中国国内日语语料库研究热点与趋势

－以CiteSpace对CNKI文献的关键词分析为例－

王雯婷 马小兵

Visualization analysis of Chinese:

Japanese corpus research in CNKI based on CiteSpace

Wang Wenting, Ma Xiaobing

In order to explore the research trends and hotspots related to the Chinese-Japanese corpus research, this study collected information on 527 publications from CNKI (China National Knowledge Infrastructure) and conducted a quantitative and visual analysis. Hot keywords are "Corpus", "Transition", "Japanese language teaching", "Empirical method". Combined with keywords and cluster analysis, "Contrast", "Translation" has become new hot research directions since 2010.

Keywords：corpus, Japanese, Visualization analysis, CiteSpace

引言

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[1]。"语料库"一词来自拉丁语"corpus"，意为"汇总""文集"。一般认为，1967年美国布朗语料库的建立和相关论文的发表标志着语料库研究在现代

1 杨惠中．语料库语言学导论［M］．上海：上海外语教育出版社，2002.

语言学意义上的开端[2]。

20世纪80年代以来，随着语料库语言学在世界范围内的蓬勃发展，这门学科被引入我国[3]。宋红波等（2013）从词汇、语法、语义等十三个方面对当前语料库语言学研究进行了分类，并指出教学方面的研究成果最为显著，其次是翻译研究、语法、语义、词典、词汇研究等。

然而相对于欧美语料库研究的蓬勃发展，日语语料库研究处于相对落后的状态[4]。为了更全面地了解近年来国内在日语语料库方面研究方面的发展态势，把握该领域研究的热点和前沿问题，本研究拟运用CiteSpace软件，对现有文献进行可视化分析，以期为今后日语语料库的研究提供借鉴与参考。

研究背景与数据采集

杨本明（2018）将国内日语语料库的发展分为三个阶段，第一阶段是计算机化以前的阶段，称之为传统语料库时期，主要以卡片语料库为主。第二阶段为计算机化以后的阶段，称之为现代语料库时期。第三阶段为超级计算机存储阶段，称之为大数据语料库时期。

20世纪90年代以前，国内语料库的建设一般是以卡片存储的方式建立的，这种日记本式的语料库建设需要人工书写，占据空间大，不方便查阅，规模也极其有限。20世纪90年代以来，随着计算机存储技术和网络技术的发展，国内的日语语料库建设开始有了起色。2000年以后，日语语料库建设进入快速发展时期。其中，北京日本学研究中心徐一平教授团队建设的《中日对译语料库》和上海外国语大学毛文伟教授建设的《中国日语学习

2　杨柳．国际语料库语言学研究热点与前沿的信息可视化分析［J］．知识管理论坛，2018，3（04）：208-224.
3　宋红波，王雪利．近十年国内语料库语言学研究综述［J］．山东外语教学，2013，34（03）：41-47.
4　毛文伟．日语语料库建设的现状综述［J］．日语学习与研究，2009（06）：42-47.

者语料库》极具代表性。

本研究采集的数据来自中国知网（CNKI），包含“学术期刊”、“硕博”、“会议”、“报纸”数据库。检索条件为：主题含“语料库”并含“日语”；或主题含“语料库”并含“日文”；或主题含“コーパス”并含“中日”；或主题含“语料库”并含“日汉”；或主题含“计算机”并含“日语”；或主题含“计算机”并含“日汉”；或主题含“语料库”并含“汉日”。截至2019年7月19日共检索到文献527篇，作为研究的样本。

研究方法

陈悦等（2015）[5]指出，Cite Space是应用Java语言开发的一款信息可视化软件，它主要基于共引分析理论（co-ciation）和寻径网络算法（path Finder）等，对特定领域文献（集合）进行计量，以探寻出学科领域演化的关键路径及其知识拐点，并通过一系列可视化图谱的绘制来形成对学科演化潜在动力机制的分析和学科发展前沿的探测。

杨柳（2018）指出，关键词共现知识图谱能够将具有相同关键词的文章进行聚类，进而体现出同一研究领域的关键节点，集中展现一段时间内相关文献的研究热点，有利于从整体上把握已有研究内容。同时，通过对关键词共现产生的中心性分析可以揭示出研究热点之间的转化关系。

在数据可视化方面，Cite Space软件提供了三种可视化方式：聚类视图（cluster）、时间线视图（Timeline）和时区视图（timezone）。分别侧重体现聚类间的结构特征，突出关键节点及重要连接；勾画聚类之间的关系和某个聚类中文献的历史跨度；从时间维度上表示知识演进，清晰地展示出文献的更新和相互影响。

5 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（02）：242-253.

