语言数据科学的学科图景

——“语言智能学科”理论与方法论构建(之四)

姜孟[1] 曹婷[2,3]

(1. 四川外国语大学 语言智能学院〔通识教育学院〕/语言脑科学研究中心,重庆 400031;
2. 四川外国语大学 英语学院,重庆 400031;3. 重庆师范大学 外国语学院,重庆 401331)

摘 要:数据科学是大数据催生的一个新兴交叉学科,是当今学术和产业研究的热点。国内学者提出了建立“语言数据科学”的设想,但尚缺乏对其建立的资格条件、学科属性、研究工具和学科功能等问题的深入探讨。基于数据科学的一般功能与内涵,创建语言数据科学既需要有相当成熟的语言产业作支撑,即具备产业条件,也需要语言大数据给语言研究带来极大挑战为前提,即具备学术条件。作为语言学与数据科学的交叉学科,语言数据科学兼具方法论科学(语言学中的数据科学)与实质性科学(语言学数据科学)的双重属性:它一方面自然传承其母源学科的理论与技术工具,另一方面也会发展出自己独有的研究工具体系。语言数据科学将在产业应用、学术应用和人才培养应用等领域发挥产业支持、范式支持、算料支持、学科支持与专业支持五大功能。

关键词:数据科学;语言产业;语言数据科学;学科图景;数据挖掘

中图分类号:H030 **文献标志码**:A **文章编号**:1674-6414(2026)01-0060-19

0 引言

随着互联网、云计算和物联网技术的发展,数据采集、存储、传输、处理等技术不断取得突破,人类社会已经不知不觉跨过“小数据”的疆域,进入“大数据”的天地。大数据不仅正在改变人们的工作、生活与思维模式(Mayer-Schönberger et al., 2013)同时,也正在给科学界、产业界、教育界带来深刻的影响与变革。大数据具有量大、异质、高维、多元、动态、价值密度低等新特征,特别引起了计算机科学、统计学等诸多科学领域的关注,催生了一个新的学科领域——“数据科学”(data scierce)。大数据在科学领域的深刻体现就是数据科学的

收稿日期:2025-10-15

基金项目:四川外国语大学2021年哲学社会科学重大项目“语言智能新文科体系构建及其范式创新研究”(sisuzd202104)和重庆市教委人文社科项目“面向AI知识图谱的语言具身概念智能研究”(22SKGH236)的阶段性成果

作者简介:姜孟,男,四川外国语大学语言智能学院教授,博士,博士生导师,主要从事语言智能、二语习得与认知神经语言学和语言病理学研究。

曹婷,女,重庆师范大学外国语学院讲师,博士研究生,主要从事语言病理学、语言脑科学和语言智能研究。

引用格式:姜孟,曹婷. 语言数据科学的学科图景——“语言智能学科”理论与方法论构建(之四)[J]. 外国语文,2026(1):60-78.

兴起(赵国栋 等, 2013: 285),数据科学将成为科研体系中的关键组成部分,其地位有望与化学、物理、生命科学等传统自然科学比肩。

在语言学界,一些学者第一时间洞察到了大数据正在给语言学科带来深远的影响,明确提出建立“语言数据科学”的设想,呼吁着眼语言数据的“数据共性”和“语言特性”,加强语言数据研究,清楚界定语言数据科学的发展内涵与外延,探究其功能与实现方式,建立语言数据(资源)学科及人才培养体系(李宇明 等,2022)。语言数据科学被界定为:依托信息技术与统计学,研究语言数据的类型、状态、属性及其演变规律,以揭示人类语言及语言行为的内在规律,并探索其在智慧教育、人工智能等领域的应用(胡开宝,2022)。这些学者关于语言数据科学的思想创意十分可贵,今天的一小步可能会决定明天能迈出怎样的一大步。然而,由于出于对相关研究的立意与重心的考虑,加之新学科创建的复杂性,因而这方面还有诸多问题需要探究与厘清。笔者认为,建立语言数据科学必须回答三个最基本的问题:语言数据科学建立的基本要求是什么,目前具备哪些基础与条件,语言数据科学的总体样貌应该怎样?

