OSID

2009—2019年中国语料库研究的可视化分析

李 慧
(重庆第二师范学院 外国语言文学学院,重庆 400067)

摘要:从中国知网上筛选出2311篇主题为“语料库”的核心期刊论文,通过CiteSpace可视化软件,从研究作者、研究机构、高频关键词、聚类分析、研究前沿等方面对其进行了分析。国内语料库研究成果较为丰富,但是仍然有一些不足之处,如缺少对语料库软件开发的研究、研究前沿仍然多是单模态语料库应用研究、对于语料库的研究主要以英语语言为主。

关键词:语料库;可视化;中国
中图分类号:H313 文献标识码:A 文章编号:1008-6390(2021)03-0056-05

语料库是遵循特定的语言学原则、结合随机抽样的统计方法、将自然出现的连续语言运用文本及话语片段采集收纳,进而构建出具备一定容量的大型电子文库[1]。此电子文库中所收纳的都是实际应用过程中典型的真实语言材料,能反映语言现象的特点,具有准确、客观的特性。1967年,美国布朗语料库率先建立,随后一系列相关文章陆续发表,标志着语料库的开端。以语料库为基础的语言学研究在欧洲和美国快速发展,并逐渐在世界范围内得到重视。在中国,自20世纪80年代以来,随着计算机科学技术的蓬勃发展,语料库在语言学研究领域中的地位日益显著。30多年来,语料库在中国的发展经历了四个阶段:1. 语料库1.0,经手工采集的语料集合;2. 语料库2.0,初步经过计算机处理而构建的语料库;3. 语料库3.0,随着计算机处理能力的提升,研究者能够依靠其进行大规模数据采集与加工,形成了计算机化的大型语料库,语料性质也从文本扩展到音频等,此阶段语料库被广泛应用于语言学、人文社科等多个领域,并上升到语料库语言学的理论地位;4. 语料库4.0,随着现代计算机多媒体技术的发展,以及人们对语言活动本质认识的提升,多模态语料库应运而生。

一、研究样本与研究方法

本研究的数据来自中国知网(CNKI),在CNKI检索栏中,检索时间段选定为2009年1月1日至2019年12月31日,检索主题选定为“语料库”,专辑导航为“哲学与人文科学”‘社会科学Ⅰ辑’‘社会科学Ⅱ辑”,数据类型选定为“核心期刊”,在检索结果中剔除与研究不相关的项目,最终通过筛选得到2311篇论文;然后将这些文章的题录信息下载,以备载入CiteSpace软件中进行分析。

本文所用的数据分析工具是CiteSpace 5.6.R4。CiteSpace也叫引文空间,是在科学计量学、数据可视化背景下逐渐发展起来的一款引文可视化分析软件,由美国德雷塞尔大学信息科学与技术学院的华人学者陈超美博士于2004年开发,2007年首次被引入国内。由于是通过可视化的手段来呈现科学知识的结构、规律和分布情况,因此也将通过此类方法分析得到的可视化图形称为“科学知识图谱”。CiteSpace软件可用来构建学科知识图谱,分析学科热点和揭示前沿趋势。

本研究主要通过CiteSpace软件对研究数据的施引文献的合作图谱和贡献图谱进行可视化分析,包括对研究作者的可视化分析、研究机构的可视化

收稿日期:2020-10-30
基金项目:重庆市教育科学“十三五”规划一般课题“多模态语料库在商务英语听说教学中的应用及其效果研究”(2019-GX-399);重庆第二师范学院教改项目“基于多模态语料库的商务英语听说课程教学改革与探索”(JG201932)
作者简介:李慧,讲师,研究方向:商务英语话语、语料库语言学。

