

大语言模型发展视域下的语料库语言学研究

上海外国语大学 日本文化经济学院

毛文伟 邢修强

[摘要] 在大语言模型迅速发展并深度介入语言学研究实践的背景下,语料库语言学正面临研究范式方面的挑战与创新机遇。面对大语言模型在数据规模和生成能力上的绝对优势,作为语言学的重要分支,语料库语言学的证据锚点真实、稳定、平衡、有代表性,能够提供细致、动态且可调的观察视角,证据可追溯性、分析过程透明性与结论可解释性较强。因此,不仅在揭示语言的分布特征、变异机制以及历时演变等方面仍具有不可替代的学术价值,还可以充当衡量和校正大语言模型语言行为的重要依据。相比之下,大语言模型在语料扩充、标注加工、观察维度以及多模态处理等方面则展现出明显优势。因此,有必要在明确语料库语言学与大语言模型研究的范式差异与能力边界的基础上,通过优势互补而非简单替代,促进两者有机协同,推动语料库语言学的范式变革与技术创新。

[关键词] 大语言模型 语料库语言学 日语 句式 偏误标注

引言

自二十世纪后半叶兴起之后,语料库语言学逐步发展为语言学研究中的重要分支。其基本理念在于借助计算机技术,对大规模真实语料进行系统检索、统计与分析,从中归纳出语言形式的使用规律和分布特征(Sinclair, 1991; Biber, 1993; Stubbs, 2001)。围绕这一研究范式,学界构建了大量设计科学、语料来源清晰、体裁分布相对均衡的大规模语料库,如英国国家语料库(BNC)、北京大学汉语语料库(CCL)以及现代日本语书面语均衡语料库(BCCWJ)等,为词汇、语法、语篇以及二语习得等领域的研究提供了坚实的数据基础。长期以来,语料库语言学在强调数据真实性、研究可重复性以及结论可解释性等方面,形成了较为稳定的方法论共识。

而大语言模型(Large Language Models, LLMs)的迅速发展和广泛应用正在显著拓展语言学研究对象、方法和问题意识,对既有研究范式产生了深远影响。一方面,大语言模型作为崭新的研究工具被引入研究实践,在数据标注、语义识别、深度分析以及多模态语言处理等任务中显示出较高的应用潜力。既有研究表明,大语言模型在词汇识别、语法模式提取等相对形式化的任务中表现尚可(Uchida, 2024),对特定言语行为构成要素(如道歉言语行为成分)的识别结果亦可达到接近人工标注的水平,从而在一定程度上缓解语用—话语语料库构建过程中人力成本高、耗时长、一致性不足等问题,使语料标注更加高效且具有批量可复制性(Yu, 2024: 534)。此外,随着大语言模型多模态处理能力的提高,研究者还能高效开展“文本+图像”等多模态话语研究,丰富了技术手段,拓展了研究领域(Baker, 2025: 76)。另一方面,大语言模型本身也已成为语言学研究新的研究对象。研究者借助语料库语言学的研究方法,从词汇分布、句法结构、语篇组织以及语用策略等各个层面对模型生成语言进行多维分析,试图评估其语言能力的水平和特征。相关研究指出,大语言模型在处理语言显性信息、把握语句字面意义方面表现出较强能力,但是未必能够理解语句背后隐

微的交际意义和与具身意义(袁毓林,2025:35)。此外,尽管大语言模型展现出良好的对话连贯性和突出的知识涌现能力,但生成内容不准确、不完全、不稳定,对自身生成内容所含礼貌知识的准确度评价不到位,难以可靠地扮演礼貌知识提供者与评估者的角色(李瑶等,2024:113)。

总体而言,前一类研究主要从技术和应用层面评估大语言模型的语言处理能力和适用场景,侧重于方法创新与可行性验证(Incelli, 2025: 1),但较少反思其对语料库语言学研究范式造成的影响。后一类研究则更多关注大语言模型的语言生成能力及其与人类语言行为的异同,但往往未深入探讨这些生成结果在多大程度上可以被视作语言研究中“可验证、可复现、可解释”的研究素材。

由于大语言模型的输出在形式上高度接近人类自然语言,也能为一些语言学领域的研究课题提供答案,部分研究者开始尝试直接利用模型生成文本作为分析材料,或将模型输出视为语言知识的来源。这在一定程度上模糊了“自然语言使用的事实与规律”与“算法生成结果”之间的界限,对语料库语言学长期确立的基于“有限、真实、平衡”语料的研究范式提出了新的挑战。在此背景下,语料库语言学是否仍具有不可替代的学术价值,成为亟待回应的重要问题。

本文立足语言学研究的目标和需求,结合既有研究和实际案例,重点探讨两个相互关联的问题:其一,在大语言模型迅速发展的背景下,语料库语言学的学术价值主要体现在哪些方面?其方法论是否仍具有不可替代性?其二,大语言模型可以通过何种路径对语料库语言学研究进行赋能?应如何界定其应用原则与边界?本文拟通过对这两个问题的系统讨论,澄清当前关于语料库语言学与大语言模型关系的若干认识误区,并为二者的有机融合提供较为科学的、具有可操作性的框架思路。

1 大语言模型飞速发展背景下语料库语言学研究的价值分析

尽管大语言模型的快速发展为语料库语言学研究带来了前所未有的助力,显著强化了研究者的语料获取、标注加工以及分析能力,但大语言模型的输出在语言的准确性、代表性、视角可调节能力、过程透明性以及结论可解释性等关键问题上,尚难以完全满足语言学研究的要求。因此,语料库语言学研究并未因大语言模型的发展而失去意义,反而在以下关键层面凸显出其不可替代的重要价值。

