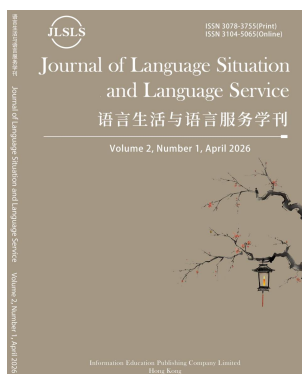




JLSLS



Journal of Language Situation and Language Service

JLSLS, Vol. 1, No.1, 2026, pp.18-30

Print ISSN: 3078-3755; Online ISSN: 3104-5065

Journal homepage: <https://www.lslsjournal.com>

DOI: <https://doi.org/10.64058/jls260102ldfeh>

语言数据要素化与我国语言产业高质量发展

——理论机制、现实瓶颈与跃迁路径

陈柏福 (Chen Baifu), 钟丽洁 (Zhong Lijie)

摘要: 数字经济时代, 语言数据正从一般性资源转化为可流通、可复用、可增值的新型生产要素, 深刻影响语言产业的组织方式与服务模式。本文基于“企业—产业—制度”的三层分析框架, 探讨语言数据要素赋能我国语言产业高质量发展的理论机制、现实瓶颈与跃迁路径。研究认为, 语言数据要素在微观层面重塑企业决策、生产与组织能力; 在中观层面促进产业链、数据链、创新链、人才链与资金链的协同; 在宏观层面支撑国家语言能力、数字治理能力与文化传播能力。当前, 我国语言数据赋能语言产业仍面临数据产权不清、流通机制不畅、质量标准不足、平台垄断风险与复合型人才短缺等现实瓶颈。未来应从企业数据能力建设、语言数据市场规范化、标准体系完善、安全伦理治理和国际化语言服务体系建设等方面推进语言产业高质量发展。

关键词: 数字经济; 语言产业; 语言数据要素; 高质量发展

作者简介: 陈柏福, 湖南衡东人, 博士(后), 广东金融学院经济贸易学院教授、文化经济研究中心主任, 南方科技大学全球城市文明典范研究院特聘研究员, 研究方向: 语言经济与文化产业、数字经济与文化出海、产业经济学与组织。邮箱: 47-175@gduf.edu.cn。钟丽洁, 通讯作者, 广东河源人, 广东金融学院经济贸易学院国际商务专业硕士研究生, 研究方向: 国际文化贸易、国际商务。邮箱: 3465986513@qq.com。

Title: Language Data Factors Empowering the High-Quality Development of China's Language Industry in the Digital Era: Theoretical Mechanisms, Realistic Bottlenecks and Leapfrog Paths

Abstract: In the digital economy era, language data is evolving from a general resource into a new production factor characterized by circulation, reusability, and value appreciation, which profoundly reshapes the organizational modes and service paradigms of the language industry. Grounded in a three-tier analytical framework of “enterprise-industry-institution”, this paper examines the theoretical mechanisms, realistic bottlenecks and leapfrog paths, through which language data elements enable the high-quality development of China's language industry. The study argues that at the micro level, language data elements reconstruct enterprises' capabilities in decision-making, production, and organizational management; at the meso level, they facilitate the synergy among industrial chains, data chains, innovation chains, talent chains, and capital chains; at the macro level, they underpin national language capacity, enhance digital governance, and strengthen cultural communication. Currently, the empowerment of the language industry by language data in China is still constrained by practical bottlenecks, including ambiguous data property rights, inefficient circulation mechanisms, inadequate quality standards, risks of platform monopoly, and a shortage of interdisciplinary talents. Moving forward, efforts should be directed toward advancing the high-quality development of the language industry through such measures as enhancing enterprises' data capabilities, standardizing the language data market, improving the standard system, strengthening security and ethical governance, and constructing an international language service system.

Keywords: Digital Economy; Language Industry; Language Data Factors; High-Quality Development

Author Biographies: **Chen Baifu**, from Hengdong, Hunan Province, holds a post-doctoral degree. He is a professor at the School of Economics and Trade, Guangdong University of Finance, and the director of the Cultural Economy Research Center. He is also a specially-appointed researcher at the Institute of Global Urban Civilization Model of Southern University of Science and Technology. His research interests include language economy and cultural industry, digital economy and cultural export, industrial economics and organization. E-mail: 47-175@gdutf.edu.cn. **Zhong Lijie**, Corresponding author, from Heyuan, Guangdong Province, is a master's student in International Business at the School of Economics and Trade, Guangdong University of Finance. Her research interests include international cultural trade and international business. E-mail: 3465986513@qq.com.

一、引言

习近平（2022）深刻指出：“数字技术作为世界科技革命和产业变革的先导力量，日益融入经济社会发展各领域全过程”，并强调“要构建以数据为关键要素的数字经济”¹。当前，数据要素已成为驱动新质生产力发展的基础性力量。根据 IDCA 研究显示，数字经济占世界名义 GDP 的 17.3%，相当于 2025 年约 119 万亿美元名义全球 GDP 中略超过 20 万亿美元²。语言数据是数字经济中具有基础支撑意义的重要数据类型之一，能有效提升全要素生产率，在经济高质量发展中起到了乘数效应（王海兰、陈柏福，2024）。在数字经济中，相当大比例的数据是以自然语言、文本、语音等形式存在的。在人工智能训练、智能交互、内容生成等核心环节，语言数据发挥着不可替代的支撑作用。数字经济时代，语言数据的重要性不言而喻：人类观察世界所形成的数据，可供计算机处理的数据，80%都是语言数据，故而语言数据是最为重要的数据（李宇明，2020）。因此，语言数据的要素化进程，构成了语言产业高质量发展的核心命题。

