语料库的尽头是什么?

詹卫东(北京大学中文系) 过去半个世纪以来,语料库在持续的规模扩张[1]中已然成为 AI 的“基础设施”,但规模更大并不必然带来更深的认知。语料库的发展方向取决于语言学的使命。若语言学以解释人脑的内在语言能力为根本任务,语料库建设就不应停留在外部语言的数量堆积,而应面向内部语言建模,将语言学洞见转化为可诊断、可训练的小规模精细化任务型语料库,并探索人机协作机制,将专家创新能力与 AI 工程技术优势相融合。

一、语言学的使命

语言学的核心使命无疑是阐释人类语言现象背后的底层逻辑。然而,对于底层逻辑的不同理解,将当代语言学划分为长期对立的两大阵营。

以乔姆斯基为首的生成语言学派将语言学的使命视为对人类心智生物属性的探寻。其演进脉络清晰:首先是以递归文法表征语言系统的无限生成性,完成了语言学技术层面的形式化转型;随后研究视野从语言表现转向对语言能力的聚焦,以“解释充分性”为语言学的最高追求;继而将“内部语言”(I-Language)确立为核心研究对象,语言学使命彻底转向人脑认知的内部语言建模;最终,生成语言学派将人类语言的普遍语法机制置于生物学与自然法则的最优设计视角下审视,语言学的使命也随之升华为在生物语言学框架内揭示人类物种特有的语言自然属性。(Chomsky 1957, 1965, 1986, 1995)跨越四十年的理论跃迁,将语言学推向终极之问:为何只有人类会说话?

功能主义、构式语法、社会语言学等非乔姆斯基阵营则认为,语言学的使命不应被禁锢在假想的孤立的语言心智模块中,而应通过语言的社会交际功能来探究其本质。不同于生成学派对先天机制的执着,非乔姆斯基阵营更侧重研究语言作为一种社会契约与交际工具的动态适应性。语言学因此被定义为解释“语言如何在使用中演化”的科学。

让语言学人尴尬的是,语言学各大门派还没争清楚语言学的圣杯到底是什么,大语言模型(Large Language Model,以下简称 LLM)就已甩出了第三种可能:人类语言或许可以脱离人而存在。靠算力与数据驱动,LLM 首次在非生物载体上涌现出类人语言行为能力。这不仅是对传统语言学使命观的冲击,更预示着语言学可能要包含对这种非生物语言智能的剖析。

二、语料库的地位

语言学不同门派关于语言学使命的观念差异,直接导致了语料库地位的天壤之别:有人视语料库为探寻语言真理的噪声,有人视语料库为揭示语言奥秘的最佳实验室。

在乔姆斯基眼中,即便语料库规模再大,它也只是有限且带有偶然性的句子集合。他反复批评把研究目标限定为从语料中抽取模式的描写主义取向,认为这种只做分类与整理的工作,会妨碍对语言机制与原则进行解释的根本追求。他也将这样的批评指向统

计式 AI 和 LLM：更多的数据与更强的统计只是让系统更好地拟合文本分布，并未能触及人类心智中作为核心生成机制的“内在语言”。（Katz 2012；Chomsky *et al.* 2023）在乔姆斯基阵营，语料库仅被视为验证形式规则的次要证据，而语言科学的主要证据则应依赖结构化的最小对比材料、母语者可接受性评价及受控实验的设计与解释。

与之相反，非乔姆斯基阵营与人工智能领域都把语料库视为语言规律的“原矿”，认为语言规律能在真实使用中自然涌现，大规模语料能揭示个体直觉难以覆盖的频率、搭配信息与语用约束。（Sinclair 1991；Tomasello 2003）LLM 通过海量语料训练获得通用语言处理能力的事实也进一步强化了这种立场：离开先验规则，仅凭大规模语料与恰当的学习算法，也能在工程上重构出高性能的语言模型。

三、解释优先与预测优先

上述语料库地位之争，其实是解释优先与预测优先两种科学文化的交锋在语言学领域的缩影。（Breiman 2001；Norvig 2011）前者更看重“知其所以然”（Know-why），后者更看重“知其然”（Know-how）。经典科学观认为，更好的解释会带来更好的预测。但在面对像自然语言这样的复杂现象时，这一信条常常失灵。语言学简洁优美的解释并不能转化为机器的预测能力。相反，基于海量语料和蛮力计算训练出来的 LLM，却涌现出惊人的语言处理能力，而其缺少科学解释的缺陷在“能干”的光环下变得无足轻重。

