Modern Linguistics 现代语言学, 2024, 12(2), 1152-1157
Published Online February 2024 in Hans. https://www.hanspub.org/journal/ml
https://doi.org/10.12677/ml.2024.122154

Hans汉斯

基于CiteSpace的国内语料库语言学可视化分析

贺玲玲

郑州商学院外国语学院，河南 郑州

收稿日期：2023年12月26日；录用日期：2024年2月22日；发布日期：2024年2月29日

摘　要

语料库语言学作为一门融合了社会学、语言学，心理学、人类学等科目的学科，于20世纪70年代末传入我国，对国内社会的发展产生了积极影响。本研究借助CiteSpace计量学软件，以中国学术网络出版总库(CNKI)在2017~2023年间所有期刊类语料库语言学论文为研究对象，通过作者和机构共现网络知识图谱、关键词共现网络知识图谱和关键词突现图、年度文献分布维度分析我国语料库语言学作者间与机构间的学术现状、学界的研究热点，并预测未来研究趋势与前沿。研究发现：2015年为语料库语言学研究的峰值年，研究热点领域为“中国形象”“话语分析”“对比分析”“语义韵”；近些年的研究主要集中在话语分析和二语习得；语料库研究出现新的发展变化，跨学科特征呈现上升趋势。

关键词

语料库语言学，研究趋势与前言，CiteSpace

A Visualized Analysis of Knowledge Map of Corpus Linguistics Research Based on CiteSpace

Lingling He

Foreign Language Department, Zhengzhou Business University, Zhengzhou Henan

Received: Dec. 26th, 2023; accepted: Feb. 22nd, 2024; published: Feb. 29th, 2024

Abstract

Corpus linguistics, as a discipline that integrates sociology, linguistics, psychology, anthropology, and other subjects, was introduced to China in the late 1970s and had a positive impact on the development of domestic society. This study used CiteSpace econometric software to analyze the academic status and research hotspots of Chinese corpus linguistics among authors and institu-

文章引用: 贺玲玲. 基于 CiteSpace 的国内语料库语言学可视化分析[J]. 现代语言学, 2024, 12(2): 1152-1157.
DOI: 10.12677/ml.2024.122154

tions, as well as among institutions, through the co-occurrence network knowledge graph, keyword co-occurrence network knowledge graph, and keyword emergence graph of all journal corpus linguistic papers in CNKI from 2017 to 2023. The study also predicted future research trends and frontiers. Research has found that the hot areas of corpus research in 2015 were "Chinese image", "discourse analysis", "comparative analysis", and "semantic rhyme". In recent years, research has mainly focused on discourse analysis and second language acquisition. Corpus research has seen new developments and changes, with interdisciplinary characteristics showing an upward trend.

Keywords

Corpus Linguistics, Research Trends and Frontiers, CiteSpace

1. 引言

20 世纪 90 年代是语料库语言学的黄金岁月。基于语料库或语料库驱动的研究范式被广泛认可，并迅速扩展，进入几乎所有语言学领域[1]。语料库语言学是在文本语料的基础上进行语言研究的一门学科[2]。语料库语言学在西方有两个源流，一个是 20 世纪 60 年代初语言学研究中英国的实证主义思潮和实践以及美国的结构主义语言学传统，另一个是自然语言处理研究，尤其是机器翻译领域中的语料库开发。从现有的文献来看，语料库语言学研究在大西洋两岸几乎同时开始[3]，而 Brown 语料库则是第一个建成后免费提供给语言学者共享的语料库。“语料库语言学”这一名称直到上世纪 70 年代末 80 年代初才真正得到使用，并赢得语言学界的尊重[4]。而国内对语料库的探索开始于 1981 年杨惠中教授的对科技英语教学问题的探究。语料库语言学作为一门新兴的研究领域，通过大量客观真实的数据对某一语言现象进行研究。其优势在于能够提供一系列统计数据，因此研究结果也更客观，更有说服力。基于语料库的计算机软件统计和相关的数据分析可避免传统翻译研究中感悟式或体验式译本方法所带来的主观或个人偏见。