语言产业的研究脉络，已从早期的文化符号价值探讨，逐步拓展至经济价值与产业形态的系统性审视。在经济价值维度，学者们聚焦于产业融合效应。李宇明（2023）在回顾人类语言技术发展历程的基础上，重点探讨了语言数据管理、数据伦理学、语言大模型数据等问题。王海兰（2022）从人力资本、制度、产业、数据要素等维度阐释了语言数据的经济属性，认为语言数据既是数字技术发展的“物质基础”，又是生成用于决策的信息和知识的“原始材料”，也是新语言职业和语言产业形成和发展的“助推器”。在机制与实践上，谢孟军（2025）将语言数据嵌入内生增长模型，揭示其可通过直接参与生产或间接嵌入传统要素的方式驱动数字经济增长，且该效应存在区域与行业异质性。王建冬和童楠楠（2020）构建的“五链协同”机制为产业融合分析提供了理论框架；谢娟娟和肖姝帆（2024）从数据流动视角论证了其促进服务贸易出口的路径。张燕、黄俊杰和陈臻（2024）提出的“技术—经济范式”理论，为理解数字技术重塑产业的阶段性过程提供了分析工具。

然而，既有研究仍存在明显的理论缺口。一方面，数字经济研究多聚焦技术应用与市场效应，对语言数据作为基础性生产要素的独特作用关注不足，对其赋能全要素生产率的具体机制探讨不够深入；另一方面，语言产业研究虽关注业态转型，但未能充分整合要素化、市场化等前沿视角，与数字经济核心理论的耦合度有待提升。此外，对数据产权界定、市场流通等现实瓶颈的回应尚不够系统，导致政策设计缺乏针对性支撑。

本文立足于数字经济时代语言产业战略升级的现实诉求，旨在构建一个整合“数据要素—产业重构—高质量发展”的分析框架，重点阐释语言数据要素如何通过资源配置优化、技术融合创新、治理机制完善等路径，推动语言产业高质量发展。本文构建“企业—产业—制度”的三层分析框架，从微观、中观、宏观三个层面揭示语言数据赋能的理论机制，在审视语言数据要素助推语言产业高质量发展的现实瓶颈的基础上，提出针对性的跃迁路径与政策建议。

1 习近平. 不断做强做优做大我国数字经济[J]. 求是, 2022(2):4-8.

2 International Data Center Authority. Global Digital Economy Report 2026[R]. New York: International Data Center Authority, 2026.

二、语言数据要素赋能语言产业高质量发展的理论逻辑

语言数据要素化的核心是从被动信息载体到主动生产要素的价值跃迁。其核心价值不在于数据本身的存储，而在于通过技术加工实现语义提取、需求匹配与价值衍生。这一跃迁的最终指向，是推动语言产业从边缘性服务产业向数字经济基础性支撑产业转型。语言产业高质量发展，本质上是从资源依赖型增长向创新驱动型增长的范式转变。这一转变的核心驱动力，并非是传统的资本积累或劳动力扩张，而是在数字经济中被确立为关键生产要素的语言数据。那么，语言数据要素究竟通过何种理论机制，实现其从微观到宏观的全面赋能呢？本文从微观企业能力重塑、中观产业生态协同、宏观国家战略支撑三个层面，系统揭示其内在逻辑。

（一）微观层面：企业能力重塑

数据要素对语言产业的驱动，最终要在微观企业层面得到体现和检验。语言数据要素并非静态地存在于数据库中等待被使用，而是作为一种能动的变革性力量，来重塑语言企业的决策范式、生产函数与组织形态。

1.决策范式：从抽样推断到全景洞察

传统语言企业的决策依据，多依赖于有限样本的市场调研和滞后信号的需求反馈，本质上是一种抽样推断。数据要素的首要价值，就在于它重构了企业的决策基础。通过对全域、实时数据流的处理与分析，企业获得了对市场需求和用户行为进行持续、精细测绘的能力。

具体而言，语言服务企业能够运用自然语言处理、情感分析、主题建模等技术，对用户生成内容、交互日志、服务反馈等大规模非结构化数据进行深度挖掘和分析。这些数据中蕴含的潜在需求信号，往往比显性的问卷调查结果更为真实和及时。例如，翻译平台可以通过持续监测用户的搜索查询日志和翻译请求特征，动态识别专业领域机器翻译后编辑、跨文化场景化语言培训等新兴需求的增长态势，从而在竞争对手尚未察觉之前，提前进行服务布局与资源储备。这不仅是决策效率的提升，更是决策逻辑的根本性转变。

更为关键的是，数据要素催生了前瞻性服务这一全新的价值创造模式。通过分析用户的实时行为序列与历史数据轨迹，系统能够构建用户能力模型和需求预测模型，预测其在语言学习或跨语言沟通中可能遇到的困难与障碍。例如，商务英语学习平台可基于学习者的行业背景、职位层级和学习行为数据，智能推送与其职业发展路径高度相关的个性化学习内容。在此过程中，语言企业的角色从均质化服务的提供者，升华为支持用户个体化语言能力发展的长期战略性伙伴。企业从一次性交易的关系模式，转向价值共生模式。