首先,语料库语言学研究具有真实、稳定、平衡、有代表性的证据锚点。

语料库语言学长期注重语料的真实性、代表性与平衡性,强调以人类产出的真实语言为基本素材,经过系统筛选,平衡体裁、规模,尽量全面、准确地反映语言的实际使用状况,确保研究结论建立在可验证、可追溯且具有可比性的经验数据基础之上。

相比之下,大语言模型训练数据的构成存在明显的结构性不平衡。从数据来源看,模型训练所用文本集中于网络公开资源,如新闻网站、维基百科、技术文档及部分论坛文本,对口语会话、方言材料以及非标准书写形式的覆盖相对不足。这使得模型更容易学习到书面语中规范化、语境依存度较低的表达模式,而难以充分掌握实际交际活动中的语言多样性和相关语用策略。此外,从时间维度来看,大语言模型训练数据以现代、当代文本为主,对古代、近代文献的覆盖明显不足。这导致模型在处理历史词汇、语法形态以及历时语言演变等问题时表现欠佳。

不仅如此,大语言模型的语言生成机制决定了其输出的文本虽然在形式上较为流畅自然,但本质是对训练语料分布概率的拟合结果,即在给定条件下出现可能性最高的语言形式,而非对自然语言事实的直接记录。在生成过程中,大语言模型往往会对真实语料中的频率差异进行平滑处理,弱化高频与低频项目之间原本具有区分意义的分布梯度,甚至可能生成在真实语料中极少出现或并不存在的搭配形式,导致出现与自然语言分布不一致的结构性偏差。尤其是在涉及句式选择、搭配限制或语用制约等对分布敏感的语言

现象时,大语言模型生成的文本往往难以准确再现其在真实语言的分布状况。此类偏差常常具有模式化特性,会在大规模语料生成过程中反复出现。同时,训练数据的规模对模型性能也有很大影响。对于低资源语言,开发者往往无法获得足够规模的高质量训练语料。例如,在GPT-3.5的训练数据集中,英语语料占92.1%。而法语仅占1.78%,德语仅占1.68%,日语、汉语的占比仅为0.16%(毛文伟等,2003:27)。这直接导致模型对这些语言的生成能力显著落后于英语。此外,大语言模型的输出还具有一定的不确定性,无法充分满足语言学研究的需要。

以表示强烈主观情感的日语构式「てならない」为例,该构式的前接词具有明显的词性、语义特征。在现代日本语书面语均衡语料库中,笔者共检索到实例811句。统计前接词发现,排名前20位高频词(以下简称BCCWJ高频词)的例句共有693句,占总体的85.5%,呈现出高度集中的分布特征。从词性来看,自发性动词或动词词组占据多数例(1),表示情感的形容词及形容动词也占较大比例例(2),其他类型的词汇几乎不与该构式搭配使用。这表明,「てならない」对前接词具有较强的共现制约。

(1) 言葉として強すぎる気がてならない。

(『血を吸う夫』新津きよみ)

(2) それにしても、ロンドンで、なつかしいしぐさを見た時は、何だか嬉しくてならなかった。

(『まかふしぎ・猫の犬』出久根達郎)

为了检验大语言模型生成文本是否能够准确再现此类构式搭配特征,笔者通过API向通义千问等四款主流大语言模型发送指令,要求其各自生成1000句包含「てならない」的例句。随后,统计这些例句中的前接词,并与BCCWJ高频词进行比对后,获得表1(因篇幅所限,高频词表从略)。结果显示,大语言模型在再现构式搭配特征方面的能力较弱。以通义千问为例,在其生成的例句中,含前20位高频前接词的例句占比仅为21.6%,远低于真实语料的85.5%,且仅有6个高频词与BCCWJ高频词一致,明显缺乏自然语言中的集中趋势。即便是表现最好的GPT-5,含前20位高频前接词的例句占比也仅为64.7%,其中与BCCWJ高频词一致的刚过半数。这说明,GPT-5的输出虽部分反映了该构式对前接词的搭配限制,但仍无法完全重现其实际使用状况。

值得注意的是,在这些模型生成的例句中,前接词的词尾活用基本正确,但出现了大量语义属性不符合构式使用规律的情形。「てならない」的前接词不仅应是[-自制性]的,而且必须是[+状态性]的(毛文伟,2002:64)。但例(3)、例(4)中的「感動する」「驚く」表示的均为情绪变化,因此是[-状态性]的。这些不自然的搭配反映出模型在处理特定构式时仍倾向于机械沿用词汇组合的泛化结构,而未能充分掌握「てならない」的前接词的独特语义特征。

表1 「てならない」的高频前接词统计表

	BCCWJ	通义千问	DeepSeek	Gemini	GPT-5
前20位高频词例句比	85.5%	21.6%	27.1%	35.3%	64.7%
共通高频词数	20	6	5	7	11

资料来源:对现代日本语书面语均衡语料库检索及大语言模型生成用例统计获得。

(3) 感動してならない。

(通义千问)

(4) その技術革新のスピードに驚いてならない。

(Gemini)

其次,语料库语言学研究具有细致、动态、可调的观察视角。

语言并非均质、静态的封闭系统。词汇、句式及语用策略等在不同语域、体裁与交际情境中的出现频率与功能负载并不均衡,且随时间不断发生历时更迭。语言研究者通常会根据研究目标,聚焦特定语言变体,对某个语言现象的分布特征与使用规律进行定量分析或定性解构。在这一过程中,语料的可控性与可区分性具有重要意义。

语料库语言学的核心价值正是在于其为捕捉这种非均质的语言分布模式提供了细致、动态、可调的观察手段。通过设定语料的语域、体裁、交际情境及时间区间等,研究者能够获得较为稳定且可比的观察框架,可以在较为明确的条件下观察语言现象,区分偶然用例与系统性倾向,揭示语言的阶段性特征以及变化的方向、速度。

