Modern Linguistics 现代语言学, 2023, 11(8), 3483-3490
Published Online August 2023 in Hans. https://www.hanspub.org/journal/ml
https://doi.org/10.12677/ml.2023.118471

Hans汉斯

基于CiteSpace的语料库语言学研究的可视化分析

杨娇娇

烟台南山学院，国学与外语学院，山东 烟台

收稿日期：2023年6月25日；录用日期：2023年8月2日；发布日期：2023年8月15日

摘　要

研究选取2011~2020年CSSCI和SSCI收录语料库语言学期刊论文，运用CiteSpace对发文量、作者、机构、高被引文献和关键词开展可视化分析。结果发现国内语料库语言学发展态势不如国际强劲；国内研究人员和机构合作关系网较为简单，合作密切度有待加强；国内研究热点涉及平行语料库和语料库翻译学，第三语码和及物性为研究前沿，而国际研究热点涉及语言习得、语言结构和语法，法语、德语、模型和词束为研究前沿，多语研究趋势明显。

关键词

语料库语言学，CiteSpace，CSSCI，SSCI，可视化分析

A CiteSpace-Based Visual Analysis of Researches on Corpus Linguistics

Jiaojiao Yang

College of Chinese Studies and Foreign Languages, Yantai Nanshan University, Yantai Shandong

Received: Jun. 25th, 2023; accepted: Aug. 2nd, 2023; published: Aug. 15th, 2023

Abstract

The study selects linguistic journal articles from the CSSCI and SSCI databases from 2011 to 2020 and uses CiteSpace for visual analysis of publication volume, authors, institutions, highly cited literature, and keywords. The results show that the growth trend of corpus linguistics in China is not as good as that in the world. Cooperative networks of domestic researchers and institutions are relatively simple and the degree of cooperation needs to be strengthened. Domestic research

文章引用: 杨娇娇. 基于 CiteSpace 的语料库语言学研究的可视化分析[J]. 现代语言学, 2023, 11(8): 3483-3490.
DOI: 10.12677/ml.2023.118471

hotspots include parallel corpora and corpus translation, and its research frontiers are third-language code and transitivity. However, international research hotspots include language acquisition, language structure and grammar. Its research frontiers are French, German, models and lexical bundles, and the trend of multilingual research is obvious.

Keywords

Corpus Linguistics, CiteSpace, CSSCI, SSCI, Visual Analysis

1. 引言

语料库(corpus 或 corpora)，McEnery 等学者认为语料库是按照严格取样标准选取的真实的机器可读的口语或书面语电子文本的集合[1]。自 20 世纪 60 年代布朗语料库建立以来，随着科学技术的进步，语料库语言学也逐步发展壮大，相关研究成果颇丰，在词典编撰[2]、语言教学[3] [4]和话语分析[5]等方面发挥着重要作用。甄凤超[6]、张继东和陈文[7]以及刘霞等[8]国内学者也对近些年语料库语言学研究进行综述，多集中于国内和国际语料库语言学单方面综述，国内外语料库语言学开展可视化对比考察研究较少，如华正雷[9]。因现有研究中少有学者对国内外语料库语言学研究进行对比综述，可视化对比国内外语料库语言学研究现状。故研究主要借助 CiteSpace 文献计量软件，选取 CSSCI (中文社会科学引文索引)和 SSCI (社会科学引文索引)收录期刊论文文献，细致地对 2011~2020 年国内外语料库语言学的研究现状和研究热点进行可视化分析，以期系统对比国内外语料库语言学研究现状并追踪发展趋势。

2. 数据来源与研究方法

2.1. 数据来源

本研究选取的国内文献数据来源于 CSSCI 期刊。国内文献检索条件为：选择“关键词”选项，检索字段为“语料库”，类型为“论文”，检索文献年限为 2011~2020 年，研究学科为语言学，共检索到 566 篇文献。研究选取的国际文献数据源自 Web of science 数据库的 SSCI 子库。国际文献检索条件为：选择“主题词”选项，检索字段为“corpus”和“corpora”，检索字段之间的关系为“or”，文献年限选为 2011~2020 年，研究类别为 linguistics 和 language linguistics，文献类型选择学术论文，共检索到 5406 篇学术论文。