1　数据科学的现状、功能与内涵

“数据科学”的概念最早由美国计算科学领域的杰出学者、图灵奖获得者彼得·诺尔(Peter Naur, 1974)在其专著《计算机方法的简明调研》(*Concise Survey of Computer Methods*)中正式提出,当时给出的界定是:“数据科学是一门基于数据处理的科学。”(Naur, 1974: 22)

1.1 数据科学的现状

从1974年算起,数据科学已经历了近半个世纪的发展。朝乐门等(2021)将数据科学的发展归纳为萌芽期、快速发展期和逐步成熟期三个阶段。笔者认为这三个阶段的划分可能仁者见仁,但它们大致缩微了数据科学在理论与技术上的内涵积聚,也折射出该学科的发展进程与成熟度。

1.1.1 数据科学萌芽期(1974—2009)

这一时期,少数具有前瞻视野的计算机科学家与统计学家基于对既有学科体系的理性审视以及对未来社会趋势的深刻洞察,系统论述了构建“数据科学”新学科的客观需要和现实基础,并深入探析了这一学科的理论根基、基础技术和应用方向。主要的标志性成果与事件有:诺尔(Naur, 1974)在学术史上首次提出“数据科学”概念,强调其作为解决数据(问题)的本质,而非仅限于数据处理技术;威廉·克利夫兰(William S. Cleveland, 2001)将数据科学视作统计学的研究新方向,推动了统计学家对该领域的关注与研究;埃里克·布鲁尔(Eric Brewer, 2000)和丹尼尔·普里切特(Daniel Pritchett, 2008)相继提出的CAP

理论与 BASE 原则,标志着数据科学“现实主义思想”的兴起,并成为数据科学区别于传统“理论完美主义”数据研究的重要标志;2002 年,第一本数据科学国际学术期刊——《数据科学学报》(*The Data Science Journal*) 创刊发行;2003 年,Google 公司相继发表关于 GFS、MapReduce 与 Bigtable 三大技术的研究论文,为云计算在数据存储、处理和管理中的应用奠定了理论基础(Ghemawat et al. , 2003);杰弗里·辛顿(Geoffrey E. Hinton)等(2006)提出“深度信念网络”的概念,推动深度学习逐步成为处理非数值型数据的关键技术;2006 年,克莱夫·汉比(Clive Humby)提出了“数据是新石油”(Data is the new oil)的“数据挖掘”理念(Arthur, 2013);2007 年,吉姆·格雷(Jim Gray)提出“数据密集型科学发现”(也称“第四范式”)的概念,为数据科学研究提供了独有的研究范式(Hey et al. , 2009);2008 年,DJ. 帕蒂尔(DJ Patil)和杰夫·哈默巴赫(Jeff Hammerbacher)分别在 LinkedIn 和 Facebook 成立“数据科学社区”,提出了“数据科学家”(Data Scientist)职业岗位名称(Patil, 2011);2009 年,我国学者朱扬勇和熊赞出版著作《数据学》。此外,这一时期也伴随着 Python 语言、R 语言、Hadoop 生态系统/数据平台、Tableau 数据可视化工具等新技术的涌现,Google 流感分析等科学认知新实践的成功,以及社会对数据科学人才需求的显著增长。

1.1.2 数据科学快速发展期(2010—2014)

这一时期,研究者们主要聚焦于数据科学的学科定位、大数据的新特征、数据科学家的专业素养与职业定位等理论命题,以及非结构化信息管理、大数据运算架构、数据管控能力测评与数据治理、大数据生态图谱等技术应用论题。

在理论探讨方面,2011 年,德鲁·康威(Drew Conway)构建了著名的数据科学维恩图(The Data Science Venn Diagram),将数据科学描述为数学统计学知识、专业实务知识、黑客精神与技能的有机结合,从学理上确立了其跨学科属性,平息了长期以来对该学科定位的争论(O' Neil et al. , 2013);2013 年,IBM 在道格·拉尼(Doug Laney, 2001)定义的大数据“3V”(Volume-海量、Velocity-高速度和 Variety-多样性) 特征基础上新增了“Veracity-真实性”第四维度,即质量维度;同年,克里斯·马特曼(Chris A. Mattmann, 2013)在《自然》(*Nature*)杂志上发表论文《数据科学愿景》(A Vision for Data Science),从研究实践中的常见数据问题出发,探讨了建立数据科学的必要性及其内在逻辑,并将其引入计算机科学与技术领域的研究范畴,引发了学术界对这一方向的广泛重视;肯尼斯·库基尔(Kenneth Cukier, 2010)、帕蒂尔(2011)、托马斯·达文波特(Thomas H. Davenport)和帕蒂尔(2012)等围绕数据科学家的专业素养与团队组建等问题展开讨论,认为理想的数据科学家应兼具编程技术、统计分析与叙事传播三方面的综合能力,必将成为备受瞩目的新型职业群体。