分析、高频关键词的可视化分析、聚类的可视化分析、中国语料库研究的前沿分析。

二、研究结果与分析

(一) 研究者的可视化分析

通过 CiteSpace 软件对发文作者的分析得出，2311 篇文章涉及 408 位作者。根据洛特卡定律，发文数量为 1 篇的作者占作者总数的比例低于 60% 时，会形成核心作者群，本研究中发文数量为 1 篇的作者为 120 人，大约占作者总数的 29.41%，说明中国语料库研究领域已经形成了核心作者群。经统计，发文数量达到 9 篇及以上的高产发文作者有 13 人，见图 1。这些研究者可谓中国语料库研究的领军人物，其中胡开宝和卫乃兴的发文量超过 30 篇，可见他们为中国语料库研究做出了突出的贡献。同时通过观察研究者的可视化图谱，发现胡开宝、王克非、卫乃兴、梁茂成等学者间的节点连线很密，说明这些作者间的合作紧密，可见中国语料库研究形成了以这些学者为代表的作者合作研究群。

Visible	Count	Centrality	Year	Auth...
✔	35	0.01	2009	胡开宝
✔	33	0.00	2009	卫乃兴
✔	23	0.01	2009	王克非
✔	19	0.00	2010	刘永兵
✔	18	0.00	2009	刘泽权
✔	18	0.00	2011	戈玲玲
✔	15	0.00	2009	何安平
✔	13	0.00	2009	秦洪武
✔	13	0.00	2010	肖忠华
✔	10	0.00	2013	许家金
✔	10	0.00	2012	梁茂成
✔	10	0.00	2011	黄立波
✔	9	0.01	2014	庞双子

图 1　发文作者和中介中心度

(二) 研究机构的可视化分析

通过对研究机构的可视化分析发现，以北京外国语大学中国外语研究中心为中心的连线较密，说明其研究团队发展较为成熟，它与北京航空航天大学外国语学院、上海交通大学、广东外语外贸大学、浙江大学外国语言文化与国际交流学院等机构有比较紧密的合作。通过观察这些研究机构的发文数据，得到语料库研究排在前十位的研究机构，见图 2。这些大学发文数量都在 19 篇以上，其中上海交通大学外国语学院发文数量多达 69 篇，排名第一。从中介中心性排序来看，北京外国语大学中国外语研究中心、北京航空航天大学外国语学院的中介中心度值最大，均为 0.07，说明这两个机构与其他机构合作广泛。

Visible	Count	Centrality	Year	Institutions
✔	69	0.06	2009	上海交通大学外国语学院
✔	59	0.05	2009	上海交通大学
✔	42	0.05	2010	北京外国语大学
✔	41	0.07	2009	北京外国语大学中国外语教育研究中心
✔	34	0.07	2010	北京航空航天大学外国语学院
✔	32	0.03	2009	上海外国语大学
✔	28	0.01	2009	广东外语外贸大学
✔	23	0.02	2009	浙江大学外国语言文化与国际交流学院
✔	22	0.00	2010	东北师范大学外国语学院
✔	19	0.00	2010	东北师范大学

图 2　研究机构和中介中心度

(三) 高频关键词分析

本研究以 1 年为一个时间区，将 2009—2019 年分为 11 个分区，选取每个时间分区内 50 个引用频率最高或者出现次数最多的关键词，通过提取关键词出现的具体频率和关键词的中介中心度得到图 3，其中关键词的高频出现率表明关键词具有较高的关注度，是研究的热点。中介中心度反映了关键词在整体网络中作为媒介的能力，即占据其他两个节点之间最短路径的能力。关键词的中介中心度值越高，那么它控制的关键词之间的信息流越多。由图 3 可见，语料库、语库、翻译、语义韵、平行语料库、语料库语言学在知识图谱中共现的频率和中介中心度的排名均较靠前，说明这几类不仅是人们的研究热点，而且其研究内容与其他研究的关联度较高。

File　Data　Visualization　Display　Nodes　Links　Labels　C

Visible	Count	Centrality	Year	Keywords
✔	694	0.77	2009	语料库
✔	79	0.13	2009	语库
✔	56	0.09	2010	翻译
✔	53	0.16	2009	语义韵
✔	48	0.07	2009	平行语料库
✔	46	0.08	2009	语料库语言学
✔	39	0.05	2009	语言学
✔	30	0.05	2012	语料库翻译学
✔	27	0.06	2010	搭配
✔	26	0.02	2009	语言科学
✔	24	0.06	2011	隐喻
✔	23	0.03	2010	翻译教学
✔	22	0.05	2009	研究方法
✔	22	0.02	2011	译者风格
✔	21	0.06	2009	基于语料库
✔	20	0.04	2010	二语习得
✔	19	0.02	2010	词块
✔	19	0.03	2010	语块
✔	18	0.07	2009	英语写作
✔	18	0.04	2010	认知语言学