本研究将从关键词共现的角度，制作可视化图谱，并结合关键词表，展示国内日语语料库领域的热点和发展动向。

图谱制作与分析

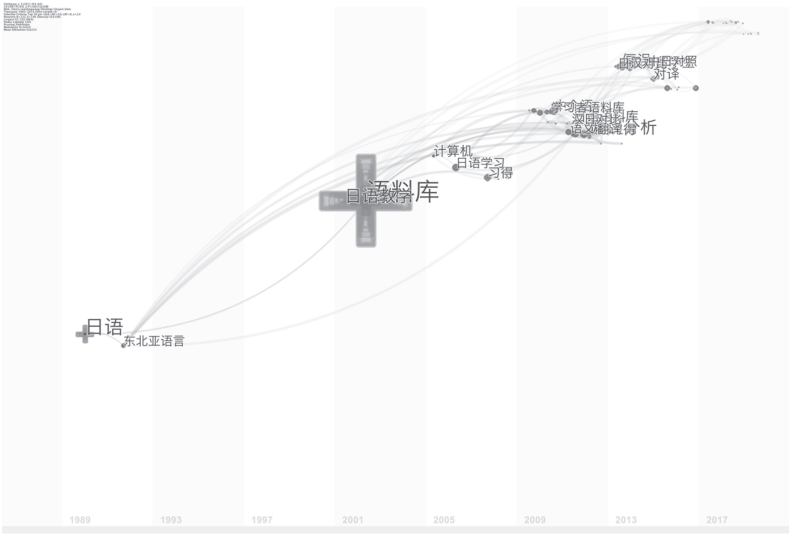
本研究以CNKI数据库1985年至2019年的527篇文献为研究样本，节点类型设置为“关键词（Keyword）”，时区分割设置为４，阈值设置为前50。结果表明，共有122个节点，146条连线，密度为0.0195。

对所得结果进行关键词聚类，可得出聚类视图（图１）、时间线视图（图２）、时区视图（图３）。

（图１）

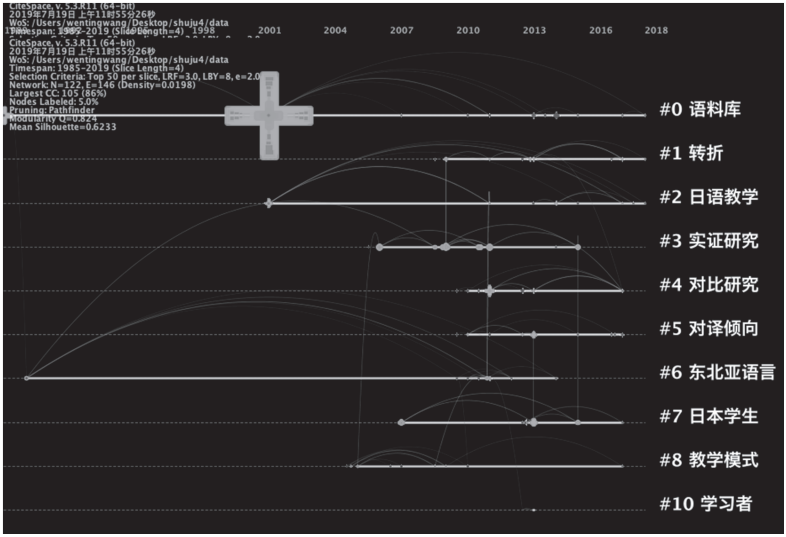
图（１）聚类视图（cluster）侧重于体现聚类间的结构特征，突出关键节点及重要连接。其中视图中的节点代表分析的对象，出现频次（或被引频次）越多，节点就越大。节点内的颜色及厚薄度表示不同时间段出

现（或被引）频次。节点之间的连线则表示共现（或共引）关系，其粗细表明共现（或共引）的强度。由图可知，国内日语语料库研究大体可以分为以下几个方面：日语教学、对译、实证研究、偏误分析、二语习得、对比研究等。