相比之下,大语言模型的生成结果并不具备明确的语域、体裁、交际情境或时间区间特性。大语言模型在训练过程中会将不同类型的语言材料混合建模,因此弱化了不同语言变体的差异,呈现一种平均化的状态。故模型生成的文本虽或可模拟某种典型用法,却无法准确反映该形式在不同变体中的真实使用状况,难以开展细致的共时或历时比较,无法据此考察语言的分布特征或变异规律。

仍以日语句式「てならない」为例,笔者按时代在日本国立国语研究所开发的日语历史语料库(<https://clrd.ninjal.ac.jp/chj/>)、昭和平成书面语语料库(<https://clrd.ninjal.ac.jp/SHC/>)以及现代日本语书面语均衡语料库下属各子库中检索实例,经人工排除表示禁止等意义的用例后,计算了该句式在不同时代语料中的使用频率,获得表2。为更直观地呈现该句式在不同时代的使用状况及变化趋势,笔者进一步运用表2中的时代信息和调整频率(每百万词的使用频率)数据绘制了下页图1。

表2 各时代「てならない」使用频率统计表

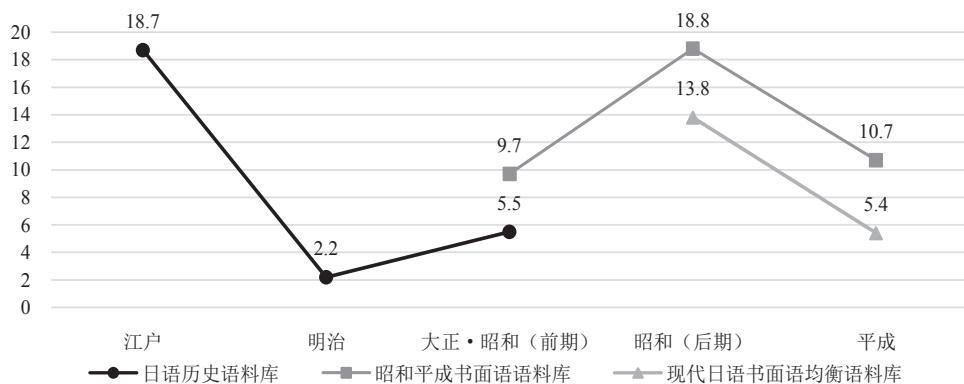
时代	语料库	子库	库容	检出数	调整频率
奈良	日语历史语料库	1 奈良	125,907	0	0.0
平安		2 平安	1,078,560	0	0.0
镰仓		3 鎌倉	1,153,191	0	0.0
室町		4 室町	367,610	0	0.0
江戸		5 江戸	1,014,710	19	18.7
明治		6 明治	9,589,092	21	2.2
大正・昭和(前期)		7 大正、8 昭和	5,418,498	30	5.5
大正・昭和(前期)	昭和平成书面语语料库	1933、1941、1949	7,841,333	76	9.7
昭和(后期)		1957、1965、1973、1981	12,728,695	239	18.8
平成		1989、1997、2005、2013	12,834,816	137	10.7
昭和(后期)	现代日本语书面语均衡语料库	1988年前出版的语料	9,451,590	130	13.8
平成		1989年之后出版的语料	95,459,870	512	5.4

资料来源:根据各语料库子库检索结果计算而得。

由表2所示数据可知,「てならない」在各个时代的使用频率存在明显差异。在奈良、平安、镰仓及室町时期的语料中均未检出相关用例,表明该句式在古代日语阶段尚未形成,或尚未进入书面语系统。进入江

户时期后,该构式首次出现。尽管检出数仅为19例,但调整频率达到18.7,说明该构式在这一时期已在书面语中初步确立,开始被反复使用。不过,由于该时期的库容相对较小,统计结果具有一定偶然性。

在明治时期,检出的例句有所增加,但由于语料规模大幅扩大,其调整频率反而下降至2.2。这表明,该构式在这一阶段尚未得到广泛应用。进入大正、昭和前期后,在日语历史语料库、昭和平成书面语语料库中,「てならない」的使用频率均明显上升。尤其是在后者中,其调整频率达到9.7,显示该构式在近代书面语中已经逐渐普及。至昭和后期,该构式的使用达到高峰,调整频率升至18.8及13.8。这说明,「てならない」已成为表达强烈主观情感的稳定构式之一。



资料来源:表2中的时代信息及对应的调整频率数据。

图1 各时代「てならない」的使用频率变化图

进入平成时期后,无论是在昭和平成书面语语料库,还是在现代日本语书面语均衡语料库中,「てならない」的使用频率都出现了一定程度的回落。总体而言,「てならない」的历时演变路径呈现出“近代产生—昭和时期高频化—近期有所降低”的发展模式,显示出语法构式在不同历史阶段使用兴衰的动态特征。此外,「てならない」的使用频率在不同文体的语料中也有较为显著的差异。篇幅所限,不再详细论述。而至少在现阶段,大语言模型尚难以准确反映此类语言现象的历时及文体分布特征。

第三,语料库语言学研究具有良好的证据可追溯性、过程透明性和结论可解释性。

语料库语言学研究以真实语言数据为基础,研究者可以明确追溯每一项分析结果所对应的具体例句、统计数据和语域语境等,从而使研究结论建立在可验证、可复现、可比较的证据之上。通过观察和比对,研究者不仅能够了解“出现了什么”,还能够进一步解释“为何如此出现”,并将分析结果与既有语言理论进行比较。这种从实例出发、逐层归纳的研究路径,使语料库语言学研究过程透明,逻辑清晰。

