NIHONGO NO
日語学習与研究
GAKUSHU TO KENKYU
2025年 第3期 总238号

语料库词典学视阈下汉日词典研编可行路径探索

山东大学 外国语学院 田润润
广东外语外贸大学 外国语言学及应用语言学研究中心 柳长青

[摘　要]　汉日词典长期处于编纂实践和理论研究极不平衡以及研编手段落后的局面，语料库词典学可为汉日词典实现现代化转型提供有效的理论启示与方法参照。本文在梳理中日两国语料库词典学重点关注课题的基础上，以汉语文化局限词“半边天”一词为例，构建了在语料库词典学、大语言模型等新兴理念和技术赋能词典学以及词典研编工作用户驱动转型的大背景下，汉日词典三位一体式研编路径，包括：基于词典内层语料库的词典质量评估、词典微观结构设计的理据探求以及探索开发汉日词典专用语料库和数据库的方法及工具；基于词典中间层语料库的汉日对比和汉日互译研究成果在词典编纂中的实际落地；基于词典外层语料库的用户日语习得特征、用户词典查检偏好及词典使用态度和需求的研究。本文可对汉日词典今后的研编工作提供适当参考。

[关键词]　汉日词典　语料库词典学　词典研编　用户驱动　大语言模型

引　言

语料库是词典研编的重要方法（胡文飞，2023）和必备资源（李宇明、王东海，2020），也是词典数据库建设和实现词典编纂现代化转型的基石（李宇明、庞洋，2006）。加强面向词典编纂的语料库的开发与建设（吕海春，2023）、推进词典学与语料库建设或技术的融合研究（于屏芳、董静文，2023）是数字时代词典研编的基础工作和必由之路。

汉日词典是指由中国本土编纂出版的以中国日语学习者为目标用户的双语学习型词典，如《简明汉日词典》（1985）、《新汉日辞典》（2011）等。从编纂出版国及目标用户来看，汉日词典要区别于在日本编纂或出版的，主要服务日本人汉语学习者的词典，如『東方中国語辞典』（2011）、『中日辞典』（2016）等；从功能来看，汉日词典是旨在服务用户进行汉译日、日语写作等产出性言语活动的编码型词典，同时对跨语言知识比较、中日文明交互也发挥着积极作用。

汉日词典在日语教育和中国文化的对日传播工作中发挥着举足轻重的作用，且其修订和编纂工作具有持续性，这便要求决定汉日词典质量和实用性的编纂理论和技术应受到相应重视。然而，汉日词典编纂长期处于“尚未进入科学、正确的轨道”[1]、缺乏理论依据[2]的局面，此外，作为非通用语种类词典，汉日词典还存在编纂出版技术急需现代化转型等问题[3]。语料库是数字时代完善词典编纂理论，提升编纂质量和效率的基础条件之一。但现实是，语料库与汉日词典的结合长久以来未引起足够重视。鉴于此，本文重点探讨语料库词典学视阈下汉日词典研编的可行路径，以期为今后相关工作提供一定参考。

1　文献综述

我国日语学界早已关注到语料库对词典研编的价值（徐一平，2003；施建军、徐一平，2003）。施

建军曾围绕该课题开展过多场学术讲座，内容涉及基于语料库的词典编纂可行性分析、基于平行语料库的译词获取、中日同形词互译性[4]、传统编纂方法的局限性、日汉词典中的翻译错误等，并指出语料库辅助日语类词典的研编工作尚处于起步阶段，有很多课题有待深入探索[5]。尽管如此，目前相关研究成果数量仍十分有限，且主要围绕日语单语词典（戴宝玉，2011；刘艳伟、刘玉琴，2015）或日汉词典（李飞菲，2023）展开，而专注于汉日词典的研究几近空白。语料库词典学是系统探讨语料库技术与词典学相融合的学科，其发展有助于增强词典研究的定量分析，提升研究结果的信度和效度[6]，中日两国在该领域已开展系列研究，相关成果能为汉日词典研编带来一定启发性。另一方面，随着人工智能技术的突飞猛进，基于深度学习的大语言模型（Large Language Models，简称LLMs）正在为词典研编带来新的技术变革（de Schryver, 2023；Lew, 2023；许家金等，2024），LLMs赋能词典编纂是语料库驱动词典编纂的升级版，探索LLMs赋能汉日词典研编是生成式人工智能时代语料库词典学的必然议题。此外，当前我国词典研编整体呈现出由“编者中心”向“用户中心”转型（杨玉玲、李宇明，2023），并最终实现编者视角和用户视角相融合（李宇明、王东海，2020）的趋势，用户的词典使用数据和需求在词典研编中的应用价值正在被挖掘。鉴于此，本文主要围绕三个问题展开：（1）中日两国在语料库词典学视阈下主要关注的研究课题是什么？（2）语料库词典学视阈下开展汉日词典研编工作的可行路径有哪些？（3）LLMs在汉日词典编纂中的应用潜力如何？结语部分将在人工智能时代以及词典研编工作用户驱动转型的大背景下，构建生成式语料库赋能汉日词典三位一体式的研编路径。

2 中日两国在语料库词典学视阈下的主要研究课题

语料库词典学由语料库语言学衍生而来，在国内词典学界已被赋予术语属性（语言学名词审定委员会，2011）和学科属性（王馥芳，2004；章宜华，2013），用以专门探讨“基于语料库的词典学理论研究和编纂技术”[7]。相比于中国，日本多以「コーパスに基づく辞書～」[8]「コーパスを利用した辞典～」[9]「コーパスと辞書」[10]等表述来开展相关研究。鉴于语料库词典学能更系统地阐释语料库技术和词典学间的交互关系，本文使用该术语来统述中日两国相关研究。

中日两国在语料库词典学视阈下产出了颇多研究成果。中国具有代表性的是李德俊（2015）的专著《语料库词典学：理论与方法探索》，该书从术语界定、发展源流、学科正名、理论依据、研究内容、研究方法、研究案例等多方面阐释了语料库词典学的学科本质，论述了语料库词典学视阈下词典研编普遍关注的问题。该研究将语料库词典学的研究内容分为三大方面：基于语料库的词典编纂技术研究、基于语料库的词典理论研究以及词典语料库的本体研究，并指出语料库赋能词典编纂自动化是语料库词典学的终极目标，但理论研究不足和技术限制使得该目标的实现任重道远。尽管该著作侧重从宏观视角综述语料库词典学的相关议题，但著者仍有意凸显了该领域的重点和前沿课题，即短语学、搭配结构和翻译对等等内容。

由传康晴和荻野纲男（2019）主编的论文集『コーパスと辞書』是日本语料库词典学领域的集大成者之一。该论文集的主要目的是从词典研编角度评价「現代日本語書き言葉均衡コーパス（BCCWJ）」（以下简称BCCWJ）的应用潜力，重点考察了语料库在日语国语词典研编中的应用价值，内容包括语料库技术与词典研编结合的个案研究、日语特定功能表达（如复合辞）的计算机处理方法与词典开发、基于语料库的日语搭配词典编纂可行性探索、基于语料库的日语国语词典描述检验等。