在方法技术方面,马修·阿斯莱特(Matthew Aslett,2011)提出 NewSQL 架构理念,为海量非结构化数据实时处理提供了新的解决方案;2011 年,内森·马兹(Nathan Marz)提出了

Lambda 大数据系统参考架构,很好地解决了大数据处理可靠性和实时性之间的矛盾(Marz et al. , 2015);2013 年,加州大学伯克利分校的 AMP 实验室开发出了新的大数据处理生态系统——Spark 商业版本(Databricks),成为与 Hadoop 生态系统并驾齐驱的两个主流技术和平台之一;哈德利·威克姆(Hadley Wickham)(2014)提出了“规整数据”(Tidy Data)理念,明确了数据规整化的原则与技术路径,有效解决了非结构化数据在主流数据分析软件或工具中的兼容性问题;2014 年,美国一些数据研究机构发布了数据管理模型或数据治理框架,如 CMMI Institute 开发的“数据管理能力成熟度模型”(Data Management Maturity Model,简称 DMM 模型)。

在应用实践方面,2012 年,FirstMark 资本公司(FirstMark Capital)的马特·特鲁克(Matt Turck)首次构建了描绘大数据产业结构与发展动态的“大数据生态图谱”(Big Data Landscape),并将其开源;2010—2014 年,数据科学专业平台 Kaggle、数据科学在线学习平台 DataCamp 以及数据科学集成开发平台 Anaconda 先后创立,美国、澳大利亚、法国、英国等国家先后推出、实施了大数据相关国家公共政策和社会治理计划与工程;2012—2013 年,美国哥伦比亚大学、纽约大学和哈佛大学相继开设了“数据科学导论”(Introduction to Data Science)课程,开始了对数据科学专业建设和课程设计进行探索;数据科学与大数据标准化工作也在这一期间启动。这一时期还出版了不少公开冠以“数据科学”字样的论文或著作。

1.1.3 数据科学逐步成熟期(2015—)

这一时期,数据科学的学科本体性研究和产业应用性研究进一步拓展,学科的独立性、完整性更加显现。

在学科本体性研究方面,一些学者更加系统地探讨、梳理、概括了数据科学的核心理论、方法与技术,为这一新学科的可持续发展奠定了基础。例如,迈克尔·乔丹(Michael I. Jordan)和汤姆·米切尔(Tom M. Mitchell)(2015)在《科学》(*Science*)上发表了题为《机器学习:趋势、前景与展望》(Machine Learning: Trends, Perspectives, and Prospects)的报告,提出机器学习是人工智能与数据科学的核心,推动了智能感知、人机对话、自然语言处理、认知计算、智能机器人等核心技术在数据科学领域中的应用;科尔·努斯鲍默·纳福利克(Cole Nussbaumer Knaflic,2015)出版专著《用数据讲故事》(*Storytelling With Data*),在“数据感知”(数据可视化)基础上提出了“数据认知”(数据故事化)这一数据科学新课题;数据科学核心竞争力和社会贡献度的瓶颈问题“数据科学模型的性能与可解释性矛盾问题”也被提出。此外,IBM 公司于 2016 年推出了 IBM 数据科学体验(IBM Data Science Experience, 简称 DSX)综合平台(后更名为 Watson Studio),集成了 Jupyter Notebooks、Scala、Python、Apache Spark 等主流开源工具,为数据科学研究提供了从数据处理到模型部

署的全流程分析环境,推动了大数据采集、计算、管理和分析技术的融合应用。戴维·多诺霍(David Donoho,2017)的专题研究《数据科学的五十年》(50 Years of Data Science)系统回顾了数据科学的发展历程。约翰·凯勒赫(John D. Kelleher)等(2018)出版学术专著《数据科学》(*Data Science*),深入探讨了学科发展史、机器学习、应用场景以及数据伦理与隐私等问题;彼得·库本(Pieter Kubben)等(2019)主编的《临床数据科学基础》(*Fundamentals of Clinical Data Science*)介绍了医疗数据的采集、建模与其在临床实践中的具体应用;阿维·布鲁姆(Avrim Blum)等(2020)出版专著《数据科学基础》(*Foundations of Data Science*),详细论述了数据科学的核心数学模型和算法原理。