Know-how 先行而 Know-why 滞后，将整个科学界推到了十字路口：若捍卫解释优先，则能力有限；若向预测优先妥协，人类又难放心拥抱“黑盒 AI”。

著名的“章鱼实验”直观地表达了语言学者的担忧：能说会道的 AI 缺少把符号与世界对接的语义锚定之道，AI 的“能干”就很难安全外推到开放环境与高风险任务。（Bender *et al.* 2020）无论语料库规模如何扩展，统计相关性终究无法提供只有科学因果性才能带来的扎实的安全感。

四、由知识驱动而内外对齐

在 1991 年第 82 届诺贝尔研讨会上，菲尔墨用“拍脑袋语言学”（armchair linguistics）与“语料库语言学”（corpus linguistics）对比，幽默地调侃了两大阵营之争。（Fillmore 1992）他给出的和解方案是“计算机辅助的拍脑袋语言学”（computer-aided armchair linguistics）。

这个思路在 LLM 时代仍具指导意义。单靠海量数据很难全面反映人类灵活的认知能力。比如，LLM 不理解“面前”和“身后”在汉语的空间表达中其实可以指同一个方位，也不知道“连博士毕业生今年找教职工作都很困难”无法用于回答“博士毕业生今年找教职工作困难吗？”（詹卫东 2025）。为此，我们亟需构建高质量的精细化任务型语料库。它不同于传统的标注语料，而是在语言学知识的指导下，通过改写与合成，构建正误对立、问答句对、推理过程等形式的自然文本，把空间关系、构式用法这样的复杂认知与领域知识，融入以自然文本为载体的交互任务中。

这意味着语料库构建方法要从“数据堆叠”进化到“知识蒸馏”。具体而言，应探索“专家指导 +LLM 辅助”的协作模式。人类专家负责任务定义、知识约束与质量控制，LLM 负责任务落实与数据增强。以中介语语料为例，可先由 LLM 自动识别偏误并修改，形成“中介语 - 目标语”的初步对齐后再由专家审核，[2]从而以较低成本将凌乱的外在语言表现（E–language）转化为具有高解释价值的训练语料资源，帮助模型实现从“分布拟合”向“逻辑诊断”的能力跨越。

探讨“语料库的尽头是什么”，并非在物理极限的意义上考虑数据规模，而是追问语料库作为外部语言的集合，能否实现与人脑内部语言的真正对齐。这种显然极度复杂的对齐关系，或许可以借用素数外延与内涵的集合表征来直观呈现，详见表 1：

表 1

外延（有限遍历枚举）	{2, 3, 5, 7, 11, 13, 17, 19, 23, 29……}
内涵（基于逻辑式的定义）	$p>1 \wedge \forall a \forall b((p=a\times b)\rightarrow(a=1 \vee b=1))$
内涵（基于正则式的计算）	^(?!1$)(?!(11+?)\1+$)1+$

正如素数集合可以从无限的数字罗列压缩为一行有限的简洁算式，语言学的使命与 AI 的目标，都是要将潜在无限的“外部语言”数据，压缩为可计算、可重复验证的“内部语言”模型。因此，语料库进阶之路的本质应是知识累积，而不只是数据堆砌。从这个意义上讲，语料库的尽头，便是它作为外延数据的使命终点：它要么说明存在从观测工具到解释系统的质变途径，要么说明内涵与外延的对齐只是一厢情愿的逻辑幻梦。

附 注

[1] enTenTen 网络英语语料库 Sketch Engine 已达到 520 亿单词规模，见 https：//www.sketchengine.eu/ententen-english-corpus/。大规模网络爬虫项目 Common Crawl 自 2008 年启动以来，截至 2025 年 10 月，已累积超过 8500TB（万亿字节）的网页数据，见 https：//commoncrawl.org/。

[2] 我的博士生周子茗用 DeepSeek-V3.2 对 865 篇汉语中介语作文（420706 字）进行自动批改的实验：模型作业速度为每分钟 3392.79 字，精确率 86.29%、召回率 87.73%、F1 综合分 87.01%。计算标准以句子为单位，即模型识别出的中介语错误如果跟标准答案在同一个句子中就算成功，否则就算失败。