总体来看，尽管语料库在词典研编中的表现仍受限于语料库的质量、规模等因素，但由于其在储存、检索、标注及统计等方面具备人工操作无法企及的优势，且能提供真实文本中动态的语义和丰富的例句，因而在中日两国的词典研编工作中表现出广泛且强大的应用性。然而，尽管与词典学相关的大部分课题都可尝试从语料库词典学视

阈下开展，但在梳理文献的过程中，仍能发现中日双方都尤其侧重了词典中的搭配问题，或展示如何用语料库方法进行搭配的词典描述，或探索如何基于语料库开发搭配词典。

3 语料库词典学视阈下汉日词典研编可行路径

3.1 理论研究可行路径

笔者认为，可从汉日词典直面的编纂难题和语料库词典学的主要学术问题出发，来探索汉日词典理论研究的可行路径。

翻译是双语词典的核心问题，双语词典的基本目的之一是给出一种语言的词汇单位在另一种语言词汇单位中意义相等的对应词（笔者译）[11]。汉日词典的编纂难题主要集中在翻译上，即词目词翻译（译义）和例句翻译（译例）。彭林宜（1982）在论述汉日词典的编纂问题时指出，处理好原语与译语的相互关系，是保证双语词典质量和研究双语词典编纂理论与方法的关键[12]。翻译策略、翻译对等是双语词典翻译研究的重点议题（万江波，2008；李明一、周红红，2011），尽管汉日词典的过往研究对这些议题都给予了一定关注（赵福堂，1982；王永全，2010），但以总结编纂经验为主的研究内容和基于个人经验判断的研究方法使得汉日词典的翻译问题并未得到理论层面的升华。通论性双语词典研究指出，双语词典主要提供源语词目的对应词，而对应词根据翻译策略可分为翻译对应词（整体对应）和解释性对应词（部分对应），二者的区别在于前者在进行翻译实践时可直接插入到译文中，而后者只是对源语词目的解释，并指出编码型双语词典应尽可能列出翻译对应词[13]。由此可认为，汉日词典作为辅助汉译日、日语写作等产出性言语活动的工具书，其翻译策略应以整体对应为主。当然，具体应采取何种翻译策略需专文论述，在此不赘述。

另一方面，如前文所述，语料库词典学重点关注的研究内容沿袭了语料库语言学对短语学及搭配结构的重视。Sinclair基于语料库提出的词汇模型即短语学，被认为是语料库词典学的理论源泉之一（李德俊，2015）。其核心观点：意义是通过上下文共同创造和解释的，搭配是意义生成的关键，自语料库词典学的首部产物Collins COBUILD English Dictionary（1987）开始，便持续影响词典编纂（Moon, 2008）。而其工作模型，即扩展意义单位模型（Model of Extended Unit of Meaning，简称EUM）（Sinclair, 1996）通过描述词项在词与词（搭配）、词与语法（类联接）、型式与意义（语义趋向）、词或语法与功能（语义韵）这四个层面上的共选关系（co-selection），实现了和语境高度关联的词项意义阐释，逐渐成为语料库词典学视阈下词典微观结构，尤其是搭配信息的主要分析框架之一（李德俊，2015；胡文飞，2023）。此外，由短语学衍生出的对比短语学旨在发现和描述对应或对等的跨语言意义单位，在翻译对等研究中应用性显著（陆军、卫乃兴，2012），继而被应用于双语词典的翻译研究中（胡文飞、刘亚兰，2023）。

基于上述分析，笔者进一步总结出语料库词典学视阈下汉日词典理论研究的可行路径之一是：通过对比短语学研究汉日词典中的翻译对应问题。下文以EUM为分析框架，以“半边天”一词在《新汉日辞典》（2011）（以下简称《新汉日》）中的描述为例，展示该路径的具体操作过程。选择“半边天”的原因是：该词是兼具本义和比喻性转义的多义词，同时也是汉语文化局限词，在汉译日过程中存在一定难度，因而易于说明汉日词典中存在的翻译问题。而选择《新汉日》，是因为该词典依托国家出版基金资助项目和“十二五”国家重点图书出版规划项目，除了具有较强权威性，也具有较强的时效性。可比语料库分别选取，目前中日两国具有代表性的综合语料库，即北京语言大学BCC语料库（下文简称BCC）和BCCWJ[14]；中日对译数据资源选取目前可访问的《中日对译语料库》、『日中対訳EGAコーパスVer.1』、“人民中国杂志社”官网。

以下是半边天”在《新汉日》中的描述：

（1）天の半分。この世の半分。

妇女能顶～

婦人はこの社会の半分を支える。

（2）婦人を指す。

本文重点分析“半边天”转义用法的两个日译，即「この世の半分」和「婦人を指す」。相较而言，「この世の半分」是采用了整体对应策略的翻译对应词组，而后者以提示词「指す」引出了与“半边天”具有相似性的「婦人」，为解释性对应表达。

以“半边天”「この世の半分」「婦人」为节点词，在相应的语料库中检索后发现，“半边天”和「婦人」均得到了一定规模的索引行，但「この世の半分」未检索出结果，这说明「この世の半分」作为一个词组在日语语境下的生成率很低。接着，分别从BCC和BCCWJ中随机提取“半边天”和「婦人」的100行索引，并启动EUM检验二者的对应性，结果如表1所示，另，EUM各要素的界定可参考相关专门研究（濮建忠，2020），在此不赘述。

基于搭配和类联接可归纳出“半边天”在具体语境中主要呈现“女性支持某个行业领域”这一语义趋向，从而推导出其具有积极语义韵。而「婦人」一词的主导语义趋向是“女性相关行业或组织名称”，因此更倾向于中性语义韵。语义韵是确立扩展意义单位边界的关键性因素，被认为是检验双语对应意义单位的“最重要的制约因素与判断标准”[15]，由于语义韵不一致，可判定“半边天”和「婦人」无对应关系。

实际上，“半边天”的主导语义趋向和积极语义韵具有深刻的文化渊源。以“半边天”一词强调女性为社会主义建设贡献力量或借喻女性的用法被证实最早由毛泽东提出（唐闻晓、陈旭楠，2023），其标志着中国女性地位在各领域中发生了重大改变。而「婦人」无论在日语国语辞典，如『岩波国語辞典 第八版』（2019），还是在日语语感辞典，如『日本語語感の辞典』（2010）中均显示因语感过时，已逐渐被「女性」所取代。另外，「婦人」在当下的日语语境中还被解读出多重消极语感[16]，从本次考察的索引行中虽未观察出消极语感的统计学意义，但仍能发现「婦人」表现出吸引有年长语义的词或词组的趋向，如「貞潔とはほど遠い」「閉経期の」「子供も何人か育てた」「中年の」等，由此可进一步证明不能用「婦人」解释“半边天”。

《新汉日》中译词的厘定方法是由编者个人判断还是依托语料库筛选尚且不得而知，但其对“半边天”的译法至少反映出两大问题，一是翻译对译词在日语语境下的低生成率使得其对实际编码活动的指导力不强；二是以语义韵不一致的词解释源语词目，易造成编码时的误用。下页表2反映了中日对译数据资源中“半边天”的日语对应情况。