2.2. 研究方法

自科学知识图谱于 2005 引进中国以来，该新颖科学计量学方法在我国迅速蓬勃发展[10]。知识图谱为科学地计量可视化相关研究数据提供了便利。其可快捷处理海量的大型数据，采取可视化方式呈现相关数据结果，绘制知识图谱，简单明了地揭示海量数据的特征。目前，文献计量软件有 Pajek、Vosviewer、HistCite、CiteSpace 和 Bibliometrix 等，这些软件在知识图谱的绘制上各有优势[11] [12] [13]。本研究主要运用 CiteSpace 软件，以时间切片为 1 年和阈值为 Top N = 30 呈现的知识图谱形式，可视化作者、机构以及关键词等相关信息，便于清晰地识别 2011~2020 年国内外语料库语言学领域研究现状和发展趋势。

3. 发文趋势

期刊论文的产出量是衡量领域知识发展水平的重要指标[14]。研究运用 Excel 统计 2011~2020 年国内外语料库语言学的年度文献数量，从时间角度呈现国内和国际语料库语言学领域发文数量的动态变化，利于把握发展的整体趋势。

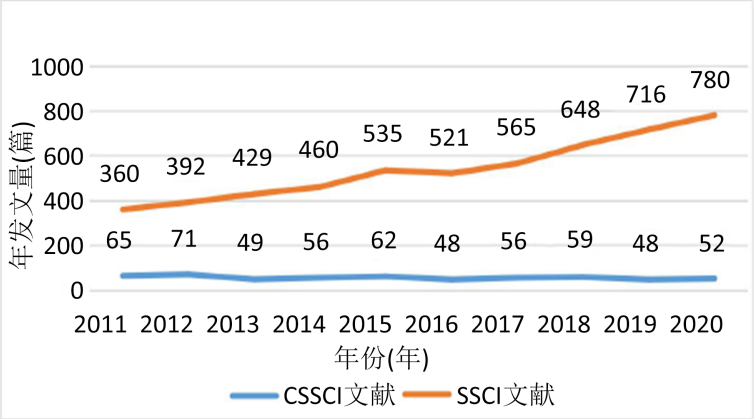

Figure 1. Line chart of the annual publication volume of corpus linguistics in CSSCI and SSCI
图 1. CSSCI 和 SSCI 语料库语言学年度文献数量

如图 1 所示，CSSCI 收录的 2011~2020 年语料库语言学发文趋势增长态势不足，整体略微缩减趋势，而 SSCI 收录的 2011~2020 年语料库语言学发文量一直保持整体增长趋势。这说明在 2011~2020 年期间，国际语料库语言学的相关研究继续蓬勃发展，而国内学者对语料库语言学的研究热度相对平稳，国内学者对语料库语言学研究的关注度稍逊于国际学者。

3.1. 主要作者和机构分析

作者和机构合作网络知识图谱可方便研究者迅速发现研究领域的核心作者和主要研究机构，结点越大的圆圈说明作者和机构的发文量和影响力越高。研究将对 2011~2020 年 CSSCI 和 SSCI 语料库语言学研究文献开展高产作者、高被引作者以及高产机构可视化分析。

3.2. 高产作者和高被引作者分析

CSSCI 高产作者合作网络知识图谱网络密度(D = 0.0026)略高于 SSCI 高产作者合作网络密度(D = 0.002)，这说明 2011~2020 年国内外语料库语言学研究比较分散，作者间合作较少，合作关系不够紧密，但国内语料库语言学高产作者间合作网络紧密度略高于国际语料库语言学。如图 2 所示，2011~2020 年 CSSCI 语料库语言学文献的高产学者有王克非，胡开宝，卫乃兴，庞双子和张威等人，王克非教授是发文量最高学者。2011~2020 年 SSCI 语料库语言学文献的高产作者主要有 Biber，Gries，Speelman，Liu HT 和 Hyland 等，发文量最高学者为 Biber。这些国内外学者为语料库语言学的发展做出了重要学术贡献。