在产业应用性研究方面,数据科学的理论方法向农业、化学、新闻传播、信息资源、医疗卫生与供应链管理等多个领域深度渗透,催生了“专业数据科学”与“专业中的数据科学”两大发展路径(朝乐门 等, 2018;朝乐门 等, 2021)。前者代表数据科学作为通用基础学科的独立性,后者代表不同学科领域对数据科学的个性化探索与实践,需要依赖于具体领域的专业知识。其它的标志性成果还包括:Gartner 公司在 2016~2021 年间连续多次更新、升级了“数据科学成长曲线和魔力象限”技术平台;一些气象公司推出了农业数字化平台 Climate FieldView,将数据科学技术应用于农作物、土壤和气候等分析,为农作物种植提供决策支持;各国政府纷纷将大数据提升为国家战略规划内容,例如,我国政府于 2015 年出台《促进大数据发展行动纲要》、美国政府于 2016 年制定《联邦大数据研究和开发战略计划》等政策文件;数据科学实践的标准化工作以及数据隐私等数据安全性问题也取得了重要进展。

在学科专业性成长方面,数据科学专业人才培养趋于体系化,课程建设逐步成熟。标志性成果有:莫尼亚·贝克(Monya Baker, 2015)在《自然》杂志上发表论文《数据科学——产业诱惑》(Data Science: Industry Allure),引发产业界对数据科学家的广泛关注;2015 年,美国政府任命帕蒂尔担任首席数据科学家,反映了社会对数据科学人才的迫切需求。此外,美国纽约大学、华盛顿大学等众多高等院校率先开设“数据科学”专业;据 StudyPortals 官方数据统计,截至 2021 年 12 月,国外“数据科学”专业的本科、硕士、博士学位项目总数分别为 5 708、5 475 和 280 个。在我国,2016 年北京大学、中南大学和对外经济贸易大学首批获批设立“数据科学与大数据技术”专业;据中国教育在线统计,截至 2021 年,全国已有 558 所高等院校开设该专业。在学科著作出版方面,2016 年,朝乐门出版了国内首部系统论述数据科学核心理论、研究方法、关键技术和分析工具的专业著作——《数据科学》;2021 年,徐宗本等又出版了数据科学新作《数据科学:它的内容、方法、意义与发展》。

由上可见,数据科学是在综合运用计算机科学、统计学等学科的理论与方法,解决产业领域大数据实务问题的过程中,发展起来的一门具有信息技术性质的学科,它既具有鲜明

的产业应用与服务导向,又兼具方法论科学与实质性科学的性质与特征。该学科以海量数据为研究对象,运用数据统计、数据可视化、机器学习等方法技术,致力于数据加工、数据管理、数据计算、数据产品等方面的研究与开发(朝乐门, 2019)。本质上,数据科学聚集于数据价值的挖掘与转化,通过构建模型、数据分析、计算处理和学习杂糅的方法,探索从数据资源到信息、从信息到知识、再从知识到决策的价值链转换过程,从而达到对现实世界的科学认知与操控(徐宗本 等, 2021)。历经近几十年的发展,尤其是最近 20 多年的发展,数据科学在理论思想、方法工具、产业应用等方面的学科内涵已经十分丰富。作为一个新兴学科,它已达到了相当的成熟度,且仍处于发展上升期,前景不可限量。

1.2 数据科学的核心功能与内涵

数据科学作为因应大数据时代际遇而产生的一门新兴学科,主要具有以下两大核心功能与内涵。

1.2.1 产业应用功能与内涵

“产业应用”是指数据科学在产业领域的应用与实践,是数据科学产生的动因和土壤,并构成其首要的功能与内涵。数据科学区别于统计学、数学等其他学科的关键在于更强调实际应用(Saltz et al., 2017);数据科学国际专业期刊《数据科学期刊》(*Journal of Data Science*)的办刊理念也同样倡导“应用”导向。在商业应用领域,数据科学就被“直白”地界定为:借助数据分析方法将原始数据转换成具有商业价值的信息资源的完整过程(魏瑾瑞等, 2014)。数据科学属于应用驱动型的新兴学科,其理论建构远远滞后于行业实践(朝乐门 等,2021)。也就是说,数据科学是一个先有实践后有学科的学科,产业领域应用是数据科学兴起与发展的先在推动性因素,没有广阔的产业应用与服务就没有今天的数据科学,它构成了数据科学的首要核心功能与内涵,是其成为一个学科的关键条件。