观察表2发现，与本文重点考察的转义用法相关的有效索引行只有3、5两项，其中的“半边天”的语义虽与「この世の半分」最为接近，但实际译文中并未使用该对译形式，而是采用了意译策略，即译成「男性並みの経済力」和「圧倒的な実力」，显然这种译法能更突出“半边天”的积极语义韵。“真正贴切的对应词大多存在于好的翻译材料之中，值得大力搜寻”[17]，平行语料库在双语词典研究中最凸显的价值就是提供真实语境下的对应关系，但从具体检索来看，即便综合使用了3个对译资源，检索结果仍十分有限，且单个数据资源无法详尽“半边天”的全部义项，必须跨越多个平台才能实现义项的全部覆盖。这也表明，受语料库的规模、语料来源等因素制约，单个语料库在提取

表1 基于EUM的“半边天”和「婦人」对应性检验

EUM构成要素	汉语词目			日语解释性对应	
节点词	半边天			婦人	
搭配	动词（v）		“顶～”（28例）；撑～（25例）	复合词（78例）	既定术语：「婦人科」
	形容词性修饰（a）	所属限定	“经济的”“人类社会的”等（33例）		组织、机构等名称：「浜中町農協婦人部」
		特征评价	“当之无愧的”（6例）		称呼词：「ご婦人」
类联接	vN vaN			n+N N+n	
语义趋向	女性支持某个行业领域			女性相关行业或组织名称	
语义韵	积极			中性	

资料来源：作者依据分别从BCC和BCCWJ中抽取的索引行制作。

表2 “半边天”在中日对译数据资源中的日译情况

序号	中日对译资源	原 文	译 文
1	《中日对译语料库》	黑云滚似的已遮黑半边天	まっ黒な雲ははやくも空のなかばにひろがっていた。
2	『日中対訳EGAコーパスVer.1』	调查：上海家庭多是“女皇当权”丈夫仅是半边天	上海の多くの家庭は「かかあ天下」最新調査
3		女性已经撑起了半边天	中国の女性も男性並みの経済力を有するようになっている。
4	“人民中国杂志社官方”	{たいかん}大寒や{はんてん}半天の{へき}碧{れいろう}玲瓏と	玲珑碧透半边天
5		“她力量”崛起，女性撑起消费市场大半边天	「女性のパワー」の台頭によって、女性は消費市場で圧倒的な実力を示している

资料来源：作者依据上述三大中日对译数据资源中“半边天”相关的中日对译例句制作。

译词方面难以具备自足性。

以上基于对比短语学的主要工作模型EUM，举例分析了《新汉日》中翻译策略和翻译对等方面存在的问题，从《新汉日》的“前言”来看，尽管该词典在翻译优化方面下了一定功夫，但深入分析后发现仍有遗留问题。其一是体现在汉语源语词目和日语目的语的态度意义、即语义韵的偏差上，语义韵分析在词典研究中备受重视（王泽鹏、张燕春，2005），胡文飞、刘亚兰（2023）从三个方面概括了语义韵分析在汉英词典研究中的应用价值，包括：有利于了解词典的编纂现状和表征效果、革新研究视角以及优化产出功能等[18]。汉日词典研究今后也可参考此路径，有效整合使用现存的中日可比语料库和平行语料库，充分挖掘对比短语学在汉日词典译义模式及译义规范研究中的应用价值。当然，本文仅列举了语料库词典学视阈下汉日词典中有关翻译的研究路径，实际上，双语词典微观结构中的其他要素也可在语料库词典学视阈下展开。如义项颗粒度的切分、义位的归纳、例句的提取以及词目文体信息标注等问题均可借力语料库的大规模性进行理据探索。

其二，不难发现中日同形词是语义韵偏差产生的一大诱因，体现了汉日词典区别于其他汉外词典的特有问题。目前依托语料库的中日同形词研究已积累了一定规模的研究成果，如在词汇对比研究领域，施建军、洪洁（2013）利用大规模语料库对中日同形词意义用法进行了阶梯式考察[19]，施建军等（2023）验证了词向量技术在中日通用汉字词计量研究中的应用潜力[20]；在汉日互译研究领域，童富智、修刚（2020）使用EUM对中央文献日译研究进行了方法论创新[21]。上述研究均摆脱了词典的桎梏，力求通过大规模语料库对中日同形词意义用法进行更为精细的区分，但研究的出发点和落脚点基本以词汇学或翻译学为主，对词典的涉及多数只出现在研究意义的阐释部分。译学于双语词典的重要性在前文中已有论述，而语言对比是双语词典学的主要研究内容之一（陈楚祥，1992），被视作双语词典学的“灵魂”[22]。在汉日语言对比研究领域，除中日同形词外，音系、句法、语用、文化背景等各个层面均积累了一定的研究成果，但相关研究于汉日词典的价值尚未得到切实关注。在接下来的研究中，可将汉日语言对比和汉日互译这两大领域中涉及词典的研究成果进行整合，构建指导汉日词典研究的文献语料库，一方面可为汉日词典研究提供语言专家视角的理据，另一方面，也可实现理论研究对具体实践的反哺价值。

此外，由于汉日词典的主要受众是中国日语学习者，因此汉日词典的收词规模如何确定、微观结构如何设计、条目信息量如何把控等问题仅由词典编纂者和研究者衡量还远远不够，对反映中国日语学习者，即汉日词典用户的语言水平和习得能力、词典查询行为以及词典需求的语料库分析也应受到同等重视。

3.2 编纂实践可行路径

语料库词典学视阈下汉日词典编纂的必要条件是要有相应的中日平行语料库，确切而言，要有专用于汉日词典编纂的词典语料库。词典语料库除了要有一般语料库的基本特征外，还会根据词典的内容、功能及规模呈现自身特征（李德俊，2015）。例如，内容方面，要注意通用词典和专科词典对语料文本类别的不同需求；功能方面，要根据编码型

和解码型的差异，调整对译语料的比例，一般认为外译汉方向的语料对汉日、汉英等汉外类词典的编纂更有价值；而在规模方面，要考虑语料规模可否覆盖所编词典收录的词目等。

现有的中日平行语料库根据构建目的，大致分两类，一类用于语言研究，尤其是汉日对比研究；另一类以提升机器翻译精准度为目的，表3列出其中几例以评述现存语料库在汉日词典编纂中的可应用性。

《中日对译语料库》是目前已公开过的、规模较大的中日双向平行语料库。其构建目的以通用性为宗旨，即以语言研究为主，同时服务于翻译、文学以及文化研究，但由于语料选材主要是文学作品，所以平衡性欠缺。在该语料库中检索“半边天”，仅检索到本义义项，而粗略观察BCC，发现转义义项多出现在新闻语料中，可见语料选材直接决定语料库能否覆盖词项的所有义项。此外，该语料库中汉译日文本和日译汉文本的规模比例也不能满足汉日词典的需求，因而无法单独用来编纂汉日词典。

