■**3.2** | Methodology

■**3.2.1** | Difference-in-differences

To examine whether firms increase their CSR following the rejection of the inevitable disclosure doctrine, we use a difference-in-differences methodology based on the 14 treatments listed in Table A2 (Online Appendix S1). Our methodology follows Bertrand and Mullainathan's (2003) application of the difference-in-differences methodology in the presence of staggered treatments at the state level. Specifically, we estimate the following regression:

$KLD_{it}=a_i+a_j$ X a_t+a_r X a_t+ 胃 X $IDD_{st}+$ Y $x'+$ 如

图3　转换后存在问题的WORD格式文献摘取部分

为保证文献格式转换质量，我们先将PDF格式文献转换为可搜索的PDF文档，再用Abbyy Finereader软件将PDF文档转换为WORD格式文本。

4.2　文本清理

PDF格式文献转换成WORD格式文献后的文本仍然存在诸多问题，如文献内容和格式不匹配、版本文字识别错误、符号标点错误、文字排版错误、图片错位等问题，这就需要由人工来完成文本清理工作。当WORD格式文献转换为TXT格式文献后，与原来的PDF格式文献的差距将进一步加大，影响该语料库的整体质量。为此，我们对转换后的WORD格式文献进行了以下三个方面的文本清理工作。一是制定文本清理手册，统一指导文本清理工作。遵循内容一致、错误排查、

整洁美观的原则展开文本清理工作。二是逐行与原PDF格式文献进行比照修改文献内容。在文献内容核对中，先是复制替换WORD格式文献中的错误内容，同时采取手动输入方式修改文献错误内容。如果文档中出现大面积的重复错误，则采取批量替换的方式进行清理，主要涉及中文字符、分页符、全角符号及部分字词等。三是保留论文的主体内容，即文章的标题、摘要和正文，删除所有与正文无关的内容，如尾注、脚注、参考文献等。至此，文本清理工作基本完成。

4.3 文本标注

语料库标注是指针对语言处理任务的需求，按照预先制定好的标注原则、规范和操作规程，为语言单位标注恰当标记符的过程，其结果是带有标注信息的语料库（邢富坤 2015）。文本信息的产生与语境条件有关且从中反映出不同的信息交流目的，所以有必要对语料进行对比分析（陈峰 2021）。因此，标注文本信息可以为语料库研究提供检索与分析的条件和依据。语料标注的内容通常分为元信息内容标注与文字结构信息内容标注两个方面。由于文本元信息的内容已由课题组收录到编写的语料库索引表中，因此本阶段标注的对象仅为文本结构信息。由于可扩展标记语言（extensible markup language，简称XML）的扩展性与交换性较好，有利于其在任何应用程序中读写数据（胡佳佳 2011），我们采用XML的方法来标注文本结构信息。需要特别注意的是，在XML语言中，所有的标志都应该成对存在，即有一个开始标志，也应该有一个终止标志，如<Tl> </Tl>，前者为起始标记，后者为结束标记。因此，在人工标注过程中，要注意对语料进行成对标注。由此，本课题组采用人工标注的方式对清洁后的文本进行标注。标注的对象包括文章的标题、摘要，以及文中的数学公式、图表等，如表3所示。

表3 文本结构信息标注对照表

标注符号	标注内容
<T>	图表
<G>	公式
<Tl> </Tl>	文章标题
<H> </H>	章节标题
<L> </L>	列表信息
<A> </A>	摘要

需要注明的是，文中章节标题标注的层级仅涉及一级标题和二级标题。另外，

如果数学公式或图表出现在一句话中间，需要在标注后将断裂的句子还原到标注符号后面。

在标注工作开始前，本课题组成员参照总项目要求和已有的语料库研究成果，界定了标注内容和规范标准，并制定了详细的标注方案。课题组成员在文本标注过程中，遵循标注的规范标准，严格执行标注工作流程，有效地保证了文本标注的精准性和完整性。由于本研究构建的语料库具有一定规模，课题组使用标注辅助工具AnnoTool（界面如图4所示）开展标注工作。该软件具有操作简单的特点，可以将固定的标签格式自动添加到文本中，并且支持用户自定义设置，可有效提高标注工作效率，减轻人工标注的负担。