一般情况下，文献被引频次体现学术影响力。研究对 2011~2020 年 CSSCI 和 SSCI 语料库语言学高被引作者开展分析。分析结果如图 3 所示，这些高被引作者为 2011~2020 年国内外语料库语言学高影响力学者。CSSCI 语料库语言学研究中前五的高影响力作学者为王克非，Biber，卫乃兴，Baker 和胡开宝。SSCI 语料库语言学研究中前五的高影响力作学者为 Biber，Anonymous，Hyland，Sinclair 和 Quirk。

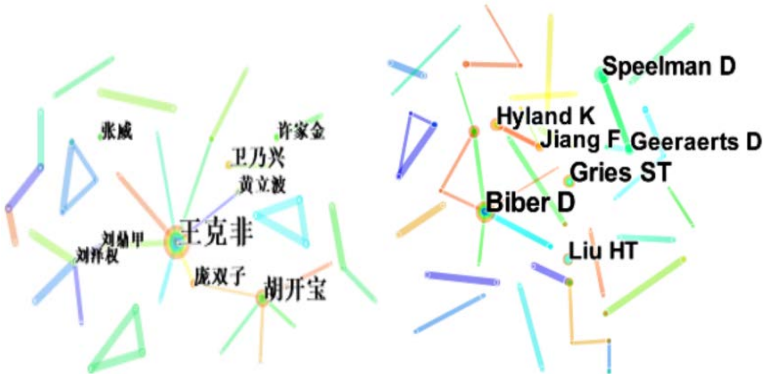

Figure 2. Knowledge map of high-yield author of corpus linguistics in CSSCI and SSCI
图 2. CSSCI 和 SSCI 语料库语言学高产作者知识图谱

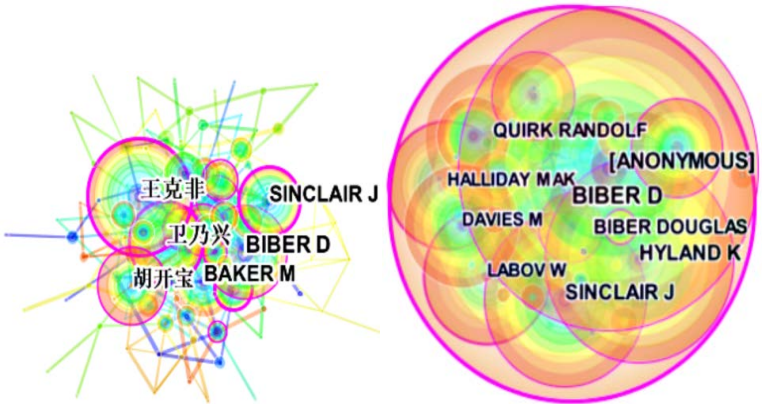

Figure 3. Knowledge map of highly-cited author of corpus linguistics in CSSCI and SSCI
图 3. CSSCI 和 SSCI 语料库语言学高被引作者知识图谱

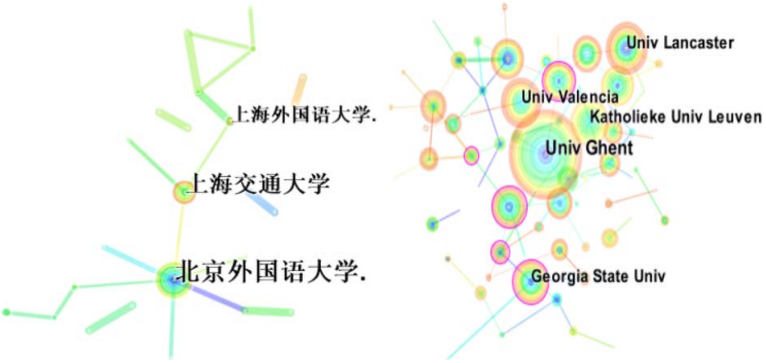

Figure 4. Knowledge map of high-yield institution of corpus linguistics in CSSCI and SSCI
图 4. CSSCI 和 SSCI 语料库语言学高产机构知识图谱