2.生产模式：基于人机协同的效率倍增

在语言产业的生产领域，数据要素与人工智能技术的深度融合，催生了人机协同的新型生产模式。其核心要义在于，通过将人类专家的领域知识与机器的计算能力进行深度耦合，实现生产效率和质量的跃升。

在这一模式下，机器依托于经过大规模、高质量、领域专属语料训练的算法模型，高效处理翻译记忆匹配、文本初稿生成、基础质量校验等标准化、重复性任务。这些任务是传统语言服务生产中大量耗费人力的工作。而人类专家则得以从繁琐的流程性工作中解放出来，将核心智力资源聚焦

于机器难以替代的环节。例如，处理语言的模糊性（如一词多义的语境消歧）、复杂性（如逻辑关系复杂的法律条文翻译）与创造性（如品牌文案的文化调适、市场营销内容的创意写作），以及对术语准确性、语境适配性、风格一致性、文化敏感性进行高阶审校和最终把关。这种分工并非机器替代人，而是机器处理规模，人类来驾驭复杂的增效闭环——两者在价值链上实现了优势互补。

这种生产模式的重构，深刻地改变了语言企业的核心竞争力构成。企业的关键资产，不再是传统意义上掌握语言技能的译员个体，而是企业所积累的、具有强不可复制的和持续增值属性的高质量领域专属语料库，以及基于这些专属数据精调、优化出的专属智能生产系统。这个“语料库+专属模型”的组合，构成了企业的核心知识库和竞争力。竞争对手可以招募译员、采购通用翻译引擎，但无法在短时间内复刻一个企业在垂直领域深耕数十年所积累的、经过精细标注和结构化处理的高质量语料资产。

3.组织形态：面向柔性化与治理内嵌的调适

数据要素的流动性推动企业组织架构向更敏捷的方向演进。为快速响应数据驱动下瞬息万变的市场信号与创新机遇，传统的刚性组织边界走向模糊化，取而代之的是以任务为中心、能够动态整合内外部资源的柔性协作网络。

在组织内部，横跨各部门的数据中台或项目制团队成为打破数据孤岛的关键机制。数据中台负责全企业的数据资产管理和共享服务，使数据一处采集、多处使用成为可能。在组织外部，企业通过与高校研究机构、云计算和 AI 技术伙伴、以及律师事务所、医疗机构、跨境电商平台等垂直行业客户，建立深度耦合的数据合作生态，构建起超越自身组织边界的开放式创新共同体。

组织柔性化也凸显了内部数据治理的重要性。当数据成为企业的核心资产，其权属如何界定、安全如何保障、质量如何控制、伦理合规如何确保，便直接关系到企业的稳健运营与合法生存，甚至构成企业的核心风险敞口。因此，建立系统化、体系化的企业内部数据治理框架——明确数据在采集、存储、使用、共享、归档、销毁全生命周期内的管理规范、责任主体和审计流程——已从可有可无的合规建设，演变为企业维系其核心竞争优势、防范系统性风险的内生性制度保障。一个缺乏有效数据治理的语言企业，犹如一个没有财务制度的金融机构，蕴含着巨大风险。

（二）中观层面：产业生态协同

数据要素对语言产业的驱动并非一蹴而就，而是需要依托中观层面的生态协同机制来传导和放大。要素融合、市场配置与治理平衡，三者层层递进，共同构成价值创造—价值实现—价值保障的产业闭环。

1.基于“五链协同”的要素融合与价值创造

数据要素赋能语言产业高质量发展的核心起点，在于其与资本、劳动、技术等传统生产要素的深度融合，这一融合过程需依托系统性机制实现价值共创。从要素配置的理论逻辑来看，数据要素参与生产分配具有重要性与必要性，且需遵循马克思主义政治经济学原理，以及以按劳分配为主体、多种分配方式并存的制度原则（李政、周希祺，2020），其赋能效应的发挥并非孤立存在，而是依赖于多链条的协同联动。王建冬与童楠楠（2020）提出的“五链协同”机制，为这一融合过程分析提供了清晰的实操框架。

具体而言,围绕语言产业链的实际需求——涵盖翻译服务、本地化工程、语言教育、语言技术研发、语言数据服务等核心环节——需要整合来自多元场景和不同主体的多源异构语言数据链,包括通用语料、行业专属语料、用户交互行为数据等。这些高质量的数据链资源,为创新链提供核心的训练素材,驱动自然语言处理算法优化、多模态大模型精调、以及新型智能语言产品的创新孵化。同时,数据链所蕴含的巨大潜在价值通过市场的价格信号和竞争机制进行传导,吸引和引导风险投资、产业基金等资金链,向关键技术领域和高潜力应用赛道倾斜和集聚。而人才链——涵盖基础研究人才、技术开发人才、产品运营人才和产业管理人才——则为整个系统的持续创新和高效运转提供了根本的智力保障。

在这一协同框架中,语言数据要素既扮演着贯通全局的要素连接器角色,又发挥着价值放大器的关键功能。作为连接器,它通过自身的高流动性穿梭于各链之间,打破了传统分工下的信息壁垒,使得孤立的要素得以跨环节优化重组。作为放大器,它通过赋能某单一传统要素,产生巨大的乘数效应。例如,资深译员的经验与高质量语料库结合后,其产出效率可能数倍于仅依靠个人记忆和搜索引擎的同行。从经济学本质来看,“五链协同”的核心价值在于,通过政府、企业、用户等多元主体的复杂互动与信息共享,系统性地降低了要素配置过程中的信息不对称与交易成本,从而突破了传统生产要素的边际收益递减规律,为语言产业的价值创造注入持续而强劲的动力。