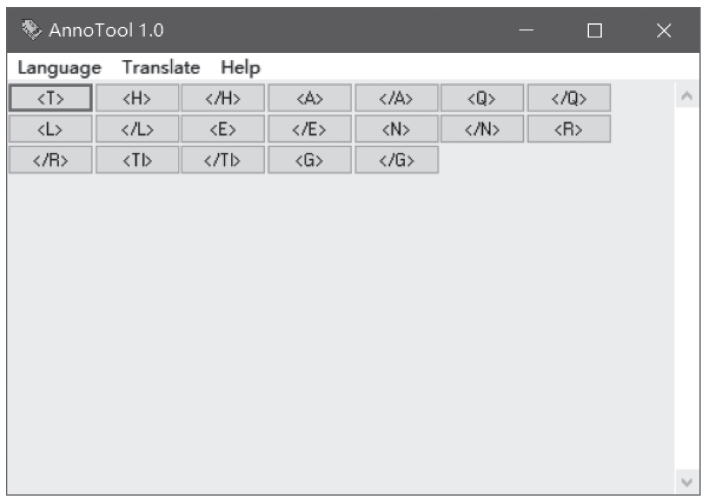


图4　标注辅助工具AnnoTool

4.4　文本校对

为确保文本标注结果的准确性和可靠性，在完成上述工作步骤后，本课题组对全部文本进行了两次全方位的校对。第一次是在完成文本标注工作以后，课题组成员对各自完成的任务进行交叉检查。检查的主要内容包括：（1）标注后的文本内容是否与原文献一致；（2）文本格式及标点符号是否符合规范；（3）文本标注内容是否完整、准确；（4）起止和终止标记的数目是否一致。交叉校对后发现标注后文本存在以下问题：一是存在乱码和中文字符；二是存在英语语法错误。相关文本负责人对有问题的文本进行了修改。第二次是在WORD格式文献转换成TXT格式文献后，课题组成员对各自完成的任务进行自查。主要检验语料能否在语料库分析软件上正确运行并显示出分析结果，为语料库分析工作做准备。

综上，本语料库的构建遵循最大程度“还原”原有文献面貌的原则，通过对

文本选取、文本转换、文本清理、文本校对等步骤的严格把控，力求保证所采集语料的完整性、真实性、代表性和精准性。需要说明的是，由于MgmtDEAP管理科学与工程学术英语语料库的构建涉及特定学科领域的专业研究文献，专业知识宽泛，知识内容比较复杂，学术性较强，该语料库建设难免存在疏漏和不当之处。希望该领域的专家学者给予谅解和批评指正，以期进一步完善该语料库的建设发展。

5 基于MgmtDEAP管理科学与工程学术英语语料库的应用研究

MgmtDEAP管理科学与工程学术英语语料库将被收录于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，可以应用于管理科学与工程学科领域的以下方面。

5.1 语言研究

根据语言材料的采集和使用途径，现代语言学研究的方法主要有三种，即内省法、诱导法和语料库研究方法（孙云莉 2021）。不同的研究方法对应着不同的语料收集方式。其中语料库研究方法遵循语言研究的规则，按照不同学科领域的文献发表情况对已有学术论文进行抽样，收集具有代表性的语料，组成一个储存于计算机的文本库（桂诗春、杨惠中 2003）。通过语料库，语言学学者可以利用计算机技术对学术论文中的语料进行拆解与研究，兼顾定量与定性分析，使研究过程更加高效、精准，使研究成果更为可靠、科学，在语言研究方面具有广阔的应用场景，为业界学者提供科学研究的方法工具。 因此，MgmtDEAP管理科学与工程学术英语语料库的建立，可以客观地反映出该学科领域国际核心期刊论文的语言特征，让研究者能直观地总结出该学科论文中经常出现的词汇、句子、语法，总结出论文的语言特点，并由此理解其语义与功能，从而支撑该领域高质量英语论文的文体研究、篇章研究和英汉双语对比研究，进而改变现有管理科学与工程语言研究方法的局限性，填补该领域大型学术英语语料库建设的空白。

5.2 学科建设

作为国内管理学中唯一按照一级学科招生的门类学科，管理科学与工程学科领域涉及不确定性决策研究、公共工程、资源优化等多个方面，在国内外学者都热衷于对其下属领域进行相关研究的同时，该语料库更多聚焦于管理科学与工程学科的整体性研究。学者们借助该语料库，通过使用计算机技术和定量分析方法，梳理出该学科的整体科研成果和国际化研究现状，了解学科的发展历史，比较学科不同时期的发展热点和存在的不足，从而对管理科学与工程学科有一个全景式的整体描述，这对于该学科的建设发展将起到重要的推进作用。

5.3　教学方面

该语料库可以梳理出管理科学与工程学科的核心理论和热点问题，从而确定学科课程教学必须涵盖的核心教学内容，进一步提升课堂教学内容质量，帮助学生们了解和掌握这门学科的内涵和外延。在此基础上，该语料库还可以为开展本领域的学术英语教学提供资源和教学素材，从而推动相关院校、相关专业的特色化英语教学建设。

5.4　教材和词典建设

由于该语料库语料覆盖管理科学与工程学的不同学科方向，可以为编写该学科教材提供真实的语言数据，并满足该学科学术英语词典编撰者的不同需求，从而增强管理科学与工程学术英语教学的针对性和实际效果。