『日中対訳EGAコーパス』力求避免使用陈旧语料，并尽量维持语料来源的均衡性，以区别现有的中日平行语料库（宫本华瑠，2023）。该语料库尚处于构建阶段，截至目前仅有559万字的规模，且汉语语料中有相当一部分来自中国台湾，源自中国大陆的语料主要是“人民网”的新闻文本，此外，现阶段语料构成只有汉译日文本，因此，无论是在语料的规模上、还是构成上都远不能满足汉日词典对语料库的需求。

『JPO中日対訳コーパス』是日本专利局（JPO）根据中日两国专利公报的明细单等语料开发的中日对译语料库，基于该语料库，构建了包含约20万条记录的词典数据库，后将其与日本专利局提供的词典数据合并，创建了『JPO中日対訳辞書』[23]。『中国語·日本語の対訳コーパス』包含400万件以上的中日对译语料，基于该语料库开发出了提高翻译精准度的科技术语词典[24]。这两个语料库虽然都衍生出了词典产品，但主要是面向自然语言处理的机用词典。

通过对现有语料库的分析发现，依托语料库的专科词典的编纂似乎较易实现，但编纂的通用类词典仍任重道远。对于今后汉日词典的编纂路径，首先，要加强对语言资源的整合使用。资源整合是数字时代词典编纂的有效路径，依托语料库编纂的“JUZI汉语”对多语言资源的整合使用充分证明了该路径的可行性（章宜华，2024）。此外，在以用户视角为主导的词典编纂理念的引领下，学习者语言资源在词典编纂中的应用价值也被挖掘（宋立文，2021）。鉴于汉日词典的学习性，在进行资源整合时，除了中日对译资源外，还应考虑对汉日词典的主要用户，即中国日语学习者语料库的整合，该语料库由其内容大体可分三类，其一是日语习得语料库，如中国日语学习者语料库（CJLC）、「北京日本語学習者縦断コーパス」（B-JAS）等。日语习得语料库记录了学习者在使用日语时常见的词汇或语法等错误及习得难点，通过分析这些错误及难点，可在词条编写时添加更多的语言使用提示和注意事项，避免学习者误用并帮助其突破难点；其二是词典查检行为语料库，在用户词典使用数据备受重视的背景下，反映日语学习者词典检索行为的相关调查，如，由日本国立国语研究所发布的『学習者の辞書資源使用の実態調査——語彙検索行動の解明』[25]及相关论作（石黑圭等，2023；石黑圭等，2024）在汉日词典编纂中的应用价值也有待挖掘；其三是词典评价语料库，用户的词典评价语料汇集了用户对词典的需求痛点，有助于完善词典编纂的评价体系（陈贤德、杨玉玲，2025）。

表3 现存中日平行语料库例举

名 称	开发者	文本类别	构建目的	对译语料构成
《中日对译语料库》	北京日本学研究中心	主要为文学作品	语言研究	中日双向
『日中対訳EGAコーパス』	一般社团法人EGA	杂志、报纸	汉日语言对比研究	中译日
『JPO中日対訳コーパス』	日本国特許庁（JPO）	专利公报	提升机器翻译精准度	中日双向
『中国語·日本語の対訳コーパス』	国立研究開発法人科学技術振興機構（JST）等	科技论文等	提升机器翻译精准度	中日双向

资料来源：作者依据表中所示几大语料库的基本信息制作。

其次，设计和开发汉日词典编纂专用语料库。开发词典专用语料库是当前我国词典学界普遍面临的难题（齐红飞、王东海，2023），专用语料库不仅要顾及语料的规模和文体种类等语料库自身问题，同时更要考虑所编词典的释义深度、例句难度和丰富度以及搭配类型等词典微观信息。因此，探索高效的词典语料库的开发技术和工具是一大课题。

最后，要开发汉日词典数据库。词典数据库是“按词典的宏观结构、微观结构、中观结构和索引结构等来分解、组织和存储词典知识信息的库”[26]，而语料库只是对语言材料最基本的加工，其目的是形成精加工的、结构化的多用途数据系统。随着词典数字化进程的深入，数字词典、尤其是手机终端的APP词典正在取代纸质词典成为日语类词典用户的首选（石黑圭等，2023）。然而，目前我国日语类APP词典存在汉日词典数据资源紧缺的情况。例如，广受用户青睐的“MOJi辞書”中没有内置的汉日词典数据库，而有内置数据库的APP词典仅实现了纸质词典的直接电子化，如将《新汉日》作为内置数据库的“日语大词典”，但这种转变并非是现代意义上的数字词典转型[27]。因此，未来还应在构建汉日词典专用语料库的基础上加强汉日词典数据库的开发。

4 LLMs在汉日词典编纂中的应用性

随着人工智能技术的突飞猛进，以数据驱动和知识驱动为导向的词典研编工作已成大势所趋，尤其是近两年来，基于深度学习的LLMs在词典编纂中的应用开始广受关注。LLMs可为语料库词典学带来技术层面的升级，而语料库词典学的原理和方法可对LLMs生成的词典产品进行质量监督，二者相辅相成互促词典质量的提升和编纂技术的革新。de Schryver（2023）指出，LLMs由于比以往任何（语言）技术都更以语言为导向，因而在词典学界被寄予厚望[28]。目前国内国际学界都在积极探索LLMs在词典编纂中的应用前景，特别是近两年的词典学研讨会都集中或重点关注了这一新兴领域。如国内的“中国辞书学会第十四届会员代表大会暨学术研讨会”（2023）[29]、“第三届计算词典学研讨会”（2023）[30]、“第九届学习词典与二语教学国际研讨会”（2024）[31]以及日本「語彙·辞書研究会」第64届研讨会[32]都将LLMs赋能词典编纂列为会议重点议题。国际方面，de Schryver & Joffe在东京以“词典编纂的终结，欢迎来到机器时代”[33]为题举办的系列讲座拉开了LLMs时代词典研编的序幕。此外，亚洲辞书学会（ASIALEX, 2023）[34]、21世纪的电子词典编纂学会（eLex, 2023）[35]也集中探讨了该议题。

数字时代的词典编纂需要大规模的自然语料、自动化的编纂技术以及高效的人机协作模式，LLMs经过大规模自然语料的预训练，不仅拥有丰富的词汇库和语言知识，同时具备理解和生成自然语言文本的能力以及类似于人类交流的对话系统，这些因素于词典编纂而言具有得天独厚的优势。然而，尽管LLMs被证明在词典编纂方面已表现出一定的类人性（Lew, 2023），但“幻觉（hallucination）”的多发性、输入和输出结果的不确定性、训练语料库的英语主体性（指GPT家族等）以及对熟练撰写引导LLMs完成大规模编纂工作通用提示词的高要求（de Schryver, 2023）等问题是当前LLMs赋能词典编纂主要面临的挑战。尽管如此，词典专家仍承认LLMs赋能词典编纂已成大势所趋，认为其已具备成熟词典编纂者的潜质，并号召加强探索LLMs技术在词典编纂中的应用，提议通过微调GPT等通用大模型或开发专门用于词典编纂的大模型（LexGPT）等手段以提高编纂效率。