2.突破产权、定价与流通壁垒的市场化配置

通过协同机制创造出的价值,需要经由高效规范的市场体系来完成分配和实现循环增值。这是驱动产业高质量发展的关键制度环节。当前,我国语言数据要素市场面临着两大核心制度性壁垒,直接制约了其价值转化效率。

第一大壁垒是产权界定模糊。在语言数据这一客体之上,往往附着着极为复杂的权益结构,涉及用户(数据的原始生产者,享有人格性的隐私权和财产性的数据收益权)、企业(数据的加工控制者,享有经营性的数据财产权)、社会(作为公共利益的代表,享有公共数据知情权和普惠使用权)等多元主体。数据要素的非竞争性(一人使用不影响他人使用)与产权制度有效运行所要求的排他性之间存在内在的理论张力(黄少安等,2022),而语言数据所涉多元权益边界又高度交叉重叠,实践中难以清晰切割。这种产权界定的制度性模糊,直接导致两个并存的负面后果:一是“数据孤岛”——企业因担心权益受损、丧失竞争优势,宁愿将宝贵的数据资源锁在自己的服务器里也不愿对外共享;二是“数据沉睡”——企业因无法可靠评估合规的风险,即使拥有海量数据也不敢进行深度开发和产品化。二者共同构成了语言数据市场化流通的核心制度阻碍。

第二大壁垒是价值定价困难。语言数据的价值增值具有极强的场景依赖性,一组医疗领域的专业平行语料对于开发医学AI的价值,远高于同等规模的一般日常对话语料。同时,语言数据产品的生产具有典型的高固定成本、低边际成本特征,初期的采集、清洗、标注投入巨大,而一旦形成标准化产品,其复制和分发的边际成本趋近于零。传统的基于生产成本的加成定价法和基于供需关系的均衡定价法,均难以直接适配和反映其真实价值。定价标准的缺失,使得交易双方难以形成公允的价格预期,严重阻碍了其交易的规模化和常规化。

值得关注的是,我国在语言数据要素市场的特色化实践上已取得关键突破。深圳数据交易所于

2023 年成功落地了全国首单语言语料数据的场内实质闭环交易，它通过设立专门的“人工智能产品专区”，并联合产业各方发布大规模高质量训练数据集，为语料数据的标准化、规模化流通探索出了一条可信的平台化解决路径。¹这一前沿实践深刻地表明，依托权威、专业的交易枢纽，并配套以清晰的准入规则、技术标准、合规框架和安全保障方案，能够有效降低交易双方的信任成本和信息搜寻成本，激活并牵引高质量语言数据要素进入规范的交易市场。

3. 平衡创新与垄断的可持续生态治理

市场化配置机制在有效激活语言数据巨大潜在价值的同时，也因其另一面的强规模经济效应，催生了一种数字经济时代特有的新型垄断风险，对健康的市场生态构成了严峻威胁。

这种新型垄断的形成遵循着数据飞轮的逻辑：少数掌握先发优势的头部企业，凭借庞大的用户基数持续、海量地积累多模态语言交互数据；这些数据反过来又用于训练和迭代出性能更优的算法模型，提供更精准、更流畅的语言服务；更优的服务体验吸引更多用户，产生更多数据，如此形成一个自我强化的“数据—算法”正反馈循环。这种循环在构筑企业核心的同时，也树起了极高的行业准入门槛，使后来者和中小企业难以进入，最终拥有支配市场的地位。

这种市场支配地位的负面影响是多重的。它可能挤压中小企业的生存空间、抑制全行业的源头创新活力；可能使占据了数据垄断地位的企业获得操控语言服务价格、固化服务质量的能力；甚至可能引发大范围数据泄露、系统性信息不对称等社会安全隐患。传统反垄断理论以“有形产品”为核心设计，难以完全适配数据要素驱动下的新型垄断形态（杨东、臧俊恒，2020）。如何在保护数据创新激励的同时，构建科学有效的市场规制体系，是当前语言产业生态治理的关键命题。

（三）宏观层面：国家战略支撑

从宏观层面看，语言数据要素通过优化资源配置与重构生产流程，系统性地提升全要素生产率，并推动语言产业从边缘性服务产业向战略性“基础设施”演进。

1. 全要素生产率的系统性提升

数据要素之所以能够成为推动长期经济增长的关键引擎，其经济学依据在于它所具有的非竞争性特征。Jones & Tonetti（2020）指出，数据的核心经济属性在于非竞争性，即其使用不具有排他性，且边际复制成本趋近于零。语言数据不仅具备一般数据的共性特征，更因其特有的文化承载、认知基础与场景嵌入属性，在产业赋能过程中除带来效率提升外，还能促进文化传播、知识积累与信任形成，从而对全要素生产率（TFP）的影响机制更为复杂与深入。该特性使数据能够以较低成本在不同生产单元和环节中重复使用，进而通过优化资源配置、促进知识溢出、激励组合创新等途径，缓解传统有形要素的边际收益递减趋势，成为推动长期经济增长与 TFP 提升的关键因素。语言数据作为数据要素的重要组成，其要素化进程正是上述机制在产业实践中的具体体现。