5.5　经济建设

该语料库的建立可以为我国经济结构建设、供给侧改革，社会治理和其他领域的科学管理提供学术性的参考意见；为上市公司、中小型企业的产业转型和制度管理提供学理依据。

6　结语

本文主要介绍了MgmtDEAP管理科学与工程学术英语语料库的建库目标、建库原则、建库过程和建库意见，详细阐述了该语料库的学科分布、期刊选取、语料下载与命名、格式转换、文本清理、文本标注和文本校对等建库步骤。在建库过程中，本课题选取了777篇发表于SCI管理科学与工程学科领域核心期刊的论文，发表时间集中于2018—2020年。本课题组成员在完成文献收集工作后，对所收集论文进行了分类整理，并对文中特定的语料单位进行清理、标注，最后转换成UTF-8编码的TXT格式文献，最终建成管理科学与工程学术英语语料库。该语料库涵盖了管理工程与科学学科下10个子方向的研究成果，文献内容时效性强，可以服务于管理科学与工程学的科研与教学工作，并为学者们提供了全新的研究视角和研究方法工具。同时，该语料库的建立进一步拓宽了国内学术英语的研究领域，增加了管理工程与科学领域的学术英语语料素材。值得肯定的是，MgmtDEAP管理科学与工程学术英语语料库的构建，将为专门用途英语语料库的建设提供有益参考，并推动管理科学与工程的应用与研究（闫鹏飞、谢文龙2020）。这也是英语语言学界人士感到受益的事情。我们相信，MgmtDEAP管理科学与工程学术英语语料库建设及其相关应用研究，必将大有可为。

参考文献

布占廷，王昕，王乐. LinDEAP语言学学术英语语料库的创建[J]. 语料库语言学，2018（2）：78-90.

陈峰. 化工英语语料库的构建与应用前景[J]. 材料保护，2021（3）：9-10.

董爱华. 专门用途语料库的建设、应用、问题与发展趋势[J]. 北京印刷学院学报，2013（5）：59-74.

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.

郭曙纶. 汉语语料库应用教程[M]. 上海：上海外语教育出版社，2011.

国务院学位委员会第六届学科评议组. 学位授予和人才培养一级学科简介[M]. 北京：高等教育出版社，2013.

胡佳佳.《说文解字》语料库的XML标注设计[J]. 社会科学论坛，2011（7）：214-223.

梁波，黄琨桢. 小型药事管理学英语笔语语料库的建设与后续研究初探[J]. 语文学刊，2013（9）：98-111.

孙云莉. 语料库语言学研究的概述、现状和前景[J]. 英语广场，2021（28）：62-65.

万雯婷. 基于语料库的管理学论文结论部分的体裁分析[J]. 语文学刊，2014（10）：165-168.

邢富坤. 面向语言处理的语料库标注：回顾与反思[J]. 解放军外国语学院学报，2015（3）：8-13.

闫鹏飞，谢文龙. MatDEAP材料科学学术英语语料库的创建[J]. 语料库语言学，2020（1）：97-106.

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.

通信地址：410000　湖南省长沙市　国防科技大学文理学院

《对比语言学研究新路径：实证与方法论的挑战》述评

北京航空航天大学 葛恬馨

Renata Enghels, Bart Defrancq & Marlies Jansegers (eds.). 2020. *New Approaches to Contrastive Linguistics: Empirical and Methodological Challenges*. Berlin: Walter de Gruyter. v+312 pp.

1 引言

传统的对比语言学研究主要依靠定性分析，借助高级统计方法的定量分析较少。自 2008 年以来，语言学研究经历了量化转向（Janda 2013），实证数据、高级统计方法以及可视化工具备受青睐。在这种背景下，对比语言学研究愈发面临着来自两个方面的挑战：语料的适用性问题以及高级统计方法的匮乏。

《对比语言学研究新路径：实证与方法论的挑战》一书，对上述挑战作出了及时回应，为对比语言学向纵深发展提供了适时补充。该书由九篇实证性研究组成，详细介绍了对比语言学研究中使用的实证数据、高级统计方法以及可视化技术。

2 内容概述

该书第一章，Renata Enghels、Bart Defrancq和Marlies Jansegers对传统对比语言学研究中使用的数据和方法进行了回顾与反思。本章重点介绍了对比语言学面临的两个挑战，即如何挖掘新型实证数据以及怎样在对比语言学研究中应用高级统计方法。本章还指出了多数对比语言学研究在选择共同对比基础时都面临的问题，即平行翻译语料中出现的翻译共性与翻译腔的问题。除了介绍传统对比语言学研究中常用翻译语料的优缺点外，本章还介绍了其他可应用于对比语言学研究的新型实证数据，如多语影视字幕语料库和多语口译语料库等。

第二章内容为Hans Boas对于语义框架能否作为对比研究中共同对比基础的探索。作者详细介绍了该研究中使用的Berkeley FrameNet项目，该项目可帮助判断由英语派生而来的语义框架能否成为通用框架，从而成为对比研究中的共同对比

基础。作者对“问询”框架进行了分析，该框架由英语词汇“to ask”及其德语对应词fragen组成。基于语义框架理论，作者提出语义框架成为通用框架需满足三个条件：翻译对等、配价对等以及文化对等。该研究对跨语言通用语义框架的对比研究具有一定启示。