在实践中,数据科学通过整合特定应用场景的海量原始数据,挖掘数据的潜在规律与特征,预测变化趋势,支持精准决策,从而提升各行业的运营效率与收益。能否抢先从海量数据中发现其深藏的价值资源,挖掘出数据“金矿”,往往决定一个企业或行业能否抢占先机,赢得优势。目前,数据科学在金融、商业、政务、医疗/健康、交通、工业、农业、生物、物流、新闻、消费等领域中广泛应用,作用主要体现在三方面:一是优化企业内部的协作机制,提高运营管理效能;二是坚持以人为本,基于客户需求提供个性化产品和服务;三是推进行业变革创新,发掘市场新需求,开展产品和服务革新,进而实现成本控制和效益提升(TalkingData, 2019)。

数据科学的产业应用功能也体现在“数据科学家”岗位的设置上。数据科学家被认为是为针对各行各业大数据现象而专门设立的一个“职业”。这涉及到两方面:一是政府对数据科学家岗位职业的官方认可,二是数据科学家岗位的知识素养与职责要求。对于前

者,以美国为例,美国国家科学基金会(National Science Foundation, 简称 NSF)于 2005 年明确提出:“NSF 需要与收集管理专员及整个社区开展协作,主动推动数据科学家的职业发展,并确保配备充足的高质量数据科学家。”(National Science Board, 2005: 48)2012 年,达文波特和帕蒂尔发表了《数据科学家:21 世纪最具吸引力的职业》(Data Scientist: The Sexiest Job of the 21st Century)一文,明确指出数据科学家已是企业争相聘用的对象(Davenport et al., 2012)。2015 年,贝克在《数据科学——产业诱惑》中指出,帕蒂尔被聘请为白宫首任数据科学家(如前文所述),表明社会对数据科学家的重要需求。对于数据科学家的职业岗位素养问题,康威(2011)将“数学与统计学知识”“领域实战技能”以及“黑客精神”(如原创性设计、批判性思考和好奇性提问的素质)视为其核心素养。除此之外,有学者认为数据科学家还应具备数据分析技术与商业洞察力等综合素养。也就是说,数据科学家既要掌握数据的获取途径、数据形态和存储方式,又要了解如何选择合适的分析手段,并能够对分析结果作出符合实际的专业解读(魏瑾瑞 等, 2014)。

1.2.2 学术应用功能与内涵

“学术应用”是指将数据科学的理论、方法、技术和工具等运用于科学研究,这是数据科学的又一核心功能与内涵。基于数据科学开展的科学研究包括“以数据方法研究科学问题”和“以科学方法研究数据问题”(Provost et al., 2013)两种不同的取向。“以数据方法研究科学问题”主要应用于天体信息学、生物信息学等领域,一个可以类比的例子是“开普勒第三定律”的发现。该发现是纯粹基于对观测数据的归纳总结,得到“行星绕太阳运转的周期平方与行星距太阳平均距离立方成正比”的结论,但开普勒本人也并不理解其内涵。这种数据驱动的研究在现代科学研究中得到系统性发展,图灵奖得主格雷(Gray, 2009)称之为“第四范式”。第四范式与“实验”“理论”、“仿真”三种科学研究范式相对。“实验”范式诞生于几千年前,以观察和实验为手段,是以描述自然现象为目的的经验科学范式,为第一范式;“理论”范式诞生于几百年前,是使用模型或归纳法进行研究的理论科学范式,为第二范式;“仿真”范式产生于几十年前,是通过计算机模拟复杂现象的仿真科学,为第三范式。第四范式被称为“数据密集型科学发现”(Data-intensive Scientific Discovery)范式,可以看作是对前三种范式的综合应用。前三种范式立足于“小数据”,即用常规的技术手段就能“应付”的数据。基于小数据,人们往往首先从一定的理论出发,提出一定的问题,然后通过小数据分析回答所提问题,即采取“理论驱动”(理论→数据→问题)的研究模式。然而,第四范式主要针对的是“大数据”。它把大数据看作是现实世界向数字世界的映射,通过运用和分析数据来发现现实世界所隐藏的科学规律。从大数据出发,人们往往采取“数据驱动”(数据→问题)的研究模式(周代数 等,2024),即在无理论先设的条件下用数据分析的结果(通常只是“信息”,还算不上“知识”)匹配现实世界中的问题。“数据驱动