数据要素对生产率的促进效应，在与通用目的技术（GPTs）的融合过程中进一步凸显。Brynjolfsson et al.（2021）提出的“生产率 J 曲线”理论表明，以数据为核心的无形资产（如专用软件、数字平台）与人工智能等 GPTs 之间存在显著的互补关系。二者在融合初期因需大量协同投

¹ 数据来源：证券时报网，<https://stcn.com/article/detail/1046428.html>

资，可能导致生产率增长暂时减缓，但一旦跨越拐点，便会引发明显的效率跃升。这一理论有助于解释语言产业当前的发展阶段：企业对语言智能技术（GPTs）与专用语料库（无形资产）的协同投入，正处于构建新型生产函数、推动 TFP 结构性提升的关键时期。

来自中国语境的实证研究为数据要素促进 TFP 提升带来了经验证据。蔡跃洲、马文君（2021）实证分析表明，数据要素的积累与有效配置能够显著推动经济高质量发展，其主要路径之一即为提升全要素生产率。该研究进一步指出，数据要素的产出弹性和赋能效应在高技术产业与知识密集型服务业中表现得尤为突出。这为本文的核心论点提供了宝贵的宏观结构层面的统计支撑：以语言数据处理、智能语言服务为核心业务的语言产业，恰是数据要素赋能效应最为显著和典型的领域之一，其高质量发展不仅体现为自身全要素生产率的跃升，更能通过深度的技术渗透与广泛的产业协同，对其他行业产生积极的正向外部溢出效应。

2. 产业属性的战略升维

在数据与智能技术的双重驱动之下，语言产业的价值逻辑和产业定位发生了演变，其功能与属性开始深度嵌合国家整体战略体系。这一过程最集中地体现在两个具有标杆意义的领域。

一是应急语言服务体系的战略化建设。进入风险社会以来，在自然灾害、公共卫生危机、地缘冲突等各类突发公共事件中，快速生成、精准翻译和广泛传播的多语种预警信息，面向特定人群的方言应急服务和特殊群体的无障碍通信服务，已构成国家应急管理体系和公共安全基础设施不可或缺的关键模块。语言产业在此过程中展现出的价值，已超越纯粹的经济交易范畴，凸显了其作为社会信任媒介与关键信息基础设施的公共品属性。

二是国家文化数字化战略中的核心支撑作用。在推动中华优秀传统文化创造性转化、创新性发展的国家战略实施进程中，大规模、高质量、多维度的中华文化语料库及其配套的智能标注、检索、分析、生成平台，构成了国家文化大数据体系的底层基座，是实现中华文化资源全景呈现、脉络系统梳理、知识深度挖掘和全球活化利用的核心技术前提。无论是古籍的数字化整理与智能标引，还是中国当代学术成果的多语种精准翻译与全球传播，无不依赖语言数据要素和语言智能技术的强力支撑。

与此同时，宏观层面的机制还体现为产业业态的整体进化方向。在语言数据要素的持续赋能下，语言产业的业态结构正加速向智能化、集群化方向演进。以可被调用的应用程序接口（API）形式输出底层自然语言处理能力的大语言模型产业，以及面向特定行业实现深度定制和优化部署的垂直领域语言智能产业，正在形成新的集群化发展态势。这些新兴业态，构成了数字经济时代新的语言基础设施。至此，语言产业的价值最终完成了根本性升华：从纯粹的市场交易和商业收益，升维至关乎总体国家安全的战略保障。

三、我国语言数据要素赋能语言产业高质量发展的现实瓶颈

在语言数据要素驱动语言产业重构的过程中，也衍生出涉及制度设计、市场治理、人才供给与国际竞争的多重矛盾，共同制约着语言产业向高质量发展阶段演进。

（一）数据产权不清与流通机制不畅

数据孤岛与数据沉睡是语言数据市场化流通和价值转化的根本障碍。这一障碍首先表现为各种类型的数据壁垒难以打破。在政企之间，政府部门和公共机构掌握着海量且高质量的公共语言数据资源，但受制于数据安全、开放规则不明和部门利益固化，这些数据的开放共享程度与产业界的迫切需求之间存在巨大落差。在企业与企业之间，出于对核心数据资产流失、竞争优势丧失的深切担忧，各市场主体普遍选择紧闭数据大门，以邻为壑，缺乏可信、有效的共享激励机制和技术方案。

更为深层的梗阻来自产权制度层面。语言数据的产权谱系极为复杂：一名用户在语言学习 APP 上发出的练习发音、留下的交互日志，其数据权益究竟应如何在用户（原始生产者，享有人格权和财产利益）、平台企业（数据控制者和加工者，享有经营权益）、以及社会公众三者之间进行合理界定和平衡？这些复杂权益的边界在现行法律和政策框架内仍属模糊地带。部分企业虽然坐拥海量数据资源，但受限于自身的技术能力不足、应用场景有限，或因担心触及合规红线而不敢进行商业化开发，造成“数据沉睡”。这种孤岛与沉睡的畸形共存，严重降低了社会中极为宝贵的语言数据要素的配置效率，使其难以在本可发挥巨大作用的领域释放应有价值。