当前，通过对比LLMs（尤指GPT家族）自动生成的词典条目与现存词典词条的质量和特点是衡量LLMs应用潜力的前沿课题（Phoodai & Rikk, 2023；Lew, 2023；张永伟、刘婷，2024）。为评估现阶段LLMs在汉日词典词条自动生成方面的应用潜力，我们在GPT-4中设置提示词引导其对“半边天”生成词典条目。由于此次调查的主要目的是尽量挖掘GPT-4自身的汉日词典编纂理念，因此在提示词的设置上侧重诱导GPT-4积极进行自我反馈，尽量控制人工意识的干预。在使用LLMs生成词典条目时，提示词与生成的结果之间并非

一步到位，多轮调整和优化是必然。本次调查主要通过三轮对话引导词条的自动生成。各轮设置的提示词及设置理由如下所示：

第一轮

提示词：“请对汉日词典进行界定”

理由：把握GPT-4对汉日词典的类型意识

第二轮

提示词：“请你充当一名专业的汉日词典编纂者，举例说明理想的汉日词典如何解释一个词”

理由：把握GPT-4所认为的理想的汉日词典表征形式

第三轮：请按照以上格式，并结合联网搜索，对汉语“半边天”一词进行释义。

理由：正式命令GPT-4自动生成词典条目

基于上述对话生成的“半边天”词条如下所示：

词条：半边天（pinyin：bàn biān tiān）

1.日语对应词

半分の空（はんぶんのそら，hanbun no sora）——字面意思，指“天空的一半”。

女性の社会的地位（じょせいのしゃかいてきちい，josei no shakaiteki chii）——比喻用法，指女性在社会中的重要地位。

2.词义解释

半边天：

字面意义：指“天空的一半”。

日语解释：空の半分を意味する。

比喻意义：指女性在社会中的重要地位，尤其是在提倡男女平等的社会中，表示女性在社会发展中与男性同样重要。

日语解释：特に男女平等を促進する社会で、女性が社会発展において男性と同等に重要な役割を果たすことを意味する。

3.常用搭配

女性が半分の空を支える（じょせいがはんぶんのそらをささえる，josei ga hanbun no sora wo sasaeru）——表示女性在社会或家庭中的重要支撑作用。

半分の空を守る（はんぶんのそらをまもる，hanbun no sora wo mamoru）——强调某人或群体在某一方面的重要作用。

4.例句

中文：“在新社会，妇女撑起了半边天。”

日语：新しい社会では、女性が半分の空を支えています。

中文：“她们在各个领域都能撑起半边天。”

日语：彼女たちは様々な分野で半分の空を支えることができます。

5.语法与文化注释

语法：在中文中，“半边天”可以表示物理意义上的天空一半，也可以用于比喻女性的社会地位。日语中通常直接翻译为“半分の空”时，多指字面意义；若强调女性的社会作用，则使用比喻意义的表达。

文化注释：“半边天”一词在中国文化中深具象征意义，特别是在提倡男女平等的背景下，用来突出女性在社会各个方面的重要性。这一表达方式在日语中也有类似用法，但表达时可能更加直接。

6.派生词与相关词

撑起半边天（しっかりと半分の空を支える，shikkari to hanbun no sora wo sasaeru）——强调女性或特定群体在某一领域的重要支撑作用。

红透半边天（真っ赤に染まる空，makkani somaru sora）——形容某人在某一领域中非常出色，取得了巨大的成功。

此词条参考了多种来源，并结合中日文化背景进行了详尽解释和例句设计（ChinesePod, Wikt-ionary, Yandex Translate）。

GPT-4在第二轮以“家”为例，例举了理想汉日词典词目的微观表征，第三轮参考第二轮的结果，自动生成了“半边天”的词典条目信息。初步浏览可发现，GPT-4生成的词条基本上满足词典条例且具有一定可读性，尤其是对“半边天”本义和转义日语对应词的翻译策略均采取了整体对应，且在表示“半边天”的女性含义时，采用了日语语境下逐渐取代「婦人」的「女性」一词，这一点要优于《新汉日》。此外，对常用搭配、文化注释等信息的纳入也较好地体现了汉日词典的学习性，体现了超大语言资源智能体在词条自动生成中强大

的应用潜力。但其缺陷也较为突出，如，“2. 词义解释”的混入，是混淆了双语词典和双解词典的概念，在汉日词典编纂中属于画蛇添足，而产生的原因主要在于GPT-4对汉日词典缺乏清晰的类型界定。除此以外，例句的配置、语法与文化注释、搭配词与派生词等条目信息都存在形式单一，内容重复的情况。另外，为改善LLMs生成词典条目时底层语料未知性这一根本缺陷（Rundell, 2023），我们在提示词中附加了指令：“请结合联网搜索”，但GPT-4所罗列的三个信息源都不是理想意义上的中日对译资源。

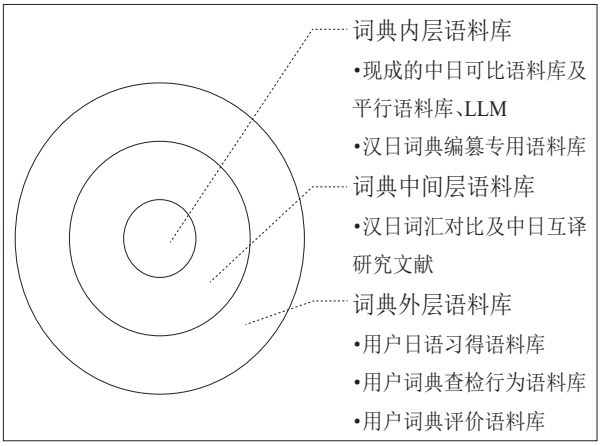
由上述分析可知，快速、批量、智能化促使LLMs为汉日词典编纂的提质增效开拓了空间，但词典自动编纂的最终产品并非只是生成符合上下文的类词典文本，而是切实服务用户语言习得的更精致的结构化了的词典文本。然而，当前LLMs在功能语言能力方面，即策略性地使用生成的文本实现各种目的的能力仍不尽人意（Mahowald et al, 2024）。词典编纂具有明确的目的导向，即在宏观上要确保权威性和规范性，在微观上要致力于将复杂的词汇、语义、语法及语用信息分类、组织成结构化的且易于访问的形式（Lew, 2023），这些问题需要词典编纂者同词典用户、开发方及出版方等进行多方协作后方可实现。因此，人类智慧在解读自动生成的语言数据中仍扮演着重要角色（Rundell, 2023）。