科学发现”的研究逻辑被认为适用于自然科学和人文社会科学研究的各个领域。“用数据的方法研究科学”与特定的学科领域联系在一起,且“数据密集型”的研究范式服务于特定学科领域的研究,可被视为一种方法与手段。在此意义上,数据科学是一门具有“方法论”性质的学科,即“方法论科学”(与“实质性科学”相对),“数据”在其中扮演着“工具性研究对象”的角色。换一个角度看,正是由于“用数据的方法研究科学”是为特定学科领域的研究服务,它对具体领域的专业知识有很强的依赖性,也同时代表着“不同学科领域对数据科学的差异性研究和应用”。在此意义上,数据科学又被称为“专业中的数据科学”(朝乐门等, 2021)。

“用科学方法研究数据”主要应用于统计学、机器学习和数据挖掘等领域,致力于探索数据处理技术和数据内在共性的规律。数据科学的兴起既促进了利用海量数据处理问题的研究突破,又为汇聚不同学科、领域的数据研究成果创造了条件,有助于各领域攻克自身难以解决的数据问题(张清华, 2022)。这一研究路径的理论逻辑是,数字世界是现实世界的映射,其中既存在它所反映的现实世界的客观共性规律(如能量守恒定律、牛顿定律等),也存在数字世界本身的类似现实世界规律的一般性规律(如对称性、黄金分割、长尾分布等常数规律,或非确定性、数据广义关联、时空演化、数据复杂性等大数据的特征与规律)。既然数据作为现实世界在数字世界中的符号化映射,构成了数字世界的核心元素,那么通过研究数据的基本属性、存在状态、分布类型及其演化过程,挖掘其内在的模式与规律,进而阐释数字世界的运行机理,应当成为数据科学研究的根本问题(张清华 等, 2022)。可见,“用科学方法研究数据”把大数据所代表的“数字世界”看作是一个与现实物理世界相平行的独立世界,数据在其中扮演着“本体性研究对象”的角色。作为一个新兴学科,数据科学获得了自己独立的本体论研究对象、认识论研究逻辑与方法论研究手段,具有实质性学科的显著特征,也是一个与前述“方法论科学”相对的“实质性科学”。同样,换一个角度看,“用科学方法研究数据”意味着数据科学不依赖于特定学科领域的知识,是一门通用性的独立于具体应用领域的“基础科学”,属于“专业数据科学”(朝乐门 等, 2021)。专业数据科学与数据本身“打交道”,主要研究内容包括:研究数据推理的理论和方法,建立数据科学实验方法,运用实验手段和理论体系开展数据的科学探索,深入理解数据的本质特征与变化规律,进而发现自然界和人类行为的现象特征与运行规律(叶鹰 等, 2015)。

上述的“产业应用”与“学术应用”两大核心功能揭示了数据科学得以创生的需求动因与发展应用空间,表明了其学科性质、地位与价值。它们也在实质上规定了数据科学建立所必须具备的两个基本条件:一是有一定成熟度的相关产业,能为数据科学的发展提供应用需求和服务对象,可称为“产业条件”;二是现有的科学研究面临一定的方法论挑战或危机,数据科学的理论、方法与技术可为其提供迫需的补足办法与途径,即“学术条件”。由

于“语言数据科学”在学科性质上必然是“数据科学”的,其创建也需遵从以上两个基本条件。“产业条件”指必须存在有相当规模、相当成熟度的语言产业,“学术条件”指语言研究急需来自数据科学的大数据统计分析方法与技术,以应对语言大数据的现实挑战与问题。目前,这两个条件均已具备。

2 语言数据科学建立的两个基本条件

近年来,伴随着知识经济的兴起,“语言产业”在我国已经获得了相当程度的发展,达到了相当大的规模。

2.1 语言产业条件

陈鹏(2012:17)将语言产业界定为以语言为核心要素或加工对象,通过生产语言产品满足社会语言需求的产业形态。贺宏志等(2012:19)则将语言产业视为涵盖语言需求、语言市场、语言产品、语言技术等若干基本要素的活动过程和组织方式。殷志平(2021:70)则从知识经济属性出发,强调语言产业本质上是从事语言知识的生产、分配、使用和消费的产业门类。尽管学者们对语言产业的界定角度各异,见仁见智,但都指向一个最通俗的含义:语言产业就是以生产和提供语言产品为主的行业。