（二）数据质量标准与治理体系不足

语言数据要实现高效的要素化流通，需以统一的质量标准和健全的治理体系为前提。然而，当前，语料采集、标注、评价的全流程缺乏公认的行业标准。在质量标准维度，语料数据的采集格式、标注精深度、内容质量评价等全流程环节，都缺乏全行业公认的、具有权威约束力的统一标准。这导致来自不同来源、不同类型的语言数据呈现出碎片化、孤岛化的技术异构状态：数据格式互不兼容，标注深度从浅层的简单文本分类到深层的复杂语义角色标注参差不齐，数据质量更是缺乏统一的合格率基准、抽样方法和验收流程。

在治理体系上，语言数据安全风险与算法伦理监管的顶层设计仍存在明显滞后问题。语言数据本身可能携带大量个人隐私和敏感信息，一旦泄露或滥用，将造成无可挽回的损害。同时，若训练所用的语料库本身即存在种族、性别、地域、文化等方面的偏见，那么在这种数据上训练出的语言模型和服务，可能会系统性地复制甚至放大这些偏见，对特定群体造成歧视和伤害。然而，针对语言数据全生命周期的分类分级安全指南、算法公平性审查规范、伦理影响评估制度等制度建设，仍处于零散起步和概念探索阶段，还未形成完整闭环。

（三）平台垄断风险与市场规制滞后

语言数据的强规模经济效应催生了产业的新型垄断风险。头部企业凭借先发优势积累海量语言数据，结合算法与算力构建技术壁垒与市场支配地位，不仅挤压中小企业生存空间、抑制行业创新活力，还可能通过数据垄断操控服务价格、固化服务质量，引发数据泄露、信息不对称等安全隐患。

这种市场力量的过度集中，可能产生一系列竞争损害。它挤压了广大中小创新主体的生存空间，抑制了颠覆性技术和差异化产品在全行业范围内的萌发。占据数据垄断地位的巨头企业，有可能获得操控细分领域服务价格、固化服务质量标准、乃至通过算法排序影响信息传播与商业话语权的能力。另外，海量语言交互数据（尤其是语音和文本）的高度集中，本身就构成极为严重的系统性用户隐私泄漏风险和国家安全信息隐患。

现有反垄断理论与规制工具以“有形产品”为核心设计，难以完全适配数据要素驱动下的新型垄断形态。如何在保护数据创新激励的同时，构建科学有效的市场规制体系，已成为当前语言产业生态治理的难题。

（四）复合型人才供给结构性短缺

产业从“语言工具服务”向“数据智能赋能”转型，亟需兼具语言学素养、数据科学能力与产业实践经验的复合型人才。但现行教育体系仍沿袭传统学科分类，语言学与数据科学、信息技术等学科融合不足，课程设置滞后于产业实际需求，缺乏针对性的跨学科实践训练。例如，语言学专业的学生普遍缺乏扎实的数学、统计学和编程能力训练，而计算机科学专业的学生则往往对语言学的核心理论和分析方法缺乏基本认知。同时，高校与产业界之间缺乏有效的产学研协同育人平台和长效机制，学生在校期间难以接触到产业一线的真实大规模数据处理任务和智能技术应用场景，导致实践能力不足。这种供需结构性错配，使企业难以有效承接数据要素赋能带来的技术转化与场景应用任务，构成制约产业向全球价值链高端攀升的短板。

（五）国际化能力与全球竞争力不足

尽管我国语言产业已形成万亿级市场体量，但在全球市场上的整体竞争力仍有待提升。一方面，语言企业随同基建、经贸项目出海的能力尚不成熟，面向“一带一路”沿线国家和地区的多语种语言服务支撑体系有待完善。另一方面，我国在 AI 训练数据版权、多语种内容生成伦理等国际数字语言治理议题上的参与度和规则制定权仍相对有限，产业实践优势尚未有效转化为制度性话语权。国际化能力的不足，制约了语言产业在全球数字经济格局中的战略地位。

四、实践路径与政策建议

围绕前述瓶颈，我们提出系统性的解决路径，力图实现理论逻辑与政策实践的有机衔接。

（一）破解数据产权与流通障碍

释放语言数据要素的巨大潜在价值，突破口在于建设一套既尊重数据特性又高效规范运行的市场化配置体系。在此，建议采用“分级、分类、分步”的确权思路。应当依据数据来源（个人、企业、公共）和加工程度（原始数据、衍生数据、数据产品），差异化地界定各方权益，形成个人享有隐私权和基础数据收益权，企业享有经加工后的数据财产权和经营权，社会享有公共数据的开放获取权的多层次，复合型产权架构。在此基础上，应加快培育和建设兼具权威性、专业性和技术先进性的国家级和区域性数据交易枢纽，借鉴深圳数据交易所在探索语言语料数据场内实质闭环交易方面的宝贵先行先试经验，尽快形成一套可复制、可推广的交易规则、定价参考、交付流程和安全保障体系，为语言数据要素安全、有序、大规模地进场流通和交易，营造良好的制度环境和市场生态。

（二）构建质量标准与治理体系

为根本解决语料市场良莠不齐和“劣币驱逐良币”的乱象，应强力推进语言数据全流程标准化体系建设。可由相关标准化主管部门联合行业协会及龙头企业，先行启动语料采集格式规范、基础标注层级指南、通用质量评价指标等基础性、通用性国家或行业标准的紧急研制。积极落实《关于