由图 2 和图 3 可知，在国内语料库语言学文献中，王克非学者不仅发文量最高，且影响力最高，其于 2006 年率先提出语料库翻译学概念，是该领域领军人物。在国外语料库语言学文献中，发文量和被引量最高学者为 Biber，其致力于语体变异、语法及语篇类型的研究，在语料库语言学领域造诣颇高。

3.3. 高产机构分析

CSSCI 高产机构合作网络知识图谱网络密度(D = 0.0011)小于 SSCI 高产机构知识图谱网络密度(D = 0.013)，这表明国内研究机构没有形成较为密切的合作关系，研究机构间交流也不够紧密，国际研究机构间的合作关系网比国内研究机构合作关系网更为密切。

由图 4 可以直观发现，2011~2020 年 CSSCI 语料库语言学前三的高产研究机构有北京外国语大学、上海交通大学和上海外国语大学；2011~2020 年 SSCI 语料库语言学发文量前三的高产研究机构有比利时 Ghent University (根特大学)、英国 Lancaster University (兰卡斯特大学)和西班牙 Universitat de València (瓦伦西亚大学)。在 2011~2020 年，这些机构处于国内外语料库语言学研究领先地位。

4. 高被引文献分析

文献共被引分析知识图谱可以清晰展示该领域的高影响力文献，便于读者清晰了解领域内的核心文献。由图 5 可知，2011~2020 年国内语料库语言学 CSSCI 前五的高被引文献有王克非(2008)，胡开宝(2010)，王克非(2009)，秦洪武(2009)和黄立波(2012)。

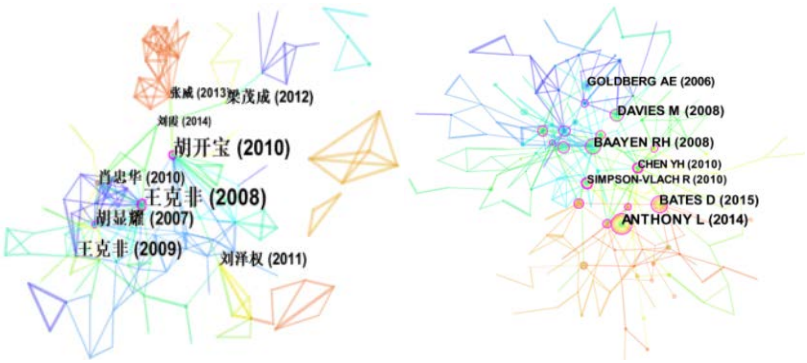

Figure 5. Co-citation network map of literatures of corpus linguistics in CSSCI and SSCI
图 5. CSSCI 和 SSCI 语料库语言学文献共被引知识图谱

王克非和胡显耀对翻译汉语词汇特征进行语料库研究，对比研究翻译汉语和汉语本身，探讨了翻译共性的问题，对语料库翻译研究提供不少借鉴[15]。胡开宝和陶庆详细介绍其团队建立的国内首个英汉会议口译语料库(CECIC)的创建与应用，突出口译研究的重要性，为后续国内汉英会议口译相关研究做出了重大贡献[16]。王克非和秦洪武从宏观视角描述和分析 CEPC 内英译汉文本的总体特征[17]，同年，两者还分析 CEPC 中英译汉文本词汇特征，发现汉语翻译文本的词类和词的组合上与汉语原创文本存在差异[18]。黄立波是国内最先探索译者风格的领军学者，其与朱志瑜运用葛浩文与戴乃迭的翻译作品建成平行语料库，来考察两者的翻译风格，对 Baker“译者风格”方法论提出质疑，指出仅语料库软件统计的标准类/形比等参数不能有效区分译者风格，认为译者风格研究分 S-型和 T-型，将源文本考虑在内的 S-型研究对翻译研究更有意义[19]。