相比之下,大语言模型的输出具有明显的黑箱特征。基于以 Transformer 为代表的深度学习技术,模型通过由多层神经网络构成的参数体系,对大规模语料中的词序列进行学习,在此过程中不断调整权重以最小化预测误差。其内部参数和决策路径难以直接解读。即便模型能够针对某个研究课题提供解答,研究者也很难判断其依据,难以进行验证。同样,当模型输出与语言学直觉或理论预期不符时,也无法分析原因、修正完善。即便思维链技术(Chain-of-Thought, CoT)在一定程度上显化了模型的思考过程,但这更适宜被视为一种性能增强与分析辅助技术,而非模型内部认知过程的反映。对于语言学研究而言,证据的不可追溯性和过程的不透明性构成了方法论层面的重要制约。

第四,语料库语言学研究能够成为评估和校正大语言模型语言行为的重要参照依据。

在大语言模型快速发展的过程中,如何评估其语言能力已成为语言学与语言技术研究中亟待回应的问

题。基于真实语言数据构建的语料库语言学研究能够帮助研究者准确判断模型输出是否存在分布差异、过度泛化或语域混淆等问题,为模型能力的评估与校正提供参照与支点。例如,毛文伟、朱海莹(2024:63)指出,与人工译文语料相比,大语言模型生成的汉日翻译文本语气较为平淡,情感色彩不足,接续词使用不如人类译者丰富、准确,且表现出更强的规范化翻译共性特征。通过将模型输出与语料库中的真实语料进行对比分析,研究者不仅能够更清晰地界定大语言模型语言能力的水平和边界,也有助于推动模型在语言多样性、语域适配性与历时敏感性等方面的改进。这一检验与校正功能正是语料库语言学在人工智能时代展现出的新的应用价值。

2 语料库语言学与大语言模型研究的范式对比

从学科渊源来看,语料库语言学属于语言学内部的方法论革新,而大语言模型研究则发端于计算机科学的人工智能分支。两者的理论基础、研究方法、数据取向和目标设定迥异。在探讨大语言模型赋能语料库语言学研究的路径与前景之前,有必要从这些维度出发,对二者的研究范式进行全面对比。只有充分认识二者的特点及差异,才能更加准确地设计赋能路径,预判在此过程中可能出现的认识错位和矛盾,并为后续构建大语言模型与语料库语言学互补共生的研究体系提供理论支撑与方法论准备。

首先,在理论基础方面,两者之间存在着是以经验归纳还是以概率预测为核心的显著差异。

语料库语言学遵循典型的经验主义和功能主义的语言观,强调语言知识的经验性与可验证性,注重描述语言实际而非预设规范,贯彻定性定量分析与语言直觉相结合,坚持通过对真实语言材料进行系统观察与统计分析,归纳语言形式的意义、功能、分布及使用规则。例如,Sinclair(1991)指出,词汇的意义往往不在于其本身,而在于它与其周围词汇的共同出现,强调通过索引(Concordance)观察真实语料中的搭配。Stubbs(2001)则认为,意义是通过重复的使用模式构建的,只能通过语料库数据来研究。

与之相对,大语言模型研究的理论基础来自统计学中的概率模型和认知科学中的人工神经网络思想,即将语言视为可由大规模数据近似建模的统计分布。开发者基于海量文本进行预训练,构建庞大的词汇与序列表征的统计分布模型,通过优化损失函数提升对下一个词的预测能力,以便产生与训练语料分布高度一致的语言形式。因此,大语言模型虽然能够生成形式上流畅且符合表达习惯的句子,但其输出并不必然契合人类交际的真实语境、语用规则或社会文化规范。这种以统计拟合替代语义推理的机制,使大语言模型所呈现的“语言知识”在本质上不同于人类基于经验归纳而形成的语言知识体系。

其次,在研究方法方面,两者之间存在着“探索、验证结合”还是“训练、评测循环”的显著差异。

语料库语言学研究通常首先对语料库进行检索和统计,获取研究对象的共现词表或频率分布等。随后,审视所获实例或数据,结合既有理论及成果,在具体语境中探究其中蕴含的语言规律。研究方法呈现自下而上的探索与自上而下的验证相结合。既讲究数据驱动,又注重理论支撑以及研究者对语义和语境等的直觉判断。

大语言模型研究的路径则高度自动化和工程化。首先,开发者确定算法和训练目标,让模型在大规模语料上自动学习。训练完成后,在标准基准任务中开展测试,以量化指标评估模型在各项任务上的得分,或用可解释性工具或探针试验探究模型学到了哪些语言知识。若未达标,则调整参数再次训练。如此循环,直至达成预定目标。研究方法以自上而下的验证为主。研究者很少去逐一阅读训练语料中的句子,更可能阅读的是模型生成的文本或注意力权重分配模式等模型产物。

再次,在数据取向方面,尽管两者均以大规模语言数据为基础,但呈现出“有限而精准”与“以大为美”的显著差异。

语料库语言学强调数据的真实性、代表性与均衡性。在语料库构建过程中,研究者会根据研究目标,严格控制语料来源。通过对不同语域、体裁与话题的合理配置,确保语料库在规模扩展的同时保持内部结构的相对均衡,避免单一来源或特定文类对研究结论造成过度影响(Mackawa et al, 2014)。与此同时,语料库开发还强调数据采集与处理的透明性与可追溯性。语料的选取标准、标注方式及清洗规则均需明确记录,使其能够被他人复核与再现(小木曾智信等,2004)。这种以规范化构建为前提的语言数据为后续研究提供了可靠保障,也构成语料库语言学研究可信度与解释力的重要来源。

与之相对,大语言模型的训练遵循一种“以大为美”的数据策略,即追求绝对规模,并优先考虑数据可获得性。例如,GPT-3.5的训练数据集已包含1953亿词。后续开发的主流大语言模型的训练语料规模更加庞大。这些语料主要来自网络,辅以通过版本授权或合作获取的数据。优势在于覆盖面极大,但在语言质量、信息可靠性以及规范性等方面差异明显,导致模型在训练过程中可能吸收不正确的分布特征或不可靠信息,并在输出中出现系统性偏差或谬误。