加强数字中文建设推进语言文字信息化发展的意见》中关于建设国家语言文字大数据中心和国家关键语料库的重大战略部署，以此为国家级基础设施和中立平台，率先推出一批高质量、可信赖的国家级标杆语料数据集，形成示范效应。同时，应前瞻性地研究制定语言数据分类分级安全保护指南和算法伦理审查规范，鼓励企业积极采用隐私计算（如联邦学习、多方安全计算）、数据水印、差分隐私等前沿技术手段，为语言数据市场构筑坚实的信任根基。

（三）规制平台垄断与促进融通发展

应对数据驱动的新型垄断风险，需在法律制度和市场治理层面进行工具革新。一方面，建议在《中华人民共和国反垄断法》的修订与执法实践中，积极探索将数据资产集中度、算法透明度、数据迁移成本、以及平台互操作性等新型指标，纳入数字经济领域市场支配地位认定和竞争损害评估的考量框架。可考虑引入守门人制度，对达到特定规模、掌握关键语言数据入口的超级平台，施加更加严格的数据开放、公平接入和互操作性义务。另一方面，应采取更加积极有为的产业组织政策，主动塑造大中小企业融通共生的健康发展生态。通过设立产业引导基金、税收优惠、联合研发补贴等方式，大力支持和鼓励头部企业向中小企业开放其基础算法、算力平台和一定规模的通用语料数据。引导广大中小企业聚焦特定语种、专业垂直领域或特定文化场景，走“专精特新”的差异化发展道路，最终形成大企业建共享平台、小企业强专业应用的协同分工格局。

（四）培育复合型语言数据人才

破解人才结构性短缺，需从教育体系与产教融合两端发力。顶层设计上，应依据国务院办公厅《关于全面加强新时代语言文字工作的意见》中关于加强语言产业规划研究、发展语言智能的明确战略导向，设立专项资金，优先支持有条件的高校打破现有院系壁垒，先行先试设立“语言学+数据科学”“翻译+人工智能”等深度融合的交叉学科专业或实验班。培养过程上，建立强制性的产学研协同育人机制。可借鉴德国双元制职业教育和专业硕士双导师制的经验，将企业真实的语言数据处理项目、算法开发任务和产品运营场景，作为学生课程实训、考核和毕业论文的核心内容，确保人才培养从知识本位转向能力本位。

（五）建设国际化语言服务体系

针对国际化能力不足的问题，应系统提升语言产业的全球服务能力。鼓励语言企业随同基建、经贸项目出海，为“一带一路”高质量发展提供覆盖项目全周期的语言服务。国家层面，应以“一带一路”倡议为战略牵引，通过政策性金融、信息服务、国别风险指引等方式，大力支持国内语言服务龙头企业通过设立海外分支机构、实施跨国并购、建立国际战略联盟等多种方式，在全球主要节点城市和目标市场国家构建起完善的属地化交付网络和本地化服务能力。在行业层面，应鼓励和支持我国相关研究机构、企业和标准化组织，更加主动、更加自信地登上国际舞台，积极参与到联合国教科文组织、国际标准化组织（ISO）等国际平台关于 AI 训练数据版权、多语种内容生成伦理规范等前沿议题的讨论和标准制定中去。力争将在中国大规模、多场景市场实践中验证过的成功模式和技术方案，升华为可供全球参考和采用的“中国方案”，在构建更加公平、包容、普惠的全球数字语言治理新秩序的历史进程中，贡献更多中国智慧，赢得更多制度性话语权。

五、结 语

本文立足于数字经济时代语言产业深刻转型的现实背景，聚焦语言数据从“资源”向“要素”的价值跃迁，构建了“企业—产业—制度”的三层分析框架，系统揭示了语言数据赋能语言产业高质量发展的内在逻辑。主要结论如下：

第一，语言数据要素是语言产业高质量发展的核心引擎。语言数据作为新型生产要素，通过与资本、劳动、技术等传统要素深度融合，系统性提升了全要素生产率，为产业向创新驱动转型奠定了效率基础。

第二，语言产业形态实现从工具服务向战略基础的重塑。在数据与智能技术驱动下，语言产业超越传统中介服务角色，演进为支撑数字经济、赋能文化数字化与国家治理的战略性基础设施。

第三，语言产业高质量发展是多重机制协同的系统工程。微观上，数据要素重塑企业决策、生产与组织模式；中观上，要素协同、市场配置与治理平衡构成产业生态基石；宏观上，需构建统筹数据流动与安全、衔接国内国际规则的制度体系。

本文的理论贡献在于提出了整合微观、中观、宏观三层视角的系统性分析框架，在阐释语言数据要素赋能语言产业高质量发展的理论机制的同时，系统梳理了当前的现实瓶颈，并形成了问题到路径的对应政策建议。未来研究可在以下方面进一步深化：一是构建语言产业高质量发展的可量化评价指标体系，开展系统而深入的实证研究；二是对语言数据交易、应急语言服务等新兴细分领域进行深入的案例跟踪与比较研究；三是关注国际语言数据治理规则的演化动态，为我国参与全球数字语言治理提供更具时效性的政策参考。

Funding: This research received no external funding.

Conflicts of Interest: The authors declare no conflict of interest.

ORCID

Chen Baifu ^{ID} <https://orcid.org/0000-0002-8265-5638>

Zhong Lijie ^{ID} <https://orcid.org/0009-0005-0606-5846>

References

习近平（2022）：“不断做强做优做大我国数字经济”，《求是》（02）：4-8。

[Xi Jinping (2022). Continuously Strengthen, Optimize and Expand China's Digital Economy. *Qiushi*, (02): 4-8.