第三章中Stefan Gries、Marlies Jansegers和Viola Miglio以跨语言近义现象为研究内容，探索如何在对比语言学研究中恰当地引入高级统计方法和可视化工具。考虑到跨语言现象的复杂性和多因素特征，该研究运用了行为概貌向量（Behavioral Profile Vectors）处理语料，对跨语言近义现象进行量化分析。该研究以西班牙语动词sendir(e)（感受/到）及其法语与意语对应词为例，探索了跨语言近义现象中的词义多样性、原型性以及差别变量的识别问题。针对词义多样性问题，作者使用了分层聚类分析、模糊聚类方法以及网络分析。对于原型性问题，作者使用随机森林模型对数据进行了分析。该研究为对比语言学研究中使用高级统计方法与可视化工具提供了具体范式，也为阐明跨语言近义现象的差异提供了重要启示。

第四章中Pauline de Baets、Lore Vandevoorde和 Gert de Sutter为验证对比语言学研究中平行翻译语料的可靠性问题，使用了行为概貌分析、对应分析和语义镜像法，对荷兰语中起始词的语义进行了可视化分析，并将其与翻译语料库中的对应词进行对比。研究发现，由于认知固化和过度标准化（Teich 2003），翻译人员倾向于过度使用目标语中出现频率较高的结构，导致原语与译文中起始词素的分布频率及其内部语义结构之间存在巨大差异。该研究向对比语言学研究中使用翻译语料库的方法提出了质疑，指出仅使用翻译语料的对比研究应慎重考虑此种数据自身存在的缺陷。

德语比英语更具名词性的说法广为流传，然而这一论断尚未得到实证研究的支持。在第五章中，Stella Neumann基于CroCo语料库对这种说法的可靠性进行了验证。该语料库包含德语及其英语译文，并以语对方式呈现。由于CroCo语料库将原语与译文进行了对齐处理，该研究选择了库中的德英对等名词作为比较基础。为了避免由于拼写差异而导致的标记化差异和不规范采样方法，该研究在三个包含100个句子的随机样本中对德语和英语中普通名词的分布进行计数，并使用泊松回归模型进行统计分析。研究发现，德英两种语言的名词频数分布规律差异较小，两种语言的名词性程度不具有显著差异，从而在一定程度上纠正了德语比英语名词性更强的说法。

第六章介绍了Bart Defrancq和Camille Collard的研究，该研究基于英荷议会记录中摘录的口语语料库、欧洲议会记录整理的口语语料库和欧洲议会同声传译语料库，考察了口译语料作为对比语言学研究数据的可行性。作者使用行为概貌法和多元线性回归模型，比较了口译和非口译语境中英语和荷兰语know/weten、

speak/zeggen两组同义动词对，分析动词语义在上述两种语境中是否保持一致。该研究发现，动词在口译语境中的用法与在原语和译文中有所不同，而且译员在口译过程中倾向于用原语中使用频率较高的语法结构取代频率较低的结构。研究表明，尽管口译效应真实存在，但其效应并不显著，因此口译语料可以作为对比语言学研究的数据来源。

第七章中Tom Bossuyt和Torsten Leuschner将小型多语对比语料库（ConverGENTie corpus）与大型单语语料库（DeReKo和SoNaR）相结合，以英语小品词“-ever”与其德语对应词immer/auch和荷兰语对应词(dan)ook为研究对象，对英语、德语及荷兰语小品词的分布模式及频率进行了比较。该研究将小型多语对比语料库与大型单语语料库相结合，为未来的对比语言学研究提供了一个新范式。前者能够保证对比语言学研究中跨语言现象的可比性，后者则为对比语言学研究提供了大量的原语数据来源。研究表明，英语、德语与荷兰语小品词的差异源于非关联标记（irrelevance-marking）语法化的不同程度，英语的语法化程度最高，德语次之，荷兰语最低。

第八章中Silvennoinen基于大规模平行翻译语料库Europarl，使用多重对应分析法（Multiple Correspondence Analysis），对 11 种欧洲语言的对比否定结构进行了跨语言对比。对比否定（contrastive negation）指将一个否定元素与一个肯定元素结合，从而使该肯定元素取代否定元素的一种否定表达方式。该研究主要考察带有纠正连词的对比否定结构。作者使用多重对应分析法，将带有纠正连词的对比否定结构与其他对比否定结构进行比较。研究表明，尽管在跨语言对比否定现象中发现了较大的翻译效应，但从整体上看，平行语料库数据仍可用于跨语言对比研究。

第九章中Åke Viberg基于大规模影视字幕语料库、英国国家语料库（BNC）以及瑞典语语料库（KORP），比较了瑞典语和英语中的“切割”类与“打破”类动词的语义特征，研究发现，“切割”类动词蕴含“完全”“彻底”的分割义，而“打破”类动词蕴含“不完整”“杂乱”的分割义，“切割”类和“打破”类动词的语义差异与所用器具和手部动作有关。另外，该研究还阐述了对比语言学研究中的三个基本问题：语料对等性、语料真实性和语料代表性。语料对等性指跨语言对应现象之间的对应程度；语料真实性指在翻译效应影响下语料代表真实语言使用的程度；语料代表性指使用的语料库对所要对比的语言具有多大代表性。