关于语料库的研究涉及多个方面，作为一种研究方法，可以与其他研究相结合。金碧希和卫乃兴[5]发表“话语研究的语料库路径：方法、挑战与前景”，其系统分析了话语研究的三条语料库研究路径：基于语料库的话语研究、语料库辅助的话语研究和语料库咨询的话语研究，对特征进行对比分析，并提出语料库研究跨学科发展是必由之径。赵冲和许家金[6]发表近百年西班牙语语料库建设与研究概述，梳理了近百年西班牙语语料库发展史，为我国语料库建设与研究提供一定的借鉴意义。姜峰和 Ken Kyland [7]发表“互动元话语：学术语境变迁中的论辩与修辞”，通过自建四个文理学科期刊论文的语料库，对互动类元话语进行系统分析，发现学术论辩与话语实践在一定情况下随着社会文化等语境的变迁而产生变化。

为了更客观地了解国内语料库语言学近几年的研究成果，本文通过 CiteSpace 对 CNKI 上面近七年的文献进行系统的总结和分析，梳理近七年语料库语言学研究的发文数量、关键词、发文作者和机构等，总结国内语料库语言学研究的发展脉络及现状，以期为今后语料库语言学的发展提供一定借鉴和参考。

2. 研究设计

2.1. 研究问题

本文以中国学术网络出版总库(CNKI)在 2017~2023 年间所有期刊类语料库语言学论文为研究对象，

拟回答以下三个问题：1) 当下语料库语言学研究的热点是什么？2) 语料库语言学研究的新趋势涉及哪些层面？3) 语料库语言学研究领域有哪些新的发展变化？

2.2. 数据来源

本文以中国学术网络出版总库(CNKI)在 2013~2023 年间所有期刊类语料库语言学论文为研究对象，以“语料库语言学”为主题进行检索，时间范围为 2013~2023 年，共计 776 篇文献。

2.3. 研究工具和方法

CiteSpace 是一款可视化分析软件，由美国德雷克塞尔大学开发，用于某一领域研究文献数据的分析，可以借助该软件对某一时期的研究热点、关键词、研究趋势等有一个清晰地了解。作为一款用 Java 语言编写而成的文献计量可视化软件，CiteSpace 集合作网络分析、共现分析、共被引分析、聚类分析、耦合分析等诸多功能于一身，能够根据研究者输入的纯文本格式文献资料和参数设置，自动生成某一指定领域的科学知识图谱，在文献可视化分析方面具有独特优势[7]。本文运用 CiteSpace 6.2.R6 作为研究工具。

本文将从知网导出的 776 篇文献以 Refworks 的形式导出后导入软件 CiteSpace 中，时间区间设置为 2017 至 2023 年，时间分区为 1 年。主题来源(Term Source)有作者(Author)、机构(Institution)、国家(Country)、关键词(Key word)、参考文献(Reference)等，在操作过程中主要选取关键词、作者和机构，之后生成关键词共现图、关键词突现图、作者和机构知识图谱。年度文献分布图是由 CiteSpace 导出的数据复制在 excel 表格中制作而成。通过数据对语料库研究的热点话题进行系统的分析。

3. 数据分析

3.1. 年度文献分布

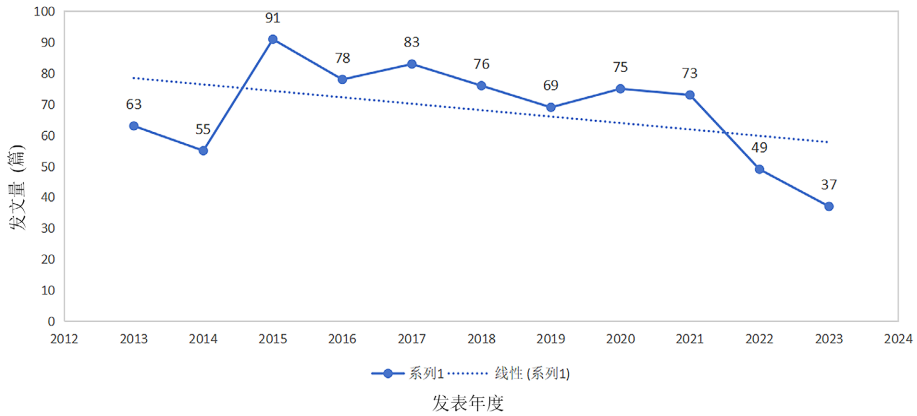

Figure 1. Annual publication volume of corpus study
图 1. 语料库研究年度发文量

年度发文量体现了某一研究领域研究热度的分布情况，图 1 为 2013~2023 年知网上关于语料库语言学论文的发表情况，从图中可以发现，语料库语言学的研究每年论文的发表数量较为持平；关于语料库语言学的研究在 2015 年为研究热度峰值，高达 91 篇。到目前为止，该数据截止到 2023 年 11 月的文献。由于文献数量不全可能导致 2023 年数量的减少。