图 3　高频关键词和中介中心度

(四) 聚类分析

聚类分析指将抽象对象的集合分组为由类似对

象组成的多个类别的过程。在 CiteSpace 中启动聚类分析功能，得到本次聚类的 Modularity Q（聚类模块值）0.5521，以及 Mean Silhouette（聚类平均轮廓值）0.3682，均大于 0.3，所以本次聚类基本合理。通过聚类分析，得到中国语料库研究的聚类分析图谱，见图 4。由图 4 可见，中国语料库研究主要包括 11 个类别，即对外汉语教学、语言科学、平行语料库、语义韵、学习者、汉语中介语语料库、自然语言处理、隐喻、共选、中国英语学习者、战略转移。

系统自动列出了排名前五位的聚类所对应的具体信息，见图 5。其中，“Silhouette”是聚类平均轮廓值，指一个聚类中关键词之间的同质性，这五个聚类序号中聚类平均轮廓均大于0.4，所以这些聚类是合理的。下面对排在前五位的聚类标识词进行详细分析。

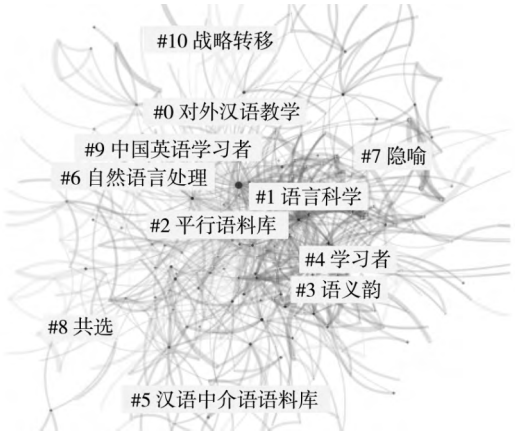

图 4　中国语料库研究的聚类分析图谱

Cluster ID	mean(Year)	Size	Silhouette	Top Terms (log-likelihood ratio, p-level)
0	2014	54	0.732	对外汉语教学 (13.24, 0.001); 语料库 (13.17, 0.001); 翻译汉语 (9.92, 0.005); ...
1	2011	52	0.842	语言科学 (30.71, 1.0E-4); 语言学 (26.83, 1.0E-4); 语库 (21.24, 1.0E-4); 文法 (...
2	2013	46	0.833	平行语料库 (16.38, 1.0E-4); 语料库翻译学 (15.38, 1.0E-4); 研究热点 (12.27, ...
3	2012	43	0.777	语义韵 (18.06, 1.0E-4); 共选理论 (10.38, 0.005); 概念迁移 (10.38, 0.005); 语...
4	2010	37	0.857	学习者 (24.25, 1.0E-4); 英语语料库 (17.07, 1.0E-4); 外语教学 (12.32, 0.001); ...

图 5　系统自动聚类的前 5 个结果

1. 对外汉语教学

近年来语料库在对外汉语教学中的应用越来越受欢迎，主要包括基于语料库的对外汉语语音、词汇、写作等领域的创新与实践。胡韧奋、朱琦、杨丽姣[2]进行了对外汉语教学领域话题语料库的研究与构建，设计了一个包含 4 个一级话题、23 个二级话题和 246 个三级话题的三层话题框架，构建了一个规模约 12 万句的对外汉语教学话题语料库。卫斓[3]研究了留学生使用汉语双音形容词与名词搭配的情况，他把收集的留学生的作文和作业，与汉语母语语料库进行了对比，分析了留学生在使用双音形容词与名词搭配时出现的各种问题、原因及其对策。杨泉[4]基于 HSK 作文语料库对留学生离合词偏误计算机自动纠错系统进行了研究，通过收集该语料库中的留学生离合词偏误句，分析了留学生易出现的问题类型和原因，同时根据每种偏误编写了纠错规则，此外还建立了计算机自动纠正留学生离合词偏误系统。