再次,在目标设定方面,两者之间存在着“解释性取向”与“性能取向”的根本差异。

语料库语言学的核心目标在于探究语言现象,解释语言规律与背后动因。因此,语料库语言学研究通常关注结论的理论贡献,通过归纳、解释语言现象,修正、丰富既有语言学理论,有时还与系统功能语言学、认知语言学等其他语言学分支充分互动,共享发现。

与之相对,大语言模型研究更加重视算法迭代与性能优化。语言学意义并非其核心目标。开发者的关注对象集中于困惑度(perplexity)、准确率、BLEU或F1等指标的提升幅度,对于模型内部如何表征语言、为何能够完成任务等兴趣较低。这种注重模型性能而弱化理论解释的倾向导致了其与语料库语言学研究之间存在着明显的兴趣差异。一方强调理解机制,一方关注结果优劣。两者长期缺乏共同的对话框架。

综上所述,研究范式的根本差异导致语料库语言学与大语言模型研究在理论基础、研究方法、数据取向和目标设定等方面发生错位。语言学家可能批评模型输出缺乏语用得体的性或逻辑一致性。但从模型开发者的角度来看,只要生成结果与训练语料在形式上高度一致,即可视为性能良好。两者的评价依据并不在同一层面。正因如此,理解并调和这两种研究范式的差异是推动大语言模型技术赋能语料库语言学研究的前提和关键所在。

3 大语言模型对语料库语言学的赋能

当然,尽管语料库语言学研究仍然具有不可替代的理论和应用价值,也并不意味着我们要故步自封。大语言模型的兴起正在深刻地改变语料库语言学的发展方向,为研究者提供了更加多元的分析工具和研究对象,显著地拓展了研究视野,丰富了研究内容。其赋能路径并非单一维度的能力提升,而是涉及语料扩充、标注加工、观察维度、研究范围等多个层面的系统性重塑。

首先,大语言模型为语料库语言学带来了前所未有的语料扩充能力。

传统语料库的构建往往需要耗费大量人力与时间。而大语言模型具备的文本生成能力能够在给定语境、风格或内容的条件下大量生成文本,从而快速补充低资源语言、稀有语言现象或专业领域语体的样本。这种可控生成语料的能力使语料库建设从被动收集走向主动扩充,也为模拟语言变异、构造极端语境或验证语言规律提供了实验条件。在低资源语言或专业领域语料匮乏的情况下,研究者开始尝试使用GPT等模型生成补充性语料,以扩展语料库规模或平衡不同类别数据的比例,取得一定的实践成效。例如,Li et al. (2023)尝试用GPT-3.5-Turbo生成了新闻、影评、金融短语等10种类型的文本,发现任务的主观性程度与大语言模型的表现密切相关。邹睿等(2025)则基于预训练的音素大语言模型XPhoneBERT和轻量化扩散语

音合成模型 LightGrad 生成越南语语音数据,显著提高了韵律表现力和合成质量。

为了提高语料采集效率以及输出的稳定性、一致性,我们通常以调用 API 的方式获取大语言模型的生成结果。当前主流模型大多提供了成熟、规范的 API 接口。由于 OpenAI 在业界起步最早,其接口设计和调用方式已经成为事实上的通用规范。研究者可在 Python 编程环境下调用 openAI 库,向模型发送提示词,并以统一的数据结构获取生成结果。该方式适用于 ChatGPT、DeepSeek、通义千问、Claude、Gemini 等绝大多数支持 OpenAI 接口规范的大语言模型。与人工逐条复制网页生成结果相比,调用 API 在效率和可控性方面具有明显优势。不仅能够显著提升语料的采集速度和稳定性,满足批量生成和重复实验的需要,还能通过控制参数、提示词内容与版本,确保语料生成条件的一致性,提高研究过程的可重复性和结果的可比性。

然而,正如前章所述,尽管大语言模型具备强大的语言产出能力,但与源自真实交际场景的语料相比,其生成文本在语言分布的准确性、语用自然性以及信息真实性等方面仍存在结构性缺陷。因此,大语言模型在语料获取中的定位应当是“辅助”而非“替代”。我们需要建立一种“人机协同”的质量控制体系,以传统语料库为基准,以生成语料为补充,确保语言数据的可靠性。

其次,大语言模型为语料标注与加工提供了更加高效、一致的技术手段。

语料库语言学长期依赖人工或半自动标注。但要完成歧义消解、语义分类、篇章结构分析以及偏误界定等复杂任务,人工标注所需时间、成本过高,且往往难以确保操作的一致性。借助大语言模型强大的语言处理能力,研究者可以实现较高精度的语料自动标注、加工,显著提高语料处理效率。更重要的是,大语言模型具备跨语言迁移能力,使得语料标注体系可以在不同语言之间更加平滑地迁移,为多语种平行语料库建设提供有力支持。

以日语格助词偏误标注任务为例,笔者通过 Python 编程调用 API,将例(5)所示提示词发送给 GPT-5,要求其对于康等(2017a, 2017b, 2018)考察的 205 例格助词典型偏误例句进行标注。这些例句筛选自 YUK 已标注汉语母语日语学习者作文语料库(YUK タグ付き中国語母語話者日本語学習者作文コーパス)。该语料库由资深日语母语研究者进行人工偏误标注,被视为高质量、可靠的基准数据。笔者以其中的人工标注为标准,对 GPT-5 的标注结果进行评估,并计算了准确率。结果如表 3 所示。