2011~2020 年国际语料库语言学高被引文献前五的文献有 Anthony (2014)，Bates (2015)，Baayen (2008)，Davies (2008)和 Goldberg (2006)。Anthony 教授研发了免费简单高效的单语语料库检索分析工具 AntConc 软件，广泛应用于语料库语言学研究[20]。Bates 介绍了其研发的免费开源程序 R 语言内 lme4 包，为语言学相关研究数据的统计分析提供计算机技术支持，推动语料库语言学进一步发展。[21] Baayen 开创线性混合模型，为语料库语言学的量化研究提供统计学技术支持[22]。而 Davies 创建的 COCA 语料

库是世界上最大免费英语在线语料库，操作简单且时效性强，为语料库语言学研究提供了强大语料库数据支持[23]。Goldberg 更深刻阐述语言概括的本质，论证构式语法理论对语言习得的解释力[24]，是语言科学的一项重大进展[25]。

由 CSSCI 和 SSCI 高被引文献可知，国内外语料库语言学都注重语料库创建。不同之处在于国内侧重于语料库翻译等应用研究，而国际语料库语言学侧重于语料库语言学技术性和理论性研究。

4.1. 研究热点

关键词一定程度上表征文章的核心内容，表达文献主题内容，属于文献计量研究的重要内容。关键词共现分析可突显研究领域的关键结点，展现一定时间内该领域的研究热点，有助于把握这段时间内相关研究的整体概况。

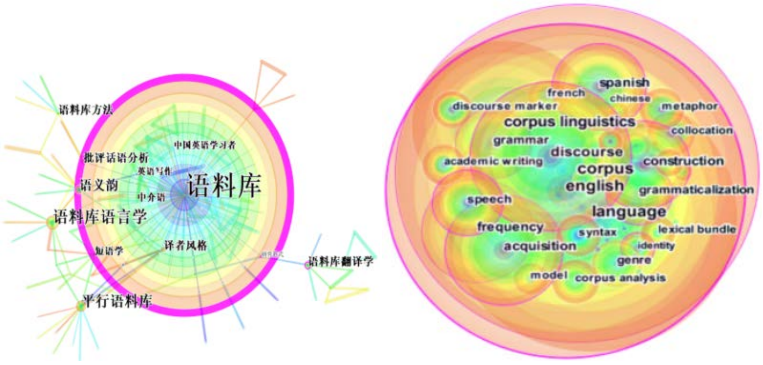

Figure 6. Keyword co-occurrence network of corpus linguistics in CSSCI and SSCI
图 6. CSSCI 和 SSCI 语料库语言学关键词共现知识图谱

如图 6 所示，CSSCI 语料库语言学共现频次最高的关键词为语料库，其共现频次为 305 次，中心度为 1.06。共现频次较高的关键词还涉及平行语料库、语义韵、语料库翻译学、批评话语分析、译者风格和短语学等。因此，2011~2020 年国内语料库语言学研究热点涉及平行语料库研究、语义韵研究、语料库翻译学研究、批评话语分析研究和短语学研究等，这说明语料库研究运用到翻译学和话语分析等更多领域中。

SSCI 语料库语言学共现频次最高的关键词为 corpus，其共现频次为 609 次，中心度为 0.17。共现频次第二的关键词为 English，说明国际语料库语言学研究语言主要为英语。除了英语外，Spanish、French 和 Chinese 等关键词共现频次也较高，西班牙语、法语和汉语相关研究也不少，这表明近些年国际语料库语言学学者们对西班牙语、法语和汉语等语料库研究关注度较高。discourse、acquisition、construction、grammar 和 genre 等关键词共现频次也较高，即 2011~2020 年国际语料库语言学的研究热点涉及语篇，语言习得、语言结构、语法以及体裁等。

4.2. 研究前沿

突显词图谱突出了研究热点关键词的演变过程，便于发现研究领域的新兴热点，可观察研究发展趋势，追踪研究前沿。由图 7 突显词知识图谱可知，2011~2020 年，国内语料库语言学文献的突显词有 21 个，突显度较强的关键词为英语写作和语料库语言学。由突显词的演变时间可知，及物性，语料库方法和第三语码等为国内语料库语言学的近几年研究热点，即 2011~2020 年国内语料库语言学研究前沿涉及及物性和语料库翻译学等。