3 简要评价

《对比语言学研究新路径：实证与方法论的挑战》为对比语言学领域提供了实证数据与方法论上的新见解与新启发。在实证数据方面，该书阐释了传统对比语

言学研究中使用平行翻译语料库的缺陷，介绍了对比语言学研究中更新颖可靠的实证数据来源，进一步拓宽了该领域的研究范围。另外，在开拓方法论方面，该书介绍了语言学领域当下时兴的多种高级统计方法和可视化工具，例如混合效应逻辑回归、泊松回归分析、多重对应分析以及行为概貌分析等，用各种实证研究详细阐释了高级统计方法与可视化工具如何应用到对比语言学研究中。

除了该书的上述优点及其为对比语言学研究提供的启发与灵感之外，对比语言学研究还可吸纳以下几条建议。第一，该书涵盖的语言类型主要来自日耳曼语和罗曼语系，对其他种类语言的探讨和研究相对匮乏，因此未来的对比语言学研究可以吸纳更多种类的语言。第二，书中介绍的高级统计方法及可视化工具对于较少接触统计学知识的语言研究者来说是相对晦涩难懂的，因此未来研究可以对复杂的多元统计方法及其原理进行更加详细的解释。第三，应用高级统计方法与可视化工具的研究应更加注重基于统计结果的理论解释，以揭示统计结果背后深层的理论问题。

总体而言，该书为未来的对比语言学研究开辟了新视野，促进了研究人员利用庞大的、新型的语料库获取大量的实证数据，以及从社交媒体中对潜在数据进行提取和汇编。另外，这本书极力主张在对比语言学研究中引入高级统计方法与可视化工具，这些技术在先前对比语言学研究中没有得到充分的重视和应用。

参考文献

JANDA L. Cognitive linguistics: the quantitative turn [C]. Berlin: De Gruyter Mouton, 2013.

TEICH E. 2003. Cross-linguistic variation in system and text: a methodology for the investigation of translations and comparable texts [M]. Berlin: Mouton de Gruyter, 2003.

通信地址：100191　北京市　北京航空航天大学外国语学院

《通过语料库方法对语言分析进行三角论证》述评*

北京师范大学 梁悦怡 王德亮

Jesse Egbert & Paul Baker (eds.). 2020. *Using Corpus Methods to Triangulate Linguistic Analysis*. New York: Routledge. xiv+286pp.

近年来，随着语料库研究方法的发展和进步，基于语料库的实证研究也得到了广泛应用（Baker & Egbert 2016）。2020年，Jesse Egbert 和 Paul Baker共同编写了《通过语料库方法对语言分析进行三角论证》一书。混合方法的研究经常会提到“三角论证”（Triangulation）这一概念，它是指使用两种或以上的研究方法或研究视角解决一项研究中提出的一个或多个问题。本书收录了九项具有代表性的、结合语料库三角论证研究方法的实证研究，集中反映了这个领域的创新性进展。

1 内容简介

本书共十一章，具体按领域可划分为三大部分（除引言和结论章节外）。首先，第二至四章应用了语篇分析的研究方法；其次，第五至七章结合了应用语言学的研究方法；最后，第八至十章借鉴了心理学实验的研究方法。这九章都属于实证研究的范畴，旨在提高人们对语料库语言学研究方法中效度的重视，每一章都对语料库中的数据进行三角论证分析，横跨了三个语言学主流领域。

在第一章引言部分，编者讨论了语料库研究结合三角论证方法的巨大潜力，阐明了两者的互补性。二人首先明确了本书使用三角论证研究的标准。“三角论证”一词的定义众说纷纭，与多方法/多领域研究之间的界限较为模糊，本书采用的规则是范围较广的定义，即三角论证研究是指“应用了两种或以上独立方法来回答一个研究问题的研究”（Egbert & Baker 2020：6）。接着，编者介绍了语料库

* 本文系教育部人文社会科学研究一般项目“对话句法理论的发展及应用研究”（20YJA740041）的阶段性成果。王德亮为本文通讯作者。
作者贡献：
梁悦怡：数据收集、数据分析、讨论结论、初稿撰写、字数占比（90%）；
王德亮：选题构思、研究方法、讨论结论、字数占比（10%）、修改润色。

语言学和三角论证的特点，说明语料库研究结合三角论证方法会更具有说服力，由此突出本书的目的和重要性。正如引言所述，Egbert和Baker二人邀请了各个领域的专家比较典型地使用语料库进行三角论证，突出其研究设计思路，结合不同领域的研究方法回答研究问题，形成了该书收录的九篇文章。

第二到第四章属于话语分析范畴。第二章为Schnur和Csomay探索学术讲座话语的结构功能，旨在寻求为学术讲座语篇进行断句编码的最佳方法。他们对比“人工编码”（MTurk软件）和“电脑编码”（TextTiler软件）两种方法，结果显示两种方法分析的侧重点不同。他们认为混合型方法最为理想，人工分析能深入地揭示构成文本结构的功能模式，电脑分析则能综合性地概括文本结构的衔接特点。