表 3 GPT-5 日语格助词偏误标注准确率统计

		正确	可接受	错误	合计
	例句数	122	46	37	205
GPT-5	准确率	59.5%	22.4%	18.0%	100%

资料来源:根据比对结果计算获得。

(5) 请按照下例的方式修改例句。修正之处请标记为[修正前の表現→修正後の表現, 誤用の種類]的形式。只要输出修改结果,不需要说明理由。如有多个错误,请全部修改。

例

#input#

昔の人々が旅行したとき、わらじが破ればまた新しいわらじに交換した。

#output#

昔の人々が[人々が→人々は, が→は]旅行したとき、わらじが破ればまた新しいわらじに交換した。

#例句#

確かに多くの中国人学習者は読み書きができるが、実際の場面での言語の運用能力は低い。

在表3中,“正确”指 GPT-5 给出的标注与人工标注完全一致。“可接受”指两者尽管不同,但 GPT-5 的修改仍然合理、准确。例如,在例(6b)中,尽管 GPT-5 未如人工标注例(6a)那样将格助词「と」改为「に」,而是替换了谓语动词,但偏误同样得到了修正,因而被视为可接受的正确处理。从整体结果来看,完全一致的标注比例达到 59.5%,可接受的标注比例为 22.4%。二者合计约 82%,达到相当高的水平。尤其是格助词偏误属于词汇、语法与句式共同作用的复杂语言现象。在大多数情况下,GPT-5 仍能做出符合语法与语用规范的判断,说明其不仅能够识别形式偏误,还能在一定程度上理解动词格关系和上下文制约。而且,GPT-5 在修改动词例(7)、副词等偏误时也表现良好。不仅准确率很高,而且与人工标注往往仅关注主要偏误不同例(7a),GPT-5 会同时做多处修改例(7b),使得结果更加准确、自然。可见,大语言模型在日语偏误标注领域具有较高的应用潜力,可充当人工标注的辅助工具,为大规模学习者语料的自动加工提供技术支撑。

(6) 中国と反して、日本では、ゴミの分類と処理は極めて緻密である。 (原句)

(6a) 中国[と→に]反して、日本では、ゴミの分類と処理は極めて緻密である。 (人工标注)

(6b) [中国と反して→中国と違って]、日本ではゴミの分類と処理は極めて緻密である。 (GPT-5 标注)

(7) 価格と価値の一致を保証して、消費者に十分な娯楽の楽しみをあげて、独特な文化を感じ取る場所を建設した。 (原句)

(7a) 価格と価値の一致を保証して、消費者に十分な娯楽の楽しみを[あげて→与え]、独特な文化を感じ取る場所を建設した。 (人工标注)

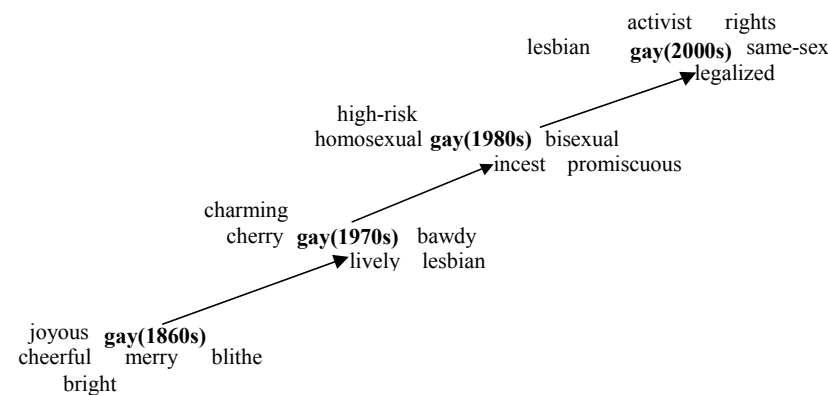
(7b) 価格と価値の一致を保証して、消費者に[十分な→十分に]娯楽の楽しみを[あげて→与えて]、独特な文化を感じ取る場所を[建設した→作り上げた]。 (GPT-5 标注)

第三,大语言模型为语料库语言学提供了强大的技术支持和全新的观察维度。

大语言模型强大的代码生成能力能够为语料库语言学研究提供高效、快捷的编程工具,显著降低复杂数据处理与统计分析的技术门槛。通过简单的任务提示,研究者就可快速生成用于执行语料清洗、格式转换、文本检索、频率统计、共现分析以及数据挖掘等任务的程序代码,减少在编程技能学习以及编程实现上的时间投入,使精力更多地集中于语言学研究本身。此外,在多语言、多语料并行处理的研究场景中,大语言模型还能辅助构建可重复执行的分析流程与脚本框架,提升研究过程的规范性与可移植性。当然,大语言模型应被定位为一种方法层面的赋能工具,而非研究分析的替代者。研究者仍需对生成代码的逻辑合理性、统计方法的适切性以及分析结果的科学性承担最终责任。只有在明确人机分工边界的前提下,大语言模型的代码生成能力才能真正发挥其在语料库语言学研究中的作用,为提升数据处理效率,加快研究迭代提供有力支持。

此外,深度学习等大语言模型相关技术也为语料库语言学提供了强大的分析工具。传统上,语料库分析主要依赖文本行分析以及统计学方法,如词频、n-gram、卡方检验、主成分分析等。这些方法擅长揭示语言的表层规律,但对探索更复杂的模式(如依存距离、隐含语义)等往往力不从心。深度学习模型则可以学习