（图2）

图（2）时间线视图（Timeline）侧重于勾画聚类之间的关系和某个聚类中文献的历史跨度。其中横轴坐标为文献发表时间，纵轴坐标为节点所属的聚类。由图可知，“日语”、“语料库”始终贯穿整条时间线。2000年以前的研究比较单一；2001年以来开始出现“日语教学”“日语学习”与“日语习得”；而到了2009年之后，研究范围进一步拓宽，出现了“中日对照”、“对译”、“汉日对比”。



（図３）

図（３）时区视图（timezone）是另一种侧重于从时间维度上来表示知识演进的视图，可以较为清晰地展示出文献的更新和相互影响。时区视图展示了领域文献的增长，某一时区的文章越多，说明这一时间段中发表的成果越多，该领域处于繁荣时期；某一时区中的文献越少，说明这一时间段中发表的成果越少，该领域处于低谷时期。由图可知，2000年以前的研究涉及范围较窄；2001年之后，“日语教学”开始与语料库研究相结合；2005年左右出现了“实证研究”与“教学模式”；而到了2010年之后，“对比研究”与“对译倾向”成为新的研究热点。

此外，对所得结果进行关键词、突现性统计，可得出关键词表（表１）。

（表1）

Freq	Burst	Centrality	Sigma	PageRank	Keyword
147	4.13	0.42	4.30	0.00	语料库
49	7.88	1.08	323.77	0.00	日语
26	3.77	0.22	2.14	0.00	日语教学
24		0.63	1.00	0.00	偏误分析
14		0.02	1.00	0.00	偏误
11		0.05	1.00	0.00	习得
11		0.27	1.00	0.00	中日对照
11		0.05	1.00	0.00	中介语
11		0.14	1.00	0.00	日语语料库
11	3.90	0.26	2.44	0.00	对译
11		0.00	1.00	0.00	日本留学生
10		0.79	1.00	0.00	日语学习
10		0.12	1.00	0.00	日汉对比
10		0.10	1.00	0.00	二语习得
10		0.83	1.00	0.00	计算机
9		0.24	1.00	0.00	汉日对比
9		0.38	1.00	0.00	语义
9		0.02	1.00	0.00	学习者语料库
8		0.00	1.00	0.00	翻译
8		0.80	1.00	0.00	实证研究

表（1）中Freq代表频率；Burst代表突现性，即一个变量的值在短期内有很大变化；Centrality代表中心性，在CiteSpace中，中心性超过0.1的节点称为关键节点；Sigma是基于中心性和突现性计算得到的，中心性和突现性越高的节点，其sigma值也越高，绝大多数的sigma值是1.00，表示结构上和引文变化中都非常重要。

由表可知，频率较高的关键词有语料库（147）、日语（49）、日语教学（26）、偏误分析（24）、偏误（14）、习得（11）；突现性较高的关键词

有日语（7.88）、语料库（4.13）、对译（3.90）、日语教学（3.77）；中心性较高的关键词有日语（1.08）、计算机（0.83）、实证研究（0.8）、日语学习（0.79）、偏误分析（0.63）。Sigma值较高的有日语（323.77）、语料库（4.30）、对译（2.44）、日语教学（2.14）。

其中，“偏误”、“中介语”、“日本留学生”几个关键词涉及对外汉语教学。如段科慧（2015）结合了中日学者对日本学生汉语习得偏误问题的研究成果，按偏误的显性与隐性特征将日本学生在“HSK动态作文语料库”中出现的偏误分为书写、词汇、语法、语用与文化偏误等类型。王桢、毕晴（2019）基于BCC中介语语料库当中日本国籍学生“比”字句的语料，主要从母语迁移以及目的语泛化的角度对其展开偏误分析和归类,将“比”字句的偏误分为八类并针对八类偏误产生的原因提出了相应的教学建议。

对上述结果进行分析，可以看出：

在研究热点方面，语料库这一关键词贯穿始终，其次分别为：转折、日语教学、实证研究、对比研究、对译倾向、东北亚语言、日本学生、教学模式、学习者。

在阶段划分方面，2000以前，有关语料库的研究相对单一。一方面这与CNKI对文献的收录年份有关，另一方面也与我国日语语料库的发展有关。在这一阶段中，施建军（1991）介绍了机器单词辞典和单词的自动切分，认为和国内相比，国外计算机早已进入了语言研究领域，日本是利用计算机从事语言研究最早的国家之一。