Top 21 Keywords with the Strongest Citation Bursts

Keywords	Year	Strength	Begin	End	2011 - 2020
英语写作	2011	2.2971	2011	2012	
红楼梦	2011	1.834	2011	2012	
中介语语料库	2011	1.4709	2011	2013	
中国英语	2011	1.4606	2012	2013	
汉语教学	2011	1.2111	2013	2013	
系统功能语法	2011	1.2111	2013	2013	
语义韵	2011	1.6107	2014	2016	
类联接	2011	1.2138	2014	2014	
翻译规范	2011	1.2138	2014	2014	
语料标注	2011	1.7826	2015	2015	
学习者语料库	2011	1.5659	2015	2017	
历时语料库	2011	1.6106	2016	2018	
可比语料库	2011	1.6106	2016	2018	
汉英翻译	2011	1.2055	2016	2016	
学术写作	2011	1.2055	2016	2016	
短语学	2011	1.5776	2016	2017	
配价结构	2011	1.2055	2016	2016	
语料库语言学	2011	2.2221	2017	2017	
及物性	2011	1.2083	2018	2018	
语料库方法	2011	1.3696	2018	2018	
第三语码	2011	1.2277	2019	2020	

Top 25 Keywords with the Strongest Citation Bursts

Keywords	Strength	Begin	End	2011 - 2020
discourse analysis	8.4447	2011	2012	
identity	4.9861	2011	2013	
context	9.606	2011	2012	
syntax	3.1133	2011	2013	
perception	5.9718	2012	2014	
phraseology	9.9184	2012	2013	
dutch	7.6434	2012	2013	
conversation	9.5097	2012	2015	
formulaic language	6.9383	2013	2015	
prosody	9.7025	2014	2015	
metaphor	4.69	2014	2015	
speaker	7.8933	2014	2016	
learner	6.6716	2014	2015	
discourse marker	3.9056	2014	2016	
conversation analysis	8.6259	2015	2017	
research article	7.7625	2016	2017	
gender	6.2215	2016	2017	
organization	9.9146	2016	2018	
pragmatics	13.7259	2016	2018	
complexity	10.0097	2017	2018	
french	3.9937	2017	2020	
word	3.4615	2017	2018	
german	6.3663	2018	2020	
model	6.8017	2018	2020	
lexical bundle	6.0967	2018	2020	

Figure 7. Keyword burstiness map of corpus linguistics in CSSCI
图 7. CSSCI 和 SSCI 语料库语言学关键词突显词图谱

SSCI 语料库语言学文献的突显词图谱清晰显示了 2011~2020 年国际语料库语言学领域的研究热点演变。国际语料库语言学文献的突显词有 25 个，其中突显强度较强的关键词为语用学(pragmatics)和复杂度(complexity)。且由突显词演变时间可知，国际语料库语言学的研究前沿为法语(French)、德语(German)、模型(model)以及词束(lexical bundle)等研究，这说明近年来关于法语、德语、模型和词束等的研究较为活跃。

5. 结论

本研究借助 CiteSpace 对 2011~2020 年间国内外语料库语言学文献开展可视化分析，较为充分展示和追踪国内外语料库语言学的研究现状和发展趋势，可为后续相关研究一定参考。在发文量方面，国际语料库语言学发展态势强劲，国内语料库语言学年度文献量则整体略呈下降态势；在机构和作者方面，国际语料库语言学研究机构间合作关系网更为密切；在高被引文献方面，国内学者侧重语料库应用研究，国际学者侧重语料库技术性和理论研究；在研究热点和研究前沿方面，国内外研究有所不同，国内语料库语言学侧重平行语料库和翻译研究等，第三语码和及物性为发展方向；国际语料库侧重于语言习得、语言结构等研究，法语、德语、模型和词束等为发展方向。基于本研究的发现，国内语料库语言学研究的增长态势有待加强，国内学者间和机构间的交流合作应更深入，加强合作，构建紧密合作关系，以及国内语料库语言学的技术性和理论性等方面研究有待进一步发展。

参考文献

[1] McEnery, T., Xiao, R. and Tono, Y. (2006) Corpus-Based Language Studies: An Advanced Resource Book. Routledge, London.