语料库与对外汉语教学已有紧密的联系，对外汉语语料库的建设为汉语教学领域的教师、研究者及教材编写者提供了较全面的话题信息参考，同时实现了对外汉语教学方法的革新。

2. 语言科学

语料库语言学是 20 世纪出现的一门语言研究科学，是一个独立的学科，有自己独到的理论体系和操作方法。语料库语言学立足于大量真实的语言数据，对语料库进行系统而详尽的观察和概括所得到的结论对语言理论建设具有无可比拟的创新意义。梁茂成[5]、李文中[6]和许家金[7]梳理了语料库语言学的学科基础、方法论等。刘明、常晨光[8]探究了语料库辅助话语研究的缘起、特征及应用，发现尽管语料库辅助话语研究目前还处于成长和发展阶段，但其为语料库语言学和话语分析的进一步交叉融合指明了前进的方向。董永义[9]研究了语料库语言学和语言测试的关系，基于语料库语言学的视角，分析了语言测试的概念并描述了语料库语言学与语言测试之间的关系，目的在于加深对语言学、语料库等概念的理解。语料库语言学拓展了语言研究的范围和广度，使语言研究的可靠性有了很大程度的提升。

3. 平行语料库

双语平行语料通过大规模平行文本的呈现革新了翻译教学方法，可以为翻译教学提供真实的翻译样例，是翻译教学的重要辅助之一。王克非、秦洪武[10]分析了平行语料库在翻译教学中的应用，认为其有助于改善译文评估方式，可以提升翻译学习效率和效果，同时有利于创建高质量的自主学习和发现式翻译教学环境。柴明颎[11]建立了翻译专业教学语料库以探讨技术时代的翻译教学改革，分析了语料库的建库原则和方法、语料的标注标记及翻译

专业教学语料库的特点等。戈玲玲、李广伟、王一鸣[12]基于本源概念建立了本源概念双语平行语料库，创建了 Web 教学平台系统，同时提出了“1 +2 + 1”翻译教学模式。平行语料库已经是翻译教学的重要方法之一，为教学方法带来了新动力，对学生学习的效果和方法产生了积极作用，未来将对翻译教学改革带来持续影响。

4. 语义韵

Louw[13]于 1993 年正式提出语义韵这一概念：“一个语言形式会被其周围搭配词稳定的语义氛围所沾染，这种现象就被称为语义韵。”他将语义韵分为“好的”和“坏的”两大类。在他提出后的二十多年里，语料库语言学家对语义韵进行了一系列丰富的研究。邵斌、王文斌[14]考察了语料库中英语词缀的语义韵，以英语析取词缀“Mc-”为个案，揭示词缀语义韵从无到有的过程，并通过概念隐喻和转喻理论探索其背后的认知机制，发现语义韵这种语义传染不仅发生在词与词之间，也可由词传染给作为词内成分的词缀，使之具有类似的语义韵。李美霞、焦瑗珲[15]基于语料库对英语逻辑结果程式语的语义韵进行了研究，通过大量数据统计分析，发现符合研究条件的 13 个英语逻辑结果程式语按其呈现的语义韵可分为三类 “as a result of、caused by、lead to”大多数情况下表达消极语义韵的趋向显著 “so that、now that、as a result、result from、thanks to”表示中性语义韵的趋向明显 “so … that、bring about、because of、result in、due to”表现糅杂型语义韵的趋向突显。陆军[16]基于语料库对学习者英语近义词搭配行为与语义韵进行了研究，发现英语近义词在搭配行为和语义韵特征上存在差异，学习者很难区别这些特征；不同类型的学习者在近义词的使用特征和发展模式上有明显差异，具体表现为不同程度的近义词替代和语义韵冲突。

