

《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

二〇二二

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

语
料
库
语
言
学

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

外
研
社

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



corpus.bfsu.edu.cn

语料库语言学

CORPUS LINGUISTICS

要 目

- | | |
|-------------------------|-------------|
| 学习者语料库研究的多因素统计方法转向 | 李元科 何安平 黄灵敏 |
| 变异语言学视角下英语情态构式多元定量研究 | 李思雨 戴雅宁 孟庆楠 |
| 贸易冲突话语中英语指责表述的局部功能研究 | 刘运锋 |
| 基于语料库的“主义”词译出译入对比研究 | 石欣玉 黄立波 |
| 国内外财经文本分析研究综述 | 牛华勇 窦一轩 夏晓雪 |
| DiSCUSS 现代汉语平衡口语语料库的创建 | 孙铭辰 |
| deGLOBE 当代德语书面语平衡语料库的创建 | 周顾盈 等 |

外研社·期刊出版分社

电话: 010-88819267
E-mail: qkzx@fltrp.com
网址: www.bfsujournals.com



记载人类文明
沟通世界文化
www.fltrp.com



北外学术期刊



iResearch 微信公众号

责任编辑: 赵 雪
责任校对: 孙凤兰
封面设计: 锋尚设计



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

编审委员会（按姓氏音序）
主 任：
梁茂成（北京航空航天大学）

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

Editorial Board (in alphabetical order)
Chair:
Liang Maocheng (Beihang University)

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

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

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

*本刊获北京外国语大学“双一流”建设经费资助

版权声明

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

语料库语言学

CORPUS LINGUISTICS

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

学习者语料库研究的多因素统计方法转向.....	李元科 何安平 黄灵敏 (1)
变异语言学视角下英语情态构式多元定量研究.....	李思雨 戴雅宁 孟庆楠 (14)
贸易冲突话语中英语指责表述的局部功能研究.....	刘运锋 (25)

研究论文

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.....	高 霞 (47)
基于语料库的“主义”词译出译入对比研究.....	石欣玉 黄立波 (61)
国内外财经文本分析研究综述.....	牛华勇 窦一轩 夏晓雪 (81)
学习者英语动词配价型式使用特征研究——以 agree 为例.....	孙海燕 牛文爽 (96)
基于语料库的美国媒体中国人口话语建构研究.....	王 琴 (109)

研制开发

DiSCUSS 现代汉语平衡口语语料库的创建.....	孙铭辰 (127)
deGLOBE 当代德语书面语平衡语料库的创建.....	周顾盈 等 (136)
MgmtDEAP 管理科学与工程学术英语语料库的创建.....	邓静子 等 (145)

书刊评介

《对比语言学研究新路径：实证与方法论的挑战》述评.....	葛恬馨 (157)
《通过语料库方法对语言分析进行三角论证》述评.....	梁悦怡 王德亮 (161)
英文摘要.....	(166)

CORPUS LINGUISTICS

Volume 9, Number 2, 2022

Table of Contents

Featured column: Contextual co-selection approach to language

- The multifactorial turn of statistical methods in learner corpus research and their applications *LI Yuanke, HE Anping & HUANG Lingmin* (1)
- A multivariate quantitative study on English modal construction from a variationist linguistic perspective..... *LI Siyu, DAI Yaning & MENG Qingnan* (14)
- A study of English accusational expressions in trade-conflict texts: A local function perspective..... *LIU Yunfeng* (25)

Research articles

- A comparable-corpus-based study of phrasal complexity in academic writing in applied linguistics *GAO Xia* (47)
- Comparison between direct and inverse translations in rendering *zhuyi* terms in Mao Zedong's works: A corpus-based investigation..... *SHI Xinyu & HUANG Libo* (61)
- A review on text analysis in the research of finance and economics *NIU Huayong, DOU Yixuan & XIA Xiaoxue* (81)
- A study of verb valency patterns used by learners: The case of "agree" *SUN Haiyan & NIU Wenshuang* (96)
- A corpus-based study of the discursive construction of China's population policy in American news coverage *WANG Qin* (109)

New corpora, tools and methods

- The construction of DiSCUSS Corpus *SUN Mingchen* (127)
- The construction of deGLOBE contemporary written German Corpus *ZHOU Guying et al.* (136)
- The construction of MgmtDEAP Corpus *DENG Jingzi et al.* (145)

Book reviews

- R. Enghels et al. (eds.). *New Approaches to Contrastive Linguistics: Empirical and Methodological Challenges* (2020)..... *GE Tianxin* (157)
- J. Egbert & P. Baker (eds.) *Using Corpus Methods to Triangulate Linguistic Analysis* (2020) *LIANG Yueyi & WANG Deliang* (161)

- English abstracts..... (166)

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穷。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1¹。

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
		D	宗教	17
	通用	E	日常技艺及消遣爱好	36
		F	通俗读物	48
		G	传记、回忆录等	75
		H	政府或机构公文及文宣	30
	学术	J	学术	80
		K	一般小说	50
		L	侦探小说	12
		M	科幻小说	12
	小说	N	历险悬疑小说	13
		P	言情小说	30
虚构类 (126篇)		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。

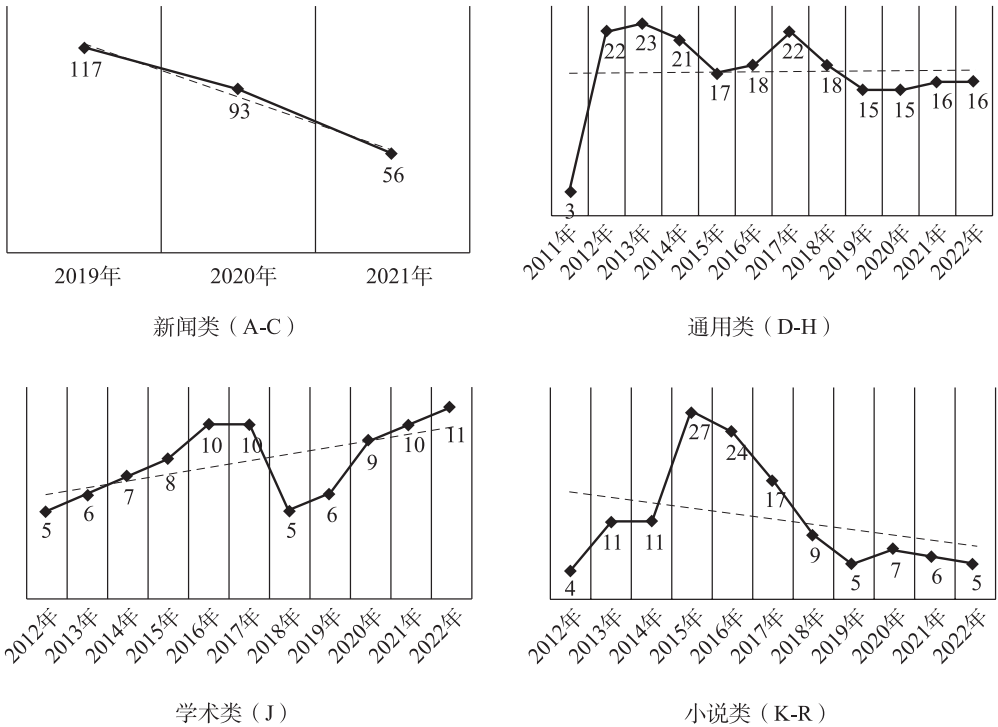


图1 四类体裁所收语料时间数量分布

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-zÄÖöÜß0-9-]+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)