3.2. 关键词共现

关键词是文章核心内容的浓缩及提炼。如果某一关键词在其所在领域的文献中反复出现，则该关键

词所表征的研究课题是该领域的研究热点[8]。本文对 776 篇文献的关键词进行了共现分析，图 2 为关键词共现情况。图 2 中每一个圆圈代表一个关键词，若圆圈越大，则表示该关键词出现的频次越高。每个圆圈之间的连线表示这些关键词在同一篇文献中存在共现关系。从该图可以发现，出现频率最高的包括话语分析、二语习得、外语教学、对比分析、中国形象、体裁分析等。其中，话语分析、二语习得、外语教学作为关键词出现频率较高，说明语料库研究的热点主要为话语分析和二语习得。张毓和卫乃兴在语料库语言学期刊上发表“基于 LDA 主题建模技术的北京冬奥会话语意义研究”，以境外英文媒体关于北京冬奥会的报道为语料来源，自建北京冬奥会英文报道语料库，采用 LDA 主题建模技术探究背景冬奥会的境外英文报道主题。该研究为 LDA 模型、语料库语言学方法和批评话语分析方法结合的研究提供了新的分析框架，为未来语料库语言学的研究提供了新的方向。孙丰果[9]在 2023 年发表博士论文“国家话语视域下中美国家身份建构对比研究——以两国在安理会涉朝核问题发言为例”，对中美两国外交官在安理会涉朝核问题的发言，考察两国国家身份的话语建构特征和差异。张红和李凯玥在外语研究与教学上发表“基于语料库的‘铸牢中华民族共同体意识’英译对外传播与接受研究”，依据批评话语分析理论，采用语料库方法，通过索引行和高频词等手段探索中国政治术语的国际传播能力建设。

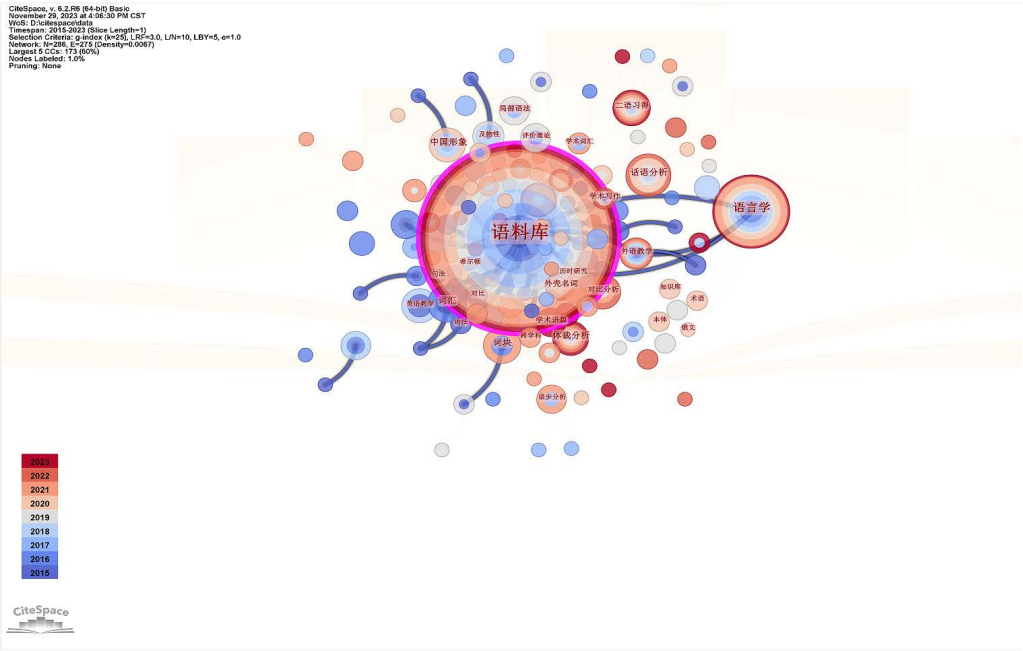

Figure 2. Keyword co-occurrence chart
图 2. 关键词共现图

3.3. 关键词突现

突现是指一个变量的值在短期内有很大变化，突现词指使用频次突然明显增多或在较短时间内突然出现的词[10]。本文在关键词共现图的基础上对关键词突现进行检测，检测出排名前十九的突现词，如图 3 关键词突现图所示。