[2] 吴先, 胡俊峰. 基于历时语料库的在线词典编纂系统设计[J]. 中文信息学报, 2020(5): 27-35.

[3] 桂诗春, 冯志伟, 杨惠中, 等. 语料库语言学与中国外语教学[J]. 现代外语, 2010, 33(4): 419-426.

[4] 何安平. 语料库与外语教学[J]. 外语教学理论与实践, 2001(3): 15-19.

[5] 卫乃兴. 语料库语言学的方法论及相关概念[J]. 外语研究, 2009, 26(5): 36-42.

[6] 甄凤超. 语料库语言学研究热点追踪与思考[J]. 当代外语研究, 2020(6): 89-100+4-5.

[7] 张继东, 陈文. 国际语料库语言学研究的可视化分析[J]. 外语电化教学, 2016(6): 66-73.

[8] 刘霞, 许家金, 刘磊. 基于 CiteSpace 的国内语料库语言学研究概述(1999-2013) [J]. 语料库语言学, 2014(1): 77-85+120.

[9] 华正雷. 基于 CiteSpace 的国内外语料库研究对比考察[J]. 安徽电气工程职业技术学院学报, 2020, 25(2): 65-78.

[10] 陈悦, 陈超美, 刘则渊, 胡志刚, 王贤文. CiteSpace 知识图谱的方法论功能[J]. 科学学研究, 2015, 33(2): 242-253.

[11] 曹增节. 艺术传播学——文献计量学方向[M]. 杭州: 中国美术学院出版社, 2014.

[12] 邱均平. 信息计量学概论[M]. 武汉: 武汉大学出版社, 2019.

[13] 周春雷, 张猛. 知识图谱软件学术影响力研究[J]. 信息资源管理学报, 2019, 9(1): 85-93.

[14] 刘璐达, 季云飞, 姜峰. 国内外体裁分析研究对比考察——基于中国知网和 Web of Science 数据库的可视化分析[J]. 中国 ESP 研究, 2020(4): 23-33+111.

[15] 王克非, 胡显耀. 基于语料库的翻译汉语词汇特征研究[J]. 中国翻译, 2008, 29(6): 16-21+92.

[16] 胡开宝, 陶庆. 汉英会议口译语料库的创建与应用研究[J]. 中国翻译, 2010, 31(5): 49-56+95.

[17] 王克非, 秦洪武. 英译汉语言特征探讨——基于对应语料库的宏观分析[J]. 外语学刊, 2009(1): 102-105.

[18] 秦洪武, 王克非. 基于对应语料库的英译汉语言特征分析[J]. 外语教学与研究, 2009(2): 131-136+161.

[19] 黄立波, 朱志瑜. 译者风格的语料库考察——以葛浩文英译现当代中国小说为例[J]. 外语研究, 2012(5): 64-71.

[20] Anthony, L. (2014) AntConc [Computer Software]. Waseda University, Tokyo. https://www.laurenceanthony.net/software

[21] Bates, D., Maechler, M., Bolker, B. and Walker, S. (2015) Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software*, **67**, 1-48. https://doi.org/10.18637/jss.v067.i01

[22] Baayen, R.H., Davidson, D.J. and Bates, D.M. (2008) Mixed-Effects Modeling with Crossed Random Effects for Subjects and Items. *Journal of Memory and Language*, **59**, 390-412. https://doi.org/10.1016/j.jml.2007.12.005

[23] Davies, M. (2008) The Corpus of Contemporary American English. https://corpus.byu.edu/coca

[24] Goldberg, A.E. (2006) Constructions at Work: The Nature of Generalization in Language. Oxford University Press, New York. https://doi.org/10.1093/acprof:oso/9780199268511.001.0001

[25] 陆俭明. 构式语法理论再议——序中译本《运作中的构式: 语言概括的本质》[J]. 外国语(上海外国语大学学报), 2013, 36(1): 16-21.