2000年以来，日语教学开始成为语料库研究领域的热点，而在随后的几年中，又增加了实证研究、教学模式等几个关键词。在这一阶段中，吴英杰（2001）认为，计算机技术的发展日新月异，其发展和普及正日益改变着人们的工作、学习和生活方式。同时计算机技术的运用也为日语教学带来了新的活力，并就日语教学实践中的计算机应用问题做出了探讨。

冯峰（2002）通过对 6 所大学223名日语学习者的计算机软件利用情况的问卷调查，分三个部分论述外语学习中计算机的使用条件和使用学习软件的有利之处、当前在校生日语学习中计算机的使用状态及对计算机日语学习软件的需求和展望。

施建军等（2003）认为，随着大规模语料库的建设及计算机性能的提高，在语言学界已经形成这样一种共识，即仅靠语言学家的内省和自己的造句不能够充分解释语言现象，并就日语研究需要什么样的语料库，怎样利用语料库所进行的日语研究进行了论述。

毛文伟（2005）从语料库素材的规模、类型、年代等各项特征出发，探讨了在运用语料库进行语言研究时需要注意的一些问题，并分析了网上信息的特征，通过与传统出版物进行比较，指出以此为基础进行语言研究时有可能遇到的一些困难。

2010年以来，研究热点有了新的转向，转折、对比、对译等研究内容走入研究者的视线。在这一阶段中，高宁（2013）通过理论研究及语料库实证的方法，认为从对比语言学视角看，汉日语序呈现出较高的相似性；从翻译实践的角度看，日语原文和汉语译文在语序上也表现出较强的趋同性。

金学江等（2014）以表示判断性原因、理由的「からには」「以上」句为研究对象，借助《中日对译语料库》考察了其汉译倾向。经考察发现，「からには」「以上」的汉译形式中表示推断行因果文的“既然P,就Q”“既然P，Q”较为常见，远远多于无标形式“P，Q”。并且出现了一小部分条件性因果推断句“只要P，Q”。

而由关键词表可以看出，国内日语语料库研究以“日语、语料库、计算机”几个关键词为中心，不断向“日语教学、翻译、实证研究、偏误分析”等方面拓展。

结论与建议

本研究对1985-2019年间有关日语语料库的文献进行了CiteSpace可视化分析，研究结果显示：

１．在研究热点方面，语料库这一关键词贯穿始终，其次分别为：转折、日语教学、实证研究、对比研究、对译倾向、东北亚语言、日本学生、教学模式、学习者。

２．在阶段划分方面，2000以前的研究相对单一；2000-2010年间，日语教学逐渐开始成为语料库研究领域的热点，而在随后的几年中，又增加了实证研究、教学模式等几个热点；2010年以来，研究热点有了新的转向，转折、对比、对译等研究内容逐渐走入研究者的视线。

３．国内日语语料库研究以“日语、语料库、计算机”几个关键词为中心，不断向“日语教学、翻译、实证研究、偏误分析”等方面拓展。

由此可见，国内关于日语语料库的研究已取得了较为可观的成绩，但相较于西方语言，仍存在一定空缺，笔者认为可以从以下几方面进行补充。

第一，新的语料库的建设与现有语料库的完善。就现阶段而已，国内日语语料库相较西方语言语料库仍存在功能与规模上的不足。比如美国当代英语语料库（Corpus of Contemporary American English简称COCA）的词汇量就高达5.2亿，且词性标注简单明了，收录了近几十年来美国的英语口语、小说、流行杂志、报纸、学术期刊五大类型，反观日语语料库，无论是从规模、速度还是词性标注等方面都有待进一步完善。

第二，专门用途语料库的建设与开发。和大型通用语料库相比，适用于特定研究方向的专门用途语料库也值得进一步挖掘与开发。比如MICASE学术口语语料库就涵盖了各层次美国大学学生和教师、职员的课堂与非课堂言语交际，合计约200小时、170万词次，而反观现阶段的日语语料库，适用于特定研究方向（如医学、法律、专利、商务、口译等）的语料库仍