在 2015 至 2016 年期间，突现的关键词为“词汇”“主位”“词块”“共词分析”“主题词”“应用”“中国英语”“学术词汇”“中国形象”“及物性”“语义韵”；在此期间突现的关键词较多，也再次证实上文语料库语言学研究在 2015 年到达研究峰值。其中中国形象的研究和及物性的突现强度较大，但是结束年份分别是 2018 和 2019 年。“对比分析”和“学术文本”不仅突现强度大，突现年份为 2021 年，

并且突现延续的周期较长，结束时间为 2023 年，这也说明语料库研究最新的研究主要集中在对比分析。

Top 19 Keywords with the Strongest Citation Bursts

Keywords	Year	Strength	Begin	End	2015 - 2023
词汇	2015	3.07	2015	2016	
主位	2015	1.75	2015	2016	
词块	2015	0.96	2015	2016	
共词分析	2015	0.87	2015	2016	
主题词	2015	0.87	2015	2016	
应用	2015	0.79	2015	2016	
中国英语	2016	1.34	2016	2017	
学术词汇	2016	0.64	2016	2017	
隐喻	2017	1.77	2017	2018	
中国形象	2015	1.01	2017	2018	
话语分析	2018	1.99	2018	2020	
及物性	2015	1.55	2018	2019	
german teaching	2018	0.67	2018	2020	
元话语	2019	0.95	2019	2020	
女性形象	2019	0.95	2019	2020	
功能	2019	0.68	2019	2021	
语义韵	2015	0.34	2019	2020	
对比分析	2019	1.59	2021	2023	
学术文本	2021	1.49	2021	2023	

Sort | Save and Display in a New Window as HTML | Save As a Tab Delimited File | Close