此外，LLMs能否胜任汉日词典的条目生成工作，与熟练编写提示词的能力以及所依托的底层语料的质量密切相关，提示词的编写需要根据汉日词典的类型特征，并结合潜在的用户类型、用户日语习得情况和词典需求等要素进行反复试错与调整，而高质量的语料很大程度上依赖大规模的中日平行语料库。因此，今后的汉日词典研编工作中，首先，要在结合多方资源的前提下，在反复试验和调整LLMs生成词条的过程中摸索出能高效率引导LLMs生成符合预期的提示词撰写方法；其次，继续构建高质量的中日平行语料库，平行语料库可帮助语言模型更好地理解源语言与目标语言之间的对应关系，从而提高翻译的准确性（Lin et al, 2024），因此，高质量的中日平行语料库的开发工作始终是技术革新背景下汉日词典编纂。此外，要开发汉日词典专用语言模型。目前有关LLMs赋能词典编纂的研究主要围绕单语词典展开，因此，所关注的问题主要是模型的理解力、生成力以及人机对话能力，而双语词典还涉及到LLMs的机器翻译能力，因此，现存的通用大模型能否有效赋能汉日词典中的翻译工作，是否必须开发领域性更强的汉日词典研编专用语言模型也是一大课题。

最后，必须承认LLMs正在为传统语料库词典学带来大程度革新，体现在：理论研究方面，LLMs强大的推理和生成能力可帮助研究者在语料库分析过程中进行自动化的数据挖掘和规律发现，从而深度优化词典理论；技术方面，LLMs可根据提示词，自动生成或者优化词条，乃至完成词典初稿的制作；而在词典语料库建设上，LLMs的底层语料库不仅可以规模性的、多领域性的扩充现存语料库，也可在语料标注方面实现智能化。但词典编纂并非只是编纂者独自面对语料和技术，还要充分考虑语言研究专家的观点以及词典用户的语言能力、词典使用习惯及需求。唯有通过多维度的综合考量，方可确保汉日词典产品的学习性、实用性、权威性及专业性。

5 结　语

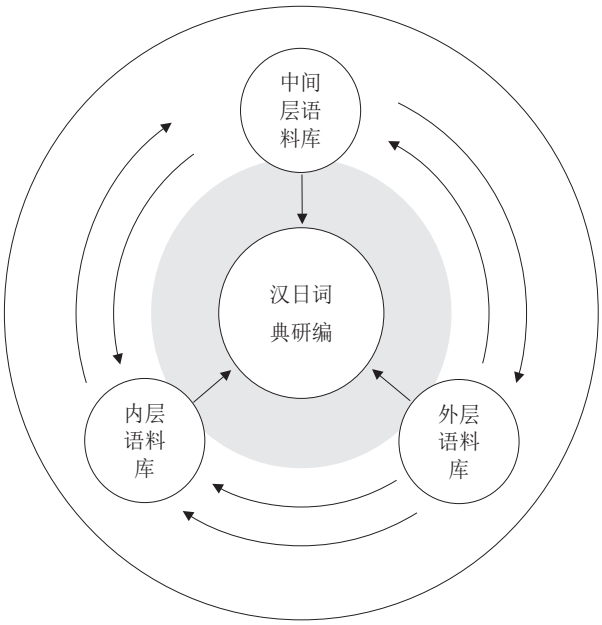
以上从中日两国在语料库词典学视阈下主要关注的研究课题入手，探讨了用户驱动词典研编转型的背景下，基于语料库词典学的汉日词典研编路径，在此基础上评估了GPT-4在汉日词典词条生成中的应用潜力。我们发现，在词典研编用户驱动转型以及LLMs带来海量语料和词条自动生成技术的背景下，汉日词典研编语料库不再局限于供词典研究使用的现成语料库和制作汉日词典文本的核心底层语料库，同时还应融合相关涉汉日词典或日语学习词典的文献语料库以及中国日语学习者语料库，我们将这三大语料库分别称为词典内层语料库、词典中间层语料库以及词典外层语料库，下页图1展示了三大语料库间的关系：



资料来源：作者依据赋能汉日词典研编的三大语料库间的关系绘制。

图1　赋能汉日词典研编的三大语料库

其中，词典内层语料库是汉日词典研编的核心语料库，基于该语料库可开展的课题有：围绕现成的中日可比语料库、平行语料库以及LLMs，并结合短语学和对比短语学等语料库词典学的主要理据，探讨汉日词典的译义模式和规范、搭配模式，基于大规模语料对现存汉日词典的义项颗粒度切分、义位归纳、义项排列和例句择取进行定量性的质量评估和理据探寻，而围绕汉日词典编纂专用语料库的研究内容主要是语料库及数据库的开发理据、技术、方法和工具的探索，微调和优化LLMs的探索。通过以上研究，一方面可实现对汉日词典微观表征的定量优化，同时还可促成汉日词典编纂自动化、智能化的早日实现。词典中间层语料库狭义上是指基于大规模语料库的汉日对比和汉日互译的研究文献，广义上包括一切有关这两大领域的研究以及日语类词典相关文献，其目的是从语言专家的观点中提取有助于汉日词典研编的理据，以确保汉日词典的专业性和权威性。此外，该层语料库还发挥着连接词典内外层语料库的中介作用，即将学习者语料库的相关情况反馈于核心语料库的开发与建设，目前该方面的研究较为薄弱，亟待加强。词典外层语料库，即学习者语料库，主要可开展的研究内容有：用户词汇语法误用、学习难点及词条难度等级的研究；用户词典查检目的、过程、偏好、困难、查得率的研究；用户对词典的态度与需求的研究，这些研究是汉日词典学习性、针对性、实用性的根本保证，其目的是从学习者视角优化词典微观表征。就此，我们构建出用户驱动和技术革新背景下，基于语料库词典学汉日词典三位一体式研编路径，如图2所示：



资料来源：作者依据赋能汉日词典研编的三大语料库间的关系绘制。

图2　语料库视阈下汉日词典三位一体式研编路径

该路径通过对词典研究语料库的内涵充实与外延拓展，一方面可为汉日词典研编提供更切实的操作路径，同时也与当下词典研编工作对语言资源整合的需求以及与词典融媒化的发展理念相吻合，从而进一步推动汉日词典的现代化转型。诚然，本文仅以“半边天”一词为例，站在传统语料库词典学与LLMs赋能词典研编的交界处，思考了当前汉日词典研编的可行路径，对LLMs在汉日词典研编工作中的理论参照和应用潜力、存在局限和解决方案等问题仍有待进一步探索。但笔者认为，坚持守正创新是生成高质量词典文本的重要保证。无论是传统语料库词典学赋能词典研编，还是LLMs为词典学带来的创新空间，首先需明确词典文本内容的准确性、权威性、规范性以及用户友好性始终是词典研编的主要出发点与落脚点，而充分利用传统语料库词典学与LLMs的互

助关系，可为生成高质量的词典文本提供保障。

当前汉日词典仍处于理论研究相对薄弱，编纂操作空间相对受限的困局。愿本文抛砖引玉，与更多同行共同探讨技术革新背景下汉日词典及日语类学习词典研编方案。同时，笔者将通过甄选具有一定规模的代表性词例，在不断试错的过程中进一步完善该路径，争取早日为汉日词典编纂方提供更具操作性的编纂方案。