语言高维表征,捕捉复杂模式。例如,基于BERT等的上下文嵌入技术可以从语料中学习词汇的语义向量,据此进行词语聚类、类比,能够揭示词汇意义的相关性及历时变化,深入挖掘语言现象之间的隐性关联。例如,Shi & Lei(2020)运用词向量技术,提取并对比“gay”的语义近邻词后发现,在十九世纪六十年代至二十世纪六十年代期间,其近邻词多为表达快乐和愉悦的积极词汇,如“merry(愉快的)”“blithe(无忧无虑的)”等(图2)。可见,在该阶段,“gay”主要表示快乐的积极语义。到了二十世纪七十年代,在保留“charming(迷人的)”等积极词汇的同时,开始出现“bawdy(粗俗的)”“lesbian(女同性恋)”等负面近邻词。这可能是“gay”的语义从“快乐”向“同性恋”转变的过渡阶段。自二十世纪八十年代开始,近邻词已变为“homosexual(同性恋的)”等,显示其已演变为专门表示同性恋的词汇。同时,与诸如“high-risk(高风险的)”等强烈负面的词汇建立了关联。到了二十一世纪初,“gay”逐渐摆脱了负面语义,除了仍保持同性恋的语义外,拥有了诸如“rights(权利)”和“legalized(合法化)”等近邻词。可见,该词所代表的社会现象在英语世界越来越被公众所接纳和保护。



资料来源:引用自 Shi & Lei(2020:4)图1(a)。

图2 “gay”的语义变化示意图

最后,大语言模型有助于促进语料库语言学研究向多模态、跨领域拓展。

大语言模型的快速发展显著提升了语料库语言学的数据处理能力,正在推动其从单一文本分析向多模态与跨领域方向拓展。受制于研究手段,传统的语料库语言学研究多依赖文字材料,而无法评估图片、音频、视频及其他模态等非语言符号对话语的影响。大语言模型通过整合多模态信息,使得语言研究不再局限于纯文本范畴,而是在更加丰富的感知维度上考察意义的生成与传播。例如,Baker et al.(2025)运用谷歌的Vertex AI为新闻图片自动设置标签并指出,因为纳入了图像标签,为分析带来了新维度。这种方法使分析者能够比仅仅考虑书面文本更清晰地理解有关伊斯兰教和穆斯林的新闻话语模式。借助具备图像、视频理解能力的大语言模型,研究者还可以探讨语言表达与视觉信息之间的对应关系,从而推动多模态语义学、手势与语言协同研究、图文对齐研究等新领域的发展。同时,具有语音识别与生成能力的模型也将为音系学、韵律学、话语分析等依赖语音材料的研究带来更加高效的数据处理与实验模拟能力。

此外,大语言模型及其技术的跨领域适应能力使语言学研究与认知科学、心理学、教育学、计算机科学、新闻传播学等学科形成更加紧密的互动关系。模型的推理能力与类人语言行为为探索语言加工机制、语言习得路径和语言产出规律提供了新的“计算被试”,从而扩大了实验语言学与认知语言学的研究空间。在文学研究、跨文化传播等人文领域,大语言模型也能够通过风格迁移、叙事分析、主题抓取等方式,为研究者提供新的分析工具和观察视角。例如,马爽、毛文伟(2025)运用BERT主题模型,对比、分析《西游记》英、日译本的读者评论发现,前者注重副文本的实用性与信息密度,关注范围更广。后者则更关注趣味性 with 思想内

涵,理解更深。因此,针对高、低文化距离市场,应分别采取以激发兴趣为核心的“文化引渡”和以深化理解为核心的“经典再激活”两大差异化传播策略,有效促进中国文学全球传播。

当然,在利用大语言模型进行语料加工和分析时,我们需谨慎对待可能出现的“幻觉”现象。大语言模型提供的某些信息可能并无依据或不够严谨。目前,较为成熟的做法是将大语言模型的输出作为参考,由研究者进行验证和解读,而非盲目信任模型输出。一些研究案例显示,大语言模型的分析结果往往需要结合真实语料统计才能确认其是否准确、可靠。例如,Uchida(2024)比较ChatGPT与语料库工具对词频和搭配的提取结果发现,虽然在整体趋势上二者有一定程度的吻合,但在细节上前者仍存在较大的遗漏和偏差。尤其是在语法模式填充测试中,两者生成的高频词重合率仅53.0%。因此,大语言模型更适合用于宏观模式的发现和辅助,而精确的量化分析仍需借助传统的语料库工具或结合统计检验来完成。

4 结 语

综上所述,大语言模型的迅速发展为语料库语言学研究提供了新的研究课题和技术手段,显著提升了研究效率,拓展了研究空间。然而,受制于其训练数据和算法,大语言模型尚难以独立承担提供语言证据及结论的角色。而以真实语言为基石的语料库语言学在论据真实性、稳定性、代表性、视角灵活性、过程透明性以及结论可解释性等诸多方面,仍具有不可替代的重要价值。

因此,大语言模型在当前的语言学研究中更适宜被定位为辅助工具,而非研究数据或结论的主体来源。在坚守基本原则的前提下,合理利用大语言模型的技术优势,有助于推动语料库语言学研究在分析能力、分析效率以及分析维度等方面进一步拓展。这种以真实语料为基础、以智能技术为支撑的研究范式将为语言学研究的持续进化提供更为稳健、高效的发展路径,促进其范式革新和理论创新,开启人工智能时代语言科学发展的新篇章。

[本文为2024年度上海外国语大学科研专项项目“文明互鉴视域下的中外语言与文化比较研究”(项目编号:41005121)阶段研究成果。项目负责人:毛文伟]