5. 学习者

自语料库引入中国以来，中国学者建立了很多学习者语料库，如中国大学生写作语料库、高中生词汇语料库、大学生口语语料库等，同时也对学习者语料库开展了很多研究。胡元江、石海漫、季萍[17]基于语料库对英语学习者与本族语者议论文词块的结构与功能特征进行了对比研究，发现英语专业从大一到大三的学生使用的词块量均比本族语者要多；学习者更多地依赖非目标词块表达各语用功能；本族语者与学习者产出的词块以动词类为主，随后是名词类和介词类，且均主要表达组篇和指示功能。黄开胜、周新平[18]基于语料库对中国英语学习者词块输出能力的趋势进行了分析，结果显示中国英语专业学习者大学阶段词块输出数量有明显增长，但是质量却没有明显提高。吴让科[19]基于中国学习者的英语作文语料库研究了代词作为嵌入词对中国英语学习者关系分句习得的影响，研究结果显示：高水平学习者比低水平学习者使用关系分句的频率高；关系分句中嵌入词为名词或代词时，两种学习者主语和宾语关系分句的使用频数有差异；二者使用各类代词为嵌入词时，其分布频率与母语者不一致。

（五）中国语料库研究的前沿分析

借助 CiteSpace 的突发词检测算法，从文献题目、摘要等提取出突变术语，可用于检测某一学科研究兴趣的突然增长，即可了解该学科的研究前沿。图 6 显示了语料库语言学研究的突变术语。由图 6 可知，外语教学、语库、语言科学、英语写作、学习者、词汇、话语、认知语言学、可比语料库、英文摘要、英译等是语料库语言学研究的前沿。从突变年份来看，外语教学、对外汉语教学、语言科学、英语写作、学习者、英语语料库最早发生突变。从突变结束年份来看，可比语料库、CiteSpace、词块、研究热点、语步、局部语法、英文摘要、英译的突变时间最接近现在。综合突变的起始年份和结束年份，英语写作、词汇突变持续的年份最久，长达 4 年。语库和学习者的突变强度排名分别为第一和第二，其中语库专指某一个语料库，说明 2009—2010 年基于自建语料库的研究激增，同时基于语料库对学习者的研究数量也迅速增长。由此可见，国内语料库语言学的研究热点为学习者、英译、语言科学、可比语料库、语步、标注、认知语言学等。

TOP 22 Keywords with the Strongest Citation Bursts

Keywords	Year	Strength	Begin	End	2009—2019
外语教学	2009	3.1947	2009	2010	
语库	2009	7.7903	2009	2011	
对外汉语教学	2009	2.8720	2009	2010	
语言科学	2009	3.6345	2009	2012	
英语写作	2009	2.6317	2009	2013	
学习者	2009	6.1645	2009	2011	
英语语料库	2009	2.8720	2009	2010	
认知语言学	2009	3.2499	2011	2012	
方法论	2009	2.9310	2011	2012	
标注	2009	3.2618	2012	2015	
词汇	2009	2.7024	2013	2017	
话语	2009	2.5573	2014	2015	
学术写作	2009	2.4852	2015	2016	
语境	2009	2.4852	2015	2016	
可比语料库	2009	3.4889	2016	2019	
CiteSpace	2009	3.1385	2016	2019	
词块	2009	3.1442	2016	2017	
研究热点	2009	2.4964	2016	2019	
语步	2009	3.5543	2017	2019	
局部语法	2009	3.0703	2017	2019	
英文摘要	2009	3.0703	2017	2019	
英译	2009	4.7902	2017	2019	

图 6　语料库语言学研究的突变术语（节选）

三、结语

语料库在中国的发展历经三十多年，已经取得了丰硕的成果。本文应用 CiteSpace 可视化软件对 CNKI 上 2311 篇主题为“语料库”的核心期刊论文进行了研究，主要对施引文献的合作图谱和贡献图谱进行了可视化分析。研究发现：首先，对于研究作者而言，中国语料库研究领域的领军人物为胡开宝、卫乃兴、王克非等，他们是高产研究作者，同时中国语料库研究形成了以这些学者为代表的作者合作研究群。其次，对于研究机构而言，语料库研究的主要研究机构分别为上海交通大学外国语学院、上海交通大学、北京外国语大学等，其中北京外国语大学中国外语研究中心、北京航空航天大学外国语学院还与其他机构合作广泛。第三，对于关键词而言，中国语料库研究的核心关键词为“语料库、语库、翻译”等，这些关键词之间的关联度比较高。第四，对于聚类而言，中国语料库研究主要包括 11 个类别，如对外汉语教学、语言科学、平行语料库等。第五，外语教学、英语写作、词汇等是中国语料库的研究前沿。