第三章中，McEnery、Baker和Dayrell三人尝试用历时语料揭示英国19世纪干旱天气的面貌。他们提出了一个引人深省的问题，即历时语料库能否反映过去的历史现实。为了回应此问题，他们选取了8个英国本地报刊，自建了一个超过50亿词的历时语料库，然后结合文本分析和地理信息系统分析方法，统计19世纪英国干旱天气出现的频率和在时空上的分布。通过语料库研究发现：英国的干旱天气发生时间比官方记录的更长、更频繁。

Paul Baker在第四章同样使用报刊语料分析英国媒体对“过度肥胖”一词的态度。Baker选取2012—2016年《每日邮报》（*Daily Mail*）对“过度肥胖”的多篇新闻报道，采用搭配分析法和对文本进行编码分类的分析方法。在处理语料数据时，利用ProtAnt快速归纳关键词，选择最合适的语料文本，以减少庞大的数据处理量。两种方法的研究均发现，“过度肥胖”一词在英国报刊媒体中常被看作是一个尚待解决的问题，而且报刊多使用程度更深的词汇，如guzzle and gorge，隐含“放纵、贪婪和缺乏自制力”等负面含义。

第五章至第七章属于应用语言学范畴。第五章中Laflair、Staples和Xun Yan三人使用英语能力测试的语料库对测试者的语料进行分析。第一次分析他们首先按测试者的程度对语料进行分类，然后用语料库方法对测试者英语写作能力评估中的41个因素进行多维分析，最后得出5个最能反映语言能力的因素。同理，第二次分析使用了测试者口语语料，发现了3个最强的主效应。在经过语料库分析后，发现测试量规与本研究的发现不匹配，意味着语言测试研究者需要考虑更多维度的语言特征。

第六章中Dana Gablasova结合心理语言学实验法和语料库分析方法探究双语者如何通过词汇联想习得第一和第二语言的新词汇。研究过程分为两步：第一步是设计词汇联想实验，让参与者一边收听朗读音频，一边阅读包含线索词的文章，最后提供线索词让参与者联想与其相关的词；第二步是利用语料库进一步分析参与者联想的词汇在语料库中出现的频率。研究发现：参与者联想的词汇是在语料库中出现的高频词，这意味着在双语者新词汇习得机制中，词频是一个很重要的

因素。

第七章中Egbert和Davies探究英语名词+名词结构（NNs）中意义关系类型和历时语义的变化，并对语料的编码处理提出了有效建议。作者对COHA语料库中1500个NNs结构语料进行编码分类和统计分析。研究发现：NNs结构随着时间推移在英语写作中使用越来越频繁，并正在从具体意义转向抽象意义。最后作者指出使用语料库对语料进行编码时，研究者需细心考虑如何增加人工编码的效度，必要时要进行多次预试验。

第八至第十章属于心理语言学范畴。第八章中Hughes和Hardie结合语料库和脑科学实验的脑电图技术（EEG）研究词汇共现的心理机制。他们试图从脑科学中寻求词汇启动理论的实证证据，探究人们加工固定词组和非固定词组时脑电图的差异。首先，作者通过BNC语料库检索形容词+名词的词组搭配，提取出现频率最高的词组。然后，让16个参与者阅读屏幕上的固定词组和非固定词组，对脑内激活区域进行定位，最后比较参与者的脑电图结果。脑科学研究结果证实了人们加工固定词组和非固定词组时大脑机制是有差异的。

第九章中，Gries使用构式搭配分析法探究英语学习者和英语母语者在双宾语和与格结构中动词使用的偏好。首先，从BNC语料库中提取英语母语者语料，使用构式搭配分析法找出母语者在这两种结构中的动词偏好。然后通过实验法招募德国英语学习者完成句子填空，找出学习者在两种结构中的动词偏好。最后通过对比发现德国英语学习者和英语母语者在双宾语和与格结构中偏好使用的动词没有显著差异，说明英语母语者和英语学习者均受到结构图式的强烈影响。

第十章中，Ellis结合心理语言学和语料库语言学的研究方法探究人们如何加工抽象的动词论元构式（VACs）。首先，他提出一种双向统计分析方法“ΔP（delta P）”对BNC语料进行分析，发现某些动词偏好出现在某种动词论元构式中，反之亦然。之后，他利用实验法探究同样的研究问题，提出的假设是高频词比低频词阅读速度更快，并招募28名大学生参与命名实验。实验结果发现：参与者偏好先阅读动词，这意味着语言使用者可能具备对动词论元构式图式的隐性知识，与语料库分析结果一致。

第十一章是本书的结论部分。Egbert和Baker 二人总结了第二至第十章的研究问题和方法，重申三角论证的优缺点和未来巨大的发展潜力。二人得出的结论是虽然三角论证的研究更费时费力，而且对研究者的要求更高，但使用这种方法能够更有效和更确切地解决研究问题，为此付出更多劳力绝对是值得的。