注

[1] 王永全.我国汉日词典编写及出版现状、存在问题及新世纪的展望[C]//张柏然，魏向清，主编，双语词典学论集，南京：江苏教育出版社，2001：120.
[2] 王永全.汉日词典与中日辞典[J].辞书研究，2010（6）：92.
[3] 魏向清，耿云冬，王东波.中国外语类辞书编纂出版30年（1978～2008）：回顾与反思 [M].上海：上海辞书出版社，2011：86.
[4] 郭家唐.语料库与词典编纂研究：谈平行语料库中译词的自动获取[EB/OL].（2015-05-12）[2024-05-15].http://www.iol.shisu.edu.cn/06/24/c1516a67108/page.psp.
[5] 杨勇.语料库应用研究中心特邀施建军、张继东两位教授讲学[EB/OL].（2019-10-30）[2024-05-15].https://wyxy.shou.edu.cn/2019/1030/c13443a258796/page.psp.
[6] 胡文飞.基于用户视角的汉英词典理论与应用研究[M].北京：科学出版社，2019：168.
[7] 李德俊.语料库词典学：理论与方法探索[M].南京：译林出版社，2015：1.
[8] 柏野和佳子.コーパスに基づく辞書記述の精緻化の研究[D].東京：東京工業大学，2016.
[9] 荻野綱男.コーパスを利用した国語辞典編集法の研究：特定領域研究「日本語コーパス」平成21年度研究成果報告書[R].文科省科学研究費特定領域研究「日本語コーパス」辞書編集班，2010.
[10] 伝康晴，荻野綱男.講座日本語コーパス：7　コーパスと辞書[M].前川喜久雄，監修.東京：朝倉書店，2019.
[11] Zgusta L. Manual of lexicography[M]. Berlin：De Gruyter Mouton, 1971：294.
[12] 彭林宜.《汉日词典》编纂中的几个问题[J].吉林大学社会科学学报，1982（4）：77.
[13] 李明一，周红红.双语词典编纂导论[M].2版.上海：上海外语教育出版社，2011：77.
[14] 在使用EUM进行翻译对等检验时，所使用的可比语料库即两个单语语料库严格意义上须保证在文本的时间、题材、分布比例、规模等层面较为接近，但同时满足这些条件的语料库目前较为稀缺，现实研究中更倾向于使用现成的大型综合单语语料库。
[15] 卫乃兴，陆军，等.对比短语学探索：来自语料库的证据[M].北京：外语教学与研究出版社，2014：109.
[16] 公益財団法人日本女性学習財団.「婦人」と「女性」[EB/OL].（1993-09）[2024-05-20].https: //www.jawe2011.jp/cgi/keyword/keyword.cgi?num=n000017&mode=detail&catlist=1&onlist=1&alphlist=1&shlist=1.
[17] 钱厚生.语义分析与双语词典[J].辞书研究，2001（3）：80.
[18] 胡文飞，刘亚兰.语义韵视域下汉英词典文化词的记忆研究[J].外语教学理论与实践，2023（6）：31.
[19] 施建军，洪洁.汉日同形词意义用法的对比方法研究[J].外语教学与研究，2013，45（4）：531-542, 640.
[20] 施建军，刘磊，周瓠.基于词向量的汉日通用汉字词语义计量研究方法探索[J].外语教学理论与实践，2023（1）.
[21] 童富智，修刚.基于语料库的中日对应词语义韵对比研究——以中央文献日译「堅持（する）」为例[J].日语学习与研究，2020（2）.
[22] 李锡胤.双语词典的灵魂——语义对比[J].辞书研究，1980（2）：68.
[23] 一般財団法人日本特許情報機構.平成２６年度中国特許文献の機械翻訳のための辞書整備及び機械翻訳の品質評価に関する調査（調査報告書）[R/OL].2015：3[2024-05-23].https://www.jpo.go.jp/system/laws/sesaku/kikaihonyaku/document/tokkyo_dictionary/h26_01.pdf.
[24] 科学技術振興機構.高精度な日中・中日機械翻訳システムの提供を開始：科学技術論文において実用に供する翻訳率を達成[EB/OL].科学技術振興機構報：第1316号，（2018-05-08）[2024-05-23].https://www.jst.go.jp/pr/info/info1316/index.html.
[25] 国立国語研究所.学習者の辞書資源使用の実態調査：語彙検索行動の解明[EB/OL].（2024-05-22）[2024-08-13].https://www2.ninjal.ac.jp/jll/word_search/survey-data/information-sheet.html.
[26] 章宜华.计算词典学[M].上海：上海辞书出版社，2013：14.
[27] 章宜华.略论融媒体辞书的技术创新和理论方法[J].语言文字应用，2022（1）：10.
[28] de Schryver G-M. Generative AI and lexicography：The current state of the art using ChatGP-T[J]. International Journal of Lexicography, 2023: 36(4): 358.
[29] 中国辞书学会.大力弘扬中国辞书精神 共话数字时代辞书生活：中国辞书学会第十四届会员代表大会暨学术研讨会胜利召开[EB/OL].（2023-11-07）[2024-05-23].http://www.zgcsxh.org.cn/subPage/newsInfo.html?id=576&type=topicActivity.
[30] 今日语言学.2023第三届计算词典学研讨会综述[EB/OL].（2023-12-15）[2024-05-23].http://ling.cass.cn/xshd/huiyi/202312/t20231215_5717271.html.
[31] 双语辞书专业委员会.AIGC时代的学习词典编纂与知识服务：第九届学习词典与二语教学国际研讨会顺利召开[EB/OL].（2024-11-01）[2024-11-06].http://www.zgcsxh.org.cn/subPage/newsInfo.html?id=733&type=societyNews.

[32] 参见：https://dictionary.sanseido-publ.co.jp/affil/goijisho/seminar/seminar64.html（2024年5月23日浏览）。
[33] 参见：https://www.youtube.com/watch?v=mEorw0yefAs（2024年5月23日浏览）。
[34] 参见：https://korealex.org/conference/asialex23（2024年5月23日浏览）。
[35] 参见：https://elex.link/elex2023/（2024年5月23日浏览）。