虽然国内语料库研究成果较为丰富，但是仍然有一些不足之处。国内对于语料库的研究主要侧重于基于语料库的应用研究，如翻译、语义韵、平行语料库等，缺少对语料库软件开发的研究。国内语料库研究起步晚，相对比较滞后，研究前沿仍然是比较基础的基于文字的单模态语料库应用研究，如英语教学、英语写作、词汇、话语等，而国外已经开始了多模态研究，即集文字、图片、声音、动画为一体的研究，今后国内也要加大多模态语料库相关研究的力度。再者，国内对于语料库的研究主要以英语语言为主，对小语种语料库的研究较少，如法语、意大利语、葡萄牙语等，未来可把语料库研究更多地推广到小语种的相关研究。

参考文献：

[1] Halliday M A K. An introduction to functional grammar [M]. 2nd. ed. London: Arnold, 1994.

[2] 胡韧奋，朱琦，杨丽姣. 对外汉语教学领域话题语料库的研究与构建 [J]. 中文信息学报，2015(6)：62-68.

[3] 卫斓. 留学生使用汉语双音形容词与名词搭配情况考察 [J]. 首都师范大学学报(社会科学版)，2014(S1)：74-81.

[4] 杨泉. 基于 HSK 作文语料库的留学生离合词偏误计算机自动纠错系统初探 [J]. 语言文字应用，2011(2)：116-124.

[5] 梁茂成. 语料库语言学研究的两种范式：渊源、分歧及前景 [J]. 外语教学与研究，2012(3)：323-335.

[6] 李文中. 语料库语言学的研究视野 [J]. 解放军外国语学院学报，2010(2)：37-40.

[7] 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.

[8] 刘明，常晨光. 语料库辅助话语研究的缘起、特征及应用 [J]. 福建师范大学学报(哲学社会科学版)，2018(1)：90-96.

[9] 董永义. 语料库语言学和语言测试的关系研究 [J]. 语文建设，2016(14)：87-88.

[10] 王克非，秦洪武. 论平行语料库在翻译教学中的应用 [J]. 外语教学与研究，2015(5)：763-772.

[11] 柴明颎，王静. 技术时代的翻译教学改革：翻译专业教学语料库的建库探索 [J]. 外语电教化，2017(6)：25-31.

[12] 戈玲玲，李广伟，王一鸣. 基于本源概念双语平行语料库的翻译教学平台建设及其教学模式研究 [J]. 外语界，2015(4)：11-17.

[13] Louw B. Irony in the text or insincerity in the writer: the diagnostic potential of semantic prosodies [M] // Baker M, Francis G, Tognini-Bonelli E (eds.). Text and Technology: in Honour of John Sinclair. Amsterdam: John Benjamins, 1993.

[14] 邵斌，王文斌. 基于语料库的英语词缀语义韵考察 [J]. 外语教学研究，2015(4)：8-12.

[15] 李美霞，焦瑗珲. 基于语料库的英语逻辑结果程式语语义韵研究 [J]. 外语教学，2013(2)：21-26.

[16] 陆军. 基于语料库的学习者英语近义词搭配行为与语义韵研究 [J]. 现代外语，2010(3)：276-286.

[17] 胡元江，石海漫，季萍. 英语学习者与本族语者议论文词块的结构与功能特征：基于语料库的对比研究 [J]. 外语研究，2017(4)：58-62.

[18] 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.

[19] 吴让科. 代词作为嵌入词对中国英语学习者关系分句习得的影响：一项基于语料库的行为研究 [J]. 外语研究，2016(3)：58-63.

责任编辑　亦　筱