受到人力、物力方面的限制，有待进一步开发。

第三，语料库的多模态化及与其他学科的结合。就目前来说，国内日语语料库仍多以文字为载体，但随着信息化时代的发展，图像、色彩、声音等都可以成为交际符号。如何将负载了这类信息的语料库与传统的语言学、文学、辞书编纂等方面的研究相结合，是今后值得我们思考的问题。

第四，语料库研究的视野有待拓展。相较于国内日语语料库方面的研究，国际语料库语言学的研究内容更复杂，并越来越呈现出跨学科、多角度的特点。除了传统的搭配与词典编撰等研究热点，近年来也有不少学者涉及了自然语言处理、专门用途语言、隐喻、话语分析，并尝试与构式语法、认知语言学等相结合。这无疑也为我们提供了新的参考与借鉴。

当然，本研究仍存在一定局限性：本文仅以CNKI数据库中收录的文献为为数据来源，并不全面，今后还可将更多的国际译学期刊纳入考察范围。

参考文献

[1] 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（02）：242-253.

[2] 李杰，陈超美. citespace科技文本挖掘及可视化 [M]. 首都经济贸易大学出版社，2016.

[3] 段科慧. 基于HSK动态作文语料库的日本学生汉语习得偏误分析 [D]. 山东大学，2015.

[4] 冯峰. 高校日语教学与日语学习软件的利用 [J]. 日语学习与研究，2002（04）：57-60.

[5] 高宁. 论鲁迅直译观的语学基础 [J]. 山东社会科学，2013（10）：75-81+88.

[6] 黄大网，秦羿，徐赛颖. 专门用途英语语料库：挑战、理据与愿景 [J]. 宁波大学学报（人文科学版），2010，23（05）：48-52.

[７] 金学江，李光赫，吴世兰，林乐青．基于语料库的因果复句日汉对比研究 ── 以「からには」「以上」句为中心[J]．语文学刊（外语教育教学），2014（04）：13-14+39.

[８] 康佳萍．语料库语言学研究动向 ── 基于中外博士论文（2006－2015年）的对比分析[J]．河南工业大学学报（社会科学版），2018，14（06）：86-91.

[９] 路邈．汉日口译语料库的构建及其在翻译教学研究中的应用 [J]．日语学习与研究，2018（06）：52-58.

[10] 罗娜．中国基于慕课的外语教学热点与趋势研究 ── 以CiteSpace对CNKI文献科学计量学分析为例 [J]．佳木斯职业学院学报，2019（03）：109-111.

[11] 毛文伟．日语语料库建设的现状综述 [J]．日语学习与研究，2009（06）：42-47.

[12] 毛文伟．试论基于语料库的实证性研究中的信度问题 [J]．日语学习与研究，2005（01）：19-23.

[13] 施建军，徐一平．语料库与日语研究 [J]．日语学习与研究，2003（04）：7-11.

[14] 施建军．用计算机对日语进行研究的基础 ── 浅谈机器单词辞典和单词的自动切分 [J]．解放军外语学院学报，1991（03）：25-28+13.

[15] 宋红波，王雪利．近十年国内语料库语言学研究综述 [J]．山东外语教学，2013，34（03）：41-47.

[16] 王桢，毕晴．基于中介语语料库的日本学生“比”字句的偏误分析 [J]．教育现代化，2019，6（64）：79-80+111.

[17] 王昱．中国译学国际影响力可视化分析（2010-2019）[J]．上海翻译，2019（06）：29-36.

[18] 吴英杰．论计算机在日语教学中的应用 [J]．日语学习与研究，

2001（03）：48-50.

[19] 杨本明. 基于大数据的日语语料库的开发和教学应用研究 [J]. 戏剧之家，2018（34）：205-206.

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

[21] 杨柳. 国际语料库语言学研究热点与前沿的信息可视化分析 [J]. 知识管理论坛，2018，3（04）：208-224.

[22] 张政，王克非. 翻译研究：现状与未来 —— 记“首届翻译学国际前沿课题高端研讨会”[J]. 中国翻译，2017，38（02）：75-78.

[23] 赵红艳. 基于COCA语料库探讨近义名词的辨析 —— 以objective，aim，goal 和purpose为例 [J]. 英语广场，2017（09）：19-22.