2　简评

结合语料库研究，利用三角论证，进一步验证各种语言现象是本书的最大

亮点。语料库的研制逐步成熟，能够为人们提供大量真实的自然语料，基于语料库的研究也逐渐增多，但语料库本质上并不是一种理论或方法，而是数据和语料的来源，我们必须融合其他领域的方法或理论才能找出零散语料中的语言规律（Gries 2016），本书正是弥补了这方面的不足。首先，它成功地为语言学研究者们提供了清晰可行的混合型研究路径和方法，能有效增加研究的效度，研究结果有较强的说服力。例如本书第七章，Egbert和Davies二人在研究名词+名词结构的历时性变化时，结合了量化和质性研究的方法，在分析数据前先进行人工编码，但发现这种方法出现了较大误差并提出质疑，于是他们重新设计分类列表并进行了3次预试验，最终不仅发现了该结构的3个规律，还为语料编码的过程提供了有效的新建议，这种严谨的精神给研究者们做了很好的示范。其次，本书也做到了兼顾各个领域的需求，涉及到语篇分析（第二至四章）、应用语言学（第五至七章）和心理语言学（第八至十章）领域，既适合语料库研究者使用，也适合对语料库感兴趣但无从入手的初学者翻阅，尤其对混合型研究者有很高的参考价值。例如在第九章，Gries把语料库搭配研究和心理语言学实验研究很好地结合起来，利用实验法拓展了原先的构式搭配研究，补充了之前研究的缺口，增强了“启动效应”的说服力。再次，本书具有很强的实用性，可以作为三角论证研究的样板。它为我们清晰细致地展现了每一步的研究过程和设计思路，指引研究者如何（how）和何时（when）进行三角论证。值得一提的是，与其他论文集类型的书籍不同，本书不是以研究领域为标准选择论文，而是按研究方法收录论文。重点突出研究者选择了哪些研究方法，为什么选择这些方法，三角论证是如何帮助他们回答研究问题等，对研究方法和步骤有非常清晰的描述。例如在第三章，McEnery、Baker和Dayrell精妙地展示了他们是如何一步步解决研究过程中发现的问题。他们试图从语料库中找出19世纪英国干旱天气的历史证据，但发现官方数据不足以解决问题，于是他们决定从报刊中寻找证据，自建语料库；过程中又发现研究范围太广，于是把研究重心集中在英国本地；为确保质量，又对语料进行人工筛选；质疑媒体报道的真实性，于是查找了所有水文学家对英国干旱天气的记录并参考了地理信息系统，整个研究设计严谨，逻辑性强。最后，本书优势在于一窥语料库语言学的前沿发展，展现了9个语料库研究的最新学术成果，创新性地运用语料库提供了新的研究视角，有助于研究人员从跨学科中受益，减少单一方法的局限性。

本书也有不足之处。第一，本书对三角论证采用了较为广泛的定义。事实上，混合型研究和采用两种或以上研究方法的研究有很多，编者未能明确地提供他们挑选文章的标准，对于初学者和新手研究人员来说难以判断编者选择的文章为什么具有代表性。第二，本书提到三角论证在使用不同的研究方法时可能会导致矛盾的研究结果，但书中的研究并没有体现这一点，也许编者在挑选文章时也有偏颇，对这个尚未解决的问题选择避而不谈。不过，尽管有这些局限，这本书还是

十分值得推荐的。语料库与其他语言学方法之间的三角论证将成为语料库语言学未来发展的关键一步（McEnery & Hardie 2012）。本书的典型性研究为语言学研究者提供了宝贵指导，推动了语料库与其他语言学方法之间的整合趋势，为结合语料库与不同的研究方法提供了新思路。

参考文献

BAKER P, EGBERT J. Triangulating methodological approaches in corpus linguistic research [M]. London: Routledge, 2016.

GRIES S. Ten lectures on quantitative approaches in cognitive linguistics (Version 2) [M]. Amsterdam: Brill, 2016.

MCENERY T, HARDIE A. Corpus linguistics: method, theory and practice [M]. Cambridge: Cambridge University Press, 2012.

通信地址：100875　北京市　北京师范大学外国语言文学学院

English abstracts

The multifactorial turn of statistical methods in learner corpus research and their applications

Learner corpus research is undergoing an expansion in statistical methods, and exhibiting a noticeable shift from monofactorial methods to multifactorial modeling. However, domestic scholars have paid insufficient attention to this trend. This article begins by explicating the drawbacks of monofactorial methods and then dwells on the advantages of building statistical models. Two studies based on the National Matriculation Essay Corpus are then reported. The first study employs multinomial logistic regression to explore the mechanism whereby ten fine-grained syntactic features jointly affect essay scores. The second study utilizes structural equation modeling to examine the complex path relation through which the frequency, association strength, and accuracy of target language bundles produced by essay writers, measured on the basis of COCA, affect essay scores. The authors conclude by arguing that corpus researchers need to enhance their statistical literacy to meet the demands of adequately researching the complexity of language systems and improve the quality of scientific studies.