Brynjolfsson, E., Rock, D., & Syverson, C. (2021). The productivity J-curve: How intangibles complement general purpose technologies. *American Economic Journal: Macroeconomics*, 13(1): 333-372.]

蔡跃洲、马文君（2021）：数据要素对高质量发展影响与数据流动制约，《数量经济技术经济研究》，38（03）：64-83。

[Cai Yuezhou, Ma Wenjun (2021). The impact of data factors on high-quality development and constraints on data flow. *The Journal of Quantitative & Technical Economics*, 38(03): 64-83. DOI: <https://doi.org/10.13653/j.cnki.jqte.2021.03.002>.]

黄少安、汪张龙、宋晖等（2022）：“语言数据与数字经济”多人谈，《语言战略研究》，（4）：61-65。

[Huang Shaoan, Wang Zhanglong, Song Hui, et al. (2022). Multi-person discussion on “language data and digital economy”. *Chinese Journal of Language Policy and Planning*, (4): 61-65.]

Jones, I. C., & Tonetti, C. (2020). Nonrivalry and the economics of data. *American Economic Review*, 110(9): 2819-2858.

李宇明 (2020): 数据时代与语言产业, 《山东师范大学学报 (社会科学版)》, 65 (05): 87-98.

[Li Yuming (2020). Language industry in the era of data. *Journal of Shandong Normal University (Social Science Edition)*, 65(05): 87-98. DOI: <https://doi.org/10.16456/j.cnki.1001-5973.2020.05.008>.]

李宇明 (2023): 人机共生时代的语言数据问题, 《华中师范大学学报 (人文社会科学版)》, 62 (05): 135-143.

[Li Yuming (2023). Language data issues in the era of human-machine symbiosis. *Journal of Central China Normal University (Humanities and Social Sciences)*, 62(05): 135-143. DOI: <https://doi.org/10.19992/j.cnki.1000-2456.2023.05.011>.]

李政、周希祺 (2020): 数据作为生产要素参与分配的政治经济学分析, 《学习与探索》, (1): 109-115.

[Li Zheng, Zhou Xizhen (2020). Political economic analysis of data participating in distribution as a production factor. *Study & Exploration*, (1): 109-115. DOI: <https://doi.org/10.3969/j.issn.1002-462X.2020.01.016>.]

王海兰 (2022): 试论语言数据的经济属性, 《语言战略研究》, (4): 26-34.

[Wang Hailan (2022). On the economic attributes of language data. *Chinese Journal of Language Policy and Planning*, (4): 26-34. DOI: <https://doi.org/10.19689/j.cnki.cn10-1361/h.20220402>.]

王海兰、陈柏福 (2024): 释放语言数据要素价值 推动新质生产力发展, 《天津师范大学学报 (社会科学版)》, (6): 55-59.

[Wang Hailan, Chen Baifu (2024). Unleashing the value of language data factors to promote the development of new quality productive forces. *Journal of Tianjin Normal University (Social Science Edition)*, (6): 55-59. DOI: <https://doi.org/10.20247/j.issn1671-1106.2024.06.015>.]

王建冬、童楠楠 (2020): 数字经济背景下数据与其他生产要素的协同联动机制研究, 《电子政务》, (3): 22-31.

[Wang Jiandong, Tong Nannan (2020). Research on the collaborative linkage mechanism between data and other production factors under the background of digital economy. *E-Government*, (3): 22-31. DOI: <https://doi.org/10.16582/j.cnki.dzzw.2020.03.003>.]

谢娟娟、肖姝帆 (2024): 数据流动与服务贸易出口——基于数据要素视角的理论与实证分析, 《南开经济研究》, (11): 46-65.

[Xie Juanjuan, Xiao Shufan (2024). Data flow and service trade export: Theoretical and empirical analysis from the perspective of data factors. *Nankai Economic Studies*, (11): 46-65. DOI: <https://doi.org/10.14116/j.nkes.2024.11.003>.]

谢孟军 (2025): 语言数据要素驱动数字经济增长研究——基于拓展内生增长模型的分析, 《上海经济研究》, (08): 15-28.

[Xie Mengjun (2025). Research on digital economic growth driven by language data factors: Analysis based on expanded endogenous growth model. *Shanghai Journal of Economics*, (08): 15-28. DOI: <https://doi.org/10.19626/j.cnki.cn31-1163/f.2025.08.010>.]

杨东、臧俊恒 (2020): 数据生产要素的竞争规制困境与突破, 《国家检察官学院学报》, (6): 143-159.

[Yang Dong, Zang Junheng (2020). Dilemma and breakthrough of competition regulation of data production factors. *Journal of National Prosecutors College*, (6): 143-159. DOI: <https://doi.org/10.3969/j.issn.1004-9428.2020.06.011>.]

张燕、黄俊杰、陈臻 (2024): 数字经济赋能制造业转型升级: 逻辑、案例与路径 —— 基于“技术—经济范式”演化的视角, 《技术经济》, 43 (11): 49-59.

[Zhang Yan, Huang Junjie, Chen Zhen (2024). Digital economy empowering the transformation and upgrading of manufacturing industry: Logic, cases and paths. *Technology Economics*, 43(11): 49-59. DOI: <https://doi.org/10.12404/j.issn.1002-980X.J24031112>.]