参考文献

- 于晖,宋金戈.知识共建视域下生成式人工智能辅助的外语教学——定位、应用与发展[J].中国外语,2025,22(3).
- 于康,林璋,于一乐,等.日语格助词的偏误研究:上[C].浙江:浙江工商大学出版社,2017a.
- 于康,林璋,于一乐,等.日语格助词的偏误研究:中[C].浙江:浙江工商大学出版社,2017b.
- 于康,林璋,张佩霞,等.日语格助词的偏误研究:下[C].浙江:浙江工商大学出版社,2018.
- 马爽,毛文伟.基于BERT主题模型的《西游记》英、日译本读者体验对比研究[J].外语教学,2025,46(6).
- 王羽迪.生成式AI在日译汉翻译中的质量评估研究——以大语言模型GPT-4o为例[J].日语学习与研究,2025,(2).
- 毛文伟,朱海莹.计量文体学视角下的汉日机器翻译语言特征研究——以《阿Q正传》日译本为例[J].日语学习与研究,2024(6).
- 毛文伟,谢冬,郎寒晓.ChatGPT赋能新时代日语教学:场景、问题与对策[J].外语学刊,2023(6).
- 毛文伟.试析复合辞“～テナラナイ”“～テショウガナイ”“～テタマラナイ”的异同——语料库统计法在语法研究中的应用一例[J].解放军外国语学院学报,2002(3).
- 冯聪,李颖.人工智能在外语教学与研究中的创新应用——基于语言数据、大语言模型和数字技术的多维度探讨[J].外语电化教学,2025(3).
- 李瑶,于富喜,毋育新.汉日语境中大语言模型的礼貌知识生成能力探析[J].日语学习与研究,2024(5).
- 何巍,韦昱妃,夏一雪,等.大模型驱动的人机审议式网络谣言治理决策框架研究[J].情报杂志,2025,44(12).
- 邹睿,杨鉴,张凯.基于音素大语言模型及扩散模型的低资源越南语语音合成[J].计算机科学,2025,52(S1).
- 张曜康,赵朝永.大语言模型之于文学翻译的适切性研究——基于多指标评估的《边城》多模型译文质量对比[J].中国外语,2025,22(4).
- 赵紫荆,张开平.广义的媒介:生成式人工智能引发的传播学范式变革[J].东南学术,2026,(1).
- 袁毓林.ChatGPT等大模型的语言处理机制及其理论蕴涵[J].外国语(上海外国语大学学报),2024,47(4).

- 袁毓林. 从三种复杂句看 ChatGPT 是不是随机鹦鹉——语言大模型能不能理解语言意义的测试与讨论[J]. 语言教学与研究, 2025(1).
- 小木曾智信, 近藤明日子, 高橋雄太, 等. 「昭和・平成書き言葉コーパス」の設計・構築・公開[J]. 情報処理学会論文誌, 2024, 65(2).
- Baker P, Schmück H, Qian Y. Automatic image tagging for corpus linguistics: A multimodal study of news representations of Islam[M]. Cambridge: Cambridge University Press, 2025.
- Biber D, Conrad S, Reppen R. Corpus linguistics: Investigating language structure and use[M]. Cambridge: Cambridge University Press, 1998.
- Incelli E. Exploring the future of corpus linguistics: Innovations in AI and social impact[J]. International Journal of Mass Communication, 2025(3).
- Li Z, Zhu H, Lu Z, et al. Synthetic data generation with large language models for text classification: Potential and limitations[J]. arXiv preprint, arXiv:2310.07849, 2023.
- Mackawa K, Yamazaki M, Ogiso T, et al. Balanced corpus of contemporary written Japanese[J]. Language Resources and Evaluation, 2014, 48.
- Shi Y, Lei L. The evolution of LGBT labelling words: Tracking 150 years of the interaction of semantics with social and cultural changes[J]. English Today, 2020, 36(4).
- Sinclair J. Corpus, concordance, collocation[M]. Oxford: Oxford University Press, 1991.
- Stubbs M. Words and phrases: Corpus studies of lexical semantics[M]. Oxford: Blackwell, 2001.
- Uchida S. Using early LLMs for corpus linguistics: Examining ChatGPT's potential and limitations[J]. Applied Corpus Linguistics, 2024, 4(1).
- Yu D, Li L, Su H, et al. Assessing the potential of LLM-assisted annotation for corpus-based pragmatics and discourse analysis: The case of apology[J]. International Journal of Corpus Linguistics, 2024, 29(4).

作者简介: 毛文伟 男 汉族 上海外国语大学日本文化经济学院教授 研究方向: 日语语言、语料库应用研究、二语习得研究

联系方式: E-mail: maowenwei@gmail.com

作者简介: 邢修强 男 汉族 上海外国语大学日本文化经济学院日语语言学专业博士研究生 研究方向: 日语语言、语料库语言学

联系方式: E-mail: xingxiuqiang@126.com

Corpus Linguistics Research in the Context of the Rapid Development of Large Language Models: Value and Empowerment

MAO Wenwei & XING Xiuqiang

(Shanghai International Studies University)

Abstract: Against the backdrop of the rapid development of large language models (LLMs) and their deep involvement in linguistic research practice, corpus linguistics is confronting both paradigm-level challenges and new opportunities for innovation. Despite the overwhelming advantages of LLMs in terms of data scale and generative capacity, corpus linguistics—an important branch of linguistics—rests on evidential anchors that are authentic, stable, balanced, and representative. It thus offers fine-grained, dynamic, and adjustable perspectives of observation, with strong traceability of evidence, transparency of analytical procedures, and interpretability of conclusions. Consequently, corpus linguistics retains irreplaceable scholarly value in revealing distributional patterns, mechanisms of variation, and diachronic change in language, and can also serve as a critical benchmark for evaluating and calibrating the linguistic behavior of LLMs. By contrast, LLMs exhibit clear advantages in corpus expansion, annotation and processing, analytical dimensionality, and multimodal handling. Accordingly, it is necessary—on the basis of clearly delineating the paradigm differences and capability boundaries between corpus linguistics and LLM research—to promote organic collaboration through complementary strengths rather than simple substitution, thereby advancing paradigm transformation and technological innovation in corpus linguistics.

Keywords: large language models; corpus linguistics; Japanese; constructions; learner error annotation

(责任编辑: 李广悦)