A multivariate quantitative study on English modal construction from a variationist linguistic perspective

This study applies a corpus-based variationist linguistic perspective to explore the major factors influencing the choice of “must” “have to”, and “have got to” and their diachronic evolution in American English using COHA corpus data. The results show that from 1810 to 2009, when expressing the meaning of “necessary to do sth.”, there was a trend whereby “must” was gradually replaced by “have to”. The main factors influencing the selection of these three constructions ordered by importance are “clause tense” “genre” and “year”. In addition, through further analysis of the evolution of these three constructions, the authors argue that there has been a tendency for central modals to be gradually replaced by semi-modals in American English. Part of the meaning of the modal “must” is increasingly expressed by the semi-modal “have to”, while the use of the semi-modal “have got to” is on the decline.

A study of English accusational expressions in trade-conflict texts: A local function perspective

Unlike the study of the generic function of describing the macro system of language, the study of local functions focuses on the specific function or meaning of language use in restricted texts, which is a developing trend in functional research in the context of corpus linguistics. Using the corpus-driven approach, this study conducts a contrastive analysis of English accusational expressions in China-U.S. anti-dumping texts from 2001 to 2017 with local functional categories and their constituent sequences. The analysis shows that these accusational expressions present regular features and development changes, indicating the differences in the understanding of anti-dumping practices between the discourse communities of China and U.S., which has applications in China's response to anti-dumping trade conflicts. Furthermore, the study also demonstrates the feasibility and usefulness of the local functional perspective in revealing subtle features of the text and provides a reference for text analysis.

A comparable-corpus-based study of phrasal complexity in academic writing in applied linguistics

This study analyzes the use and distribution of 17 complex noun phrases in L1 Chinese graduate students' theses and L1 Chinese and English expert writers' research articles in applied linguistics using a corpus-based approach. The results show that: (i) There is no significant difference in the use of 17 complex noun phrases among the three writer groups (ii) the most frequent complex noun phrases used by the three writer groups are the same; (iii) the use of three complex noun phrases is significantly different among the three writer groups; and (iv) L1 Chinese graduate students' use of five complex noun phrases is significantly different from both L1 Chinese and English expert writers. The use of complex noun phrases is a better indicator of the complexity of academic writing. Pedagogical implications are provided for academic writing and EAP teachers.

Comparison between direct and inverse translations in rendering *zhuyi* terms in Mao Zedong's works: A corpus-based investigation

The frequent occurrence of *zhuyi* terms in Mao Zedong's works reflect Mao's idiosyncratic characteristics in language use and contribute to the construction of his political discourse system. Starting with a classification of *zhuyi* terms in *Mao Zedong Xuanji*, the present paper makes a corpus-based comparison between direct and inverse translations of MaoZedong's works in terms

of their methods for translating *zhuyi* terms to explore the similarities and differences between the two types of translations in dealing with his idiosyncratic language use. The results suggest that neither type of translation displays complete inner-textual regularities, and both show some inter-textual similarities and differences. These findings can be explained from two perspectives: (a) Translated texts reflect a compromise and balance between adequacy and acceptability; (b) MaoZedong's works and their direct translations occupy a high and authoritative status in China and the world at large. This paper further points out that the direct translation of political documents should ensure correctness, promptness, authoritativeness, and maintain Chinese characteristics and standards based on acceptability.

A review on text analysis in the research of finance and economics

Previous studies in finance and economics have been based on the causal inference of econometric models. With the development of big data and computer algorithms, unstructured data represented by text information can be quantified and applied to research in finance and economics. The linguistic features of text information, such as text readability, intonation, and similarity, have gradually become the focus of scholarly research and quantification, hence, text analysis technology has been applied to research in this field. Starting from the language characteristics of text and different research issues in finance and economics, this paper combs the current domestic and foreign literature, analyzes the relationship between text and financial information, and points out direction for future development of text analysis technology to provide a reference for researchers in related fields.

A study of verb valency patterns used by learners: The case of "agree"

Valency pattern, an achievement of corpus phraseology, clearly shows the syntactic and semantic constraints of lexical items. Thus, they can help learners acquire lexical knowledge. Examining the high-frequency verb "agree", this study compares the valency patterns used by Chinese EFL learners and native speakers based on corpus data. The results demonstrate that the frequency distribution is significantly different from that of native speakers. Moreover, learners misuse some valency patterns, and the actants they use lack semantic diversity, which indicates that the learners have not fully grasped the valency patterns of "agree". This study analyzes the reasons for the learners' misuse of valency patterns from the perspective of mother tongue transfer and classroom input, and suggests that the patterns should be applied to vocabulary teaching through hands-off

data-driven learning and a valency pattern dictionary.

A corpus-based study of the discursive construction of China's population policy in American news coverage

Combining the Discourse-Historical Approach (DHA) with corpus techniques, the present study built its own corpus and conducted analysis along three dimensions: themes, discourse strategies, and historical and social context. The findings indicate that the American media employed naming, predication, perspectivization, and argumentation strategies in order to construct two opposing parties: positive "selfness" and negative "otherness." The American media's attitudes toward China's population policy underwent a dynamic diachronic change from total denial and criticism, by-stand and suspicion, to sarcasm and reluctant recognition. The present study sheds light on this frontier research methodology and theoretical framework for political discourse studies.