Figure 3. Keywords with the strongest citation bursts
图 3. 关键词突现图

3.4. 作者和机构分布知识图谱分析

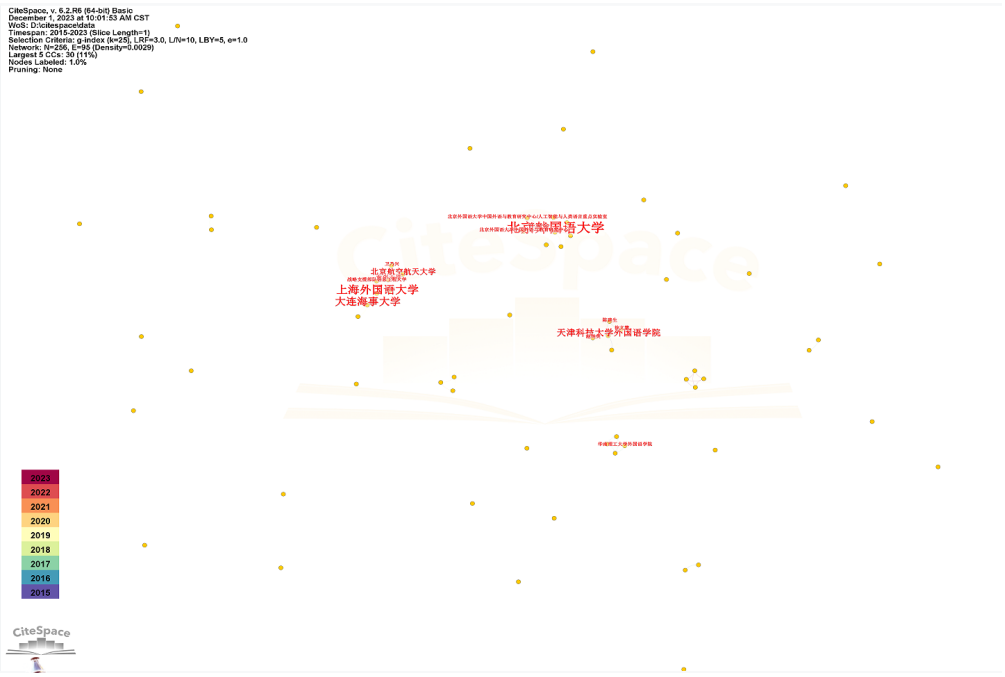

Figure 4. A knowledge map of authors and institutions
图 4. 作者和机构分布知识图谱

作者/机构的知识图谱能够帮助读者发现某一研究领域或机构之间的社会关系，了解相关领域具有影响力的学者或研究机构(陈悦，陈超美，胡志刚 2014)。利用 CiteSpace 可视化软件，勾选作者(Author)和机构(Institution)，通过可视化(Visualize)可得到图 4“作者和机构分布知识图谱”。由图 4 可知，发文最多的机构主要集中在北京外国语大学(23 篇)、上海外国语大学(13 篇)和浙江大学(13 篇)。主要年份集中在 2015 和 2016 年。其中甄凤超(6 篇)和许家金(4 篇)为该领域的主要代表。甄凤超教授在 2015 年发表语料库驱动的学习者英语动词配价研究：以 CONSIDER 为例；2016 年发表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。该研究探索了一套语料库数据驱动的动词配价结构描写体系，可以直接应用于英语学习者语料库分析中，影响深远。许家金教授在 2017 年发表论文“语料库研究学术源流考”，对语料库研究的核心概念和理论源流进行考证，并提出中国语料库的研究应立足国庆，坚持开展汉语语料库研究、汉语中介语语料库研究、中国学习者语料库研究和双语对比与翻译研究，为接下来语料库研究的发展指明方向。

4. 结语

本文基于 CiteSpace 可视化分析，对 2015~2023 年语料库研究的研究热点进行了全面系统的分析。通过对 776 篇论文的研究主题、发表年限、关键词共现、关键词突变等角度，描述了国内语料库研究的发展趋势和最新动态。结果表明，国内语料库研究在 2015 年为研究峰值，主要研究热点集中在“中国形象”和“及物性”，而在 2023 年，“对比分析”突现词延续时间较长，为语料库研究最新热点。本文对 2017~2023 年国内语料库语言学在知网上的 776 篇文献进行量化分析，较为客观，但同时也存在不足之处。如果将同期国外语料库语言学的研究同时纳入研究范围，就可以对国内外语料库语言学的研究进行对比，发现我国语料库研究的优势和不足。

参考文献

[1] 金碧希, 卫乃兴. 话语研究的语料库路径: 方法、挑战与前景[J]. 外语与外语教学, 2023(1): 1-11+144. https://doi.org/10.13458/j.cnki.flatt.004921

[2] 杨惠中. 语料库语言学导论[M]. 上海: 上海外语教育出版社, 2002.

[3] Sinclair, J.M. (1991) Shared Knowledge. In: Alatis, J.E., Ed., *Georgetown University Round Table on Language and Linguistics* 1991, Georgetown University Press, Washington, DC.

[4] Aarts, M.B. (2000) Corpus Linguistics, Chomsky and Fuzzy Tree Fragments. In: Mair, C. and Hundt, M., Eds., *Corpus Linguistics and Linguistic Theory*, Brill, Amsterdam, 5-13. https://doi.org/10.1163/9789004490758_003

[5] 张毓, 卫乃兴. 基于 LDA 主题建模技术的北京冬奥会话语意义研究[J]. 语料库语言学, 2023, 10(1): 1-13+160.

[6] 赵冲, 许家金. 近百年西班牙语语料库建设与研究概述[J]. 欧洲语言文化研究, 2023(1): 119-135+149.

[7] 姜峰, Ken Hyland. 互动元话语: 学术语境变迁中的论辩与修辞[J]. 外语教学, 2020, 41(2): 23-28. https://doi.org/10.16362/j.cnki.cn61-1023/h.2020.02.005

[8] 冯佳, 王克非, 刘霞. 近二十年国际翻译学研究动态的科学知识图谱分析[J]. 外语电化教学, 2014(1): 11-20.

[9] 孙丰果. 国家话语视域下中美国家身份建构对比研究[D]: [博士学位论文]. 北京: 北京外国语大学, 2023. https://doi.org/10.26962/d.cnki.gbjwu.2023.000033

[10] 胡敏燕, 戴运财. 国内外国语言学及应用语言学研究热点与趋势——基于外语类核心期刊的知识图谱可视化分析[J]. 浙江外国语学院学报, 2023(2): 62-71.