参考文献

万江波.双语词典的翻译研究[M].上海：复旦大学出版社，2006.
于屏方，董静文.中国双语辞书编纂与研究三十年（1992-2022）[J]. 辞书研究，2023（2）.
王馥芳.当代语言学与词典创新[M].上海：上海辞书出版社，2004.
王禾，王秋菊.新汉日辞典[M].沈阳：辽宁人民出版社，2011.
王泽鹏，张燕春.语义韵律理论[J].同济大学学报（社会科学版），2005（4）.
吉林大学《简明汉日词典》编写组.简明汉日词典[M].沈阳：辽宁人民出版社，1985.
吕海春.中国辞书数字化发展三十年[J]. 语言战略研究，2023，8（5）.
刘艳伟，刘玉琴.基于NLB检索系统词语搭配统计值的日语近义词词典编纂构想[J].东北亚外语研究，2015，3（2）.
齐红飞，王东海.中国当代辞书学研究四十年（1978—2022年）[J].辞书研究，2023（3）.
许家金，赵冲，孙铭辰.大语言模型的外语教学与研究应用[M].北京：外语教学与研究出版社，2024.
陈贤德，杨玉玲.用户视域下外向型在线汉语学习词典评价指标体系构建[J].华文教学与研究，2025（1）.
陈楚祥.对比语言学与双语词典学[J].外语研究，1992（3）.
李宇明，庞洋.关于辞书现代化的思考[J].语文研究，2006（3）.
李宇明，王东海.中国辞书历史发展的若干走势[J].鲁东大学学报（哲学社会科学版），2020，37（1）.
李飞菲.系统科学视域下日汉词典数据库编纂实践研究——以低频词条编纂为例[J].吕梁学院学报，2023，13（6）.
陆军，卫乃兴.扩展意义单位模型下的英汉翻译对等型式构成研究[J].外语教学与研究，2012，44（3）.
宋立文.外向型汉语学习词典的出版概况、编写问题与未来之路[J].北华大学学报（社会科学版），2021，22（5）.
杨玉玲，李宇明.外向型汉语学习词典的供需错位和出路[J].辞书研究，2023（6）.
语言学名词审定委员会.语言学名词[M].北京：商务印书馆，2011.
赵福堂.《汉日词典》编辑工作中的对译问题[J].日语学习与研究，1982（3）.
施建军，徐一平.语料库与日语研究[J].日语学习与研究，2003（4）.
胡文飞.基于语料库的汉英词典儒学词研究[M].北京：商务印书馆，2023.
张永伟，刘婷.汉语辞书词条自动编纂调查研究——以ChatGPT为例[J].辞书研究，2024（5）.
徐一平.コーパス言語学から見た日本語研究と辞書編纂[C]//张昌玉，主编.日汉双语辞书编纂与日语教学文集：II.北京：中国科学技术出版社，2003.
唐闻晓，陈旭楠.毛泽东“妇女能顶半边天”若干史实考证[J].档案与建设，2023（12）.
章宜华.融媒框架下多模态词典及其释义的理论模式研究[J].外语教学与研究，2024，56（4）.
斯宛•塔普，耿云冬，张喆.主持人语技术革新背景下的词典编纂创新[J].语言战略研究，2024，9（3）.
戴宝玉.日语语料库例句词典编纂构想[J].解放军外国语学院学报，2011，34（2）.
濮建忠.扩展意义单位模型再解读[J].外语研究，2020，37（2）.
中村明.日本語語感の辞典[M].東京：岩波書店，2010.
北京商務印書館，小学館.中日辞典[M].3版.東京：小学館，2016.
石黒圭，佐野彩子，吉甜.世界の日本語学習者の辞書ツール使用事情——スマホによる語彙検索行動の適切な支援のために[J].社会言語科学，2023，26（1）.
石黒圭，吉甜，佐野彩子.スマホは辞書になりうるか：日本語学習者の辞書引きの困難点と指導法[M].東京：明治書院，2024.
西尾実，岩淵悦太郎，水谷静夫，等.岩波国語辞典[M].8版.東京：岩波書店，2019.
相原茂，荒川清秀，大川完三郎.東方中国語辞典[M].東京：東方書店，2011.
宮本華瑠.日中対訳コーパスの構築と公開に向けて[C]//国立国語研究所言語資源開発センター.言語資源ワークショップ2023発表論文集.東京：国立国語研究所言語資源開発センター，2023.
Lew R. ChatGPT as a Cobuild lexicographer[J/OL]. Humanities and Social Sciences Communications, 2023, 10[2024-10-18]. https://doi.org/10.1057/s41599-023-02119-6.
Moon R. Sinclair, Phraseology, and Lexicography[J]. International Journal of Lexicography, 2008(21).
Mahowald K, Ivanova A A, Blank I A, et al. Dissociating language and thought in large language models[J]. Trends in Cognitive Sciences, 2024.
Lin P, Martins A F T, Schütze H. A Recipe of Parallel Corpora Exploitation for Multilingual Large Language Models[J/OL]. arXiv, (2024-06-29)

[2024-10-18]. https://doi.org/10.48550/arXiv.2407.00436.
Phoodai C, Rikk R. Exploring the Capabilities of ChatGPT for Lexicographical Purposes: A Comparison with Oxford Advanced Learner's Dictionary within the Microstructural Framework[C/OL]//2023: Electronic lexicography in the 21st century. Proceedings of the eLex 2023 Conference. Brno: eLex, (2023-06-29)[2024-10-18]. https://elex.link/ojs/index.php/elex/issue/view/elex2023.
Rundell M. Automating the creation of dictionaries: are we nearly there?[C/OL]//Asialex. ASIALEX 2023: The Asian Association for Lexicography. Seoul: The Asian Association for Lexicography, 2023.
Sinclair J. Collins Cobuild English Language Dictionary: Helping Learners with Real English[M]. London: Collins ELT, 1987.
Sinclair J. The Search for Units of Meaning[J].Textus, 1996, 9(1).

作者简介：田润润　女　回族　山东大学外国语学院日语语言文学专业博士研究生　研究方向：日语语言
联系方式：E-mail：tianrunrun123@163.com
作者简介：柳长青　男　汉族　广东外语外贸大学外国语言学及应用语言学研究中心　词典学研究中心教授　研究方向：计算词典学、自然语言处理
联系方式：E-mail：liucq@gdufs.edu.cn

Exploration of Feasible Approaches for the Study and Compilation of Chinese-Japanese Dictionaries from the Perspective of Corpus Lexicography

Abstract: The compilation of Chinese-Japanese dictionaries has long been characterized by a significant imbalance between practical compilation efforts and theoretical research, compounded by outdated compilation methodologies. Corpus lexicography offers valuable theoretical insights and methodological references that can facilitate the modernization of Chinese-Japanese dictionaries. This paper, drawing upon a review of key research topics in corpus lexicography from both China and Japan, uses the Chinese culture-specific term "bànbiāntiān" (literally "half the sky") as a case study to propose a tripartite framework for the research and compilation of Chinese-Japanese dictionaries. The framework is situated within the broader context of the major trends in corpus lexicography, the empowerment of lexicography and dictionary compilation through emerging concepts and technologies such as large language models, and the user-driven transformation of dictionary compilation practices. The proposed framework consists of three layers: an inner corpus layer focused on dictionary quality evaluation, the exploration of theoretical foundations for microstructure design, and the development of specialized corpora and databases for Chinese-Japanese dictionaries; a middle corpus layer dedicated to applying findings from Chinese-Japanese contrastive studies and translation research to practical dictionary compilation; and an outer corpus layer that investigates learners' characteristics in Japanese language acquisition, users' preferences for dictionary consultation, and attitudes toward dictionary usage. This study aims to contribute to the future research and development of Chinese-Japanese dictionaries by providing both theoretical and practical insights.
Keywords: Chinese-Japanese dictionary; corpus lexicography; dictionary research and compilation; user need; large language models

Authors' Information:
Tian Runrun (Female)
Doctoral Student at Shandong University, China
Japanese Literature
E-mail: tianrunrun123@163.com

Liu Changqing (Male)
Professor at Guangdong University of Foreign Studies, China
Computational Lexicography, Na
E-mail:liucq@gdufs.edu.cn

（责任编辑：李广悦）