一般认为,语言产业包括“语言产品”和“语言产业业态”两个最基本的范畴。李宇明(2019a)系统梳理了七种语言产品形态,后新增“语言数据产品”(2019b:95),形成八大类型:(1)语言、文字及相关符号,(2)语言知识产品,(3)语言文字艺术产品,(4)语言技术产品,(5)语言医疗康复产品,(6)语言咨询培训服务,(7)语言人才,(8)语言数据产品。

语言产业业态是语言产品的外观展示,反映了某一类别语言产品的整体样貌,并由语言产品的性质与功能所塑形。贺宏志等(2012)与陈鹏(2012)根据产业功能定位,将语言产业划分为“语言能力产业”“语言内容产业”和“语言处理产业”三大类别,并将之细分为语言培训业、语言出版业等九小类业态。其中,“语言能力产业”聚焦语言能力的习得、维护和测评,涵盖语言教育培训、语言康复治疗、语言能力测评等业态,“语言内容产业”致力于语言内容的整理、复制、组合、翻译与创新,包括语言出版、语言创意、语言翻译等业态,“语言处理产业”则依托软硬件技术和设备对语言进行储存、显示、识别、转换、理解等,包括语音识别、机器翻译、智能写作等业态。“语言内容产业”处于“核心层”,“语言能力产业”处于“外围层”,“语言处理产业”则属于“相关层”(贺宏志 等, 2012:47)。语言处理产业凭借其技术的“渗透性”和“活跃性”,广泛应用于语言能力产业和语言内容产业;同时,语言能力产业和语言内容产业领域的新需求,又反向推动着语言处理技术的迭代升级,助推语言处理产业的发展。三者相互依存、协同促进,给整个语言产业不断带来活力和革命性变化。

表 1　语言产品的八大类型

序号	语言产品形态名称	举 例
1	语言、文字及相关符号	共同语、民族语、方言、外语;盲文、手语;标点符号、拼音符号、国际音标、电报代码、灯语、旗语;数学、化学符号、音乐符号等
2	语言知识产品	辞书、语言教科书、字帖、语言学讲义/著作/杂志等
3	语言文字艺术产品	小说、诗歌、楹联、歌词、戏曲唱词、话剧、相声、小品、评书等;书法、字体设计、用汉字或字母设计的各种图案等
4	语言技术产品	通过印刷、雕刻、书写而成字的传统语言技术产品;与计算机软件、计算机硬件和各种运行配件相关的现代语言技术产品等
5	语言医疗康复产品	对盲、聋、儿童语言迟缓、弱智、口吃、失语症、失读症、自闭症、老年语言退化、老年痴呆等语言疾病进行研究、诊治、矫正、康复训练、咨询、教育等语言医疗康复产品等
6	语言咨询培训服务	对前五种形态的语言产品的营销、咨询、培训等
7	语言人才	语言科学家、语言技术专家、语言艺术家、语言教师、翻译、播音员、解说员、节目主持人、话剧演员、相声演员、辞书编纂者、书法家、心理咨询专家、谈判专家、调解员、校对员、广告文案作者、导游、导购、网络主播、客户服务人员等
8	语言数据产品	机器翻译双语数据库等

除此之外,语言产业还包括一个特殊的业态,即“语言数据产业”。李宇明(2020a, 2020b)将其概括为专门从事语言数据采集存储、经营管理、处理应用的行业,涵盖数据获取、数据库建设、云端存储、智能应用、产品营销、规范标准、人才培育等一系列业态。这些业态一方面将催生语言数据相关职业,另一方面通过新业态和新职业的发展,又将生产出诸多形态的语言产品。王海兰(2022)从经济学视角对此进行了补充,只是她采用了“语言数据业”的概念,强调其市场化经营属性,即通过商业化手段生产语言数据产品或提供相关服务,以回应社会各界在语言数据领域的差异化需求。语言数据产业作为语言数据产品所塑形的语言产业业态,源于语言数据的经济属性和产业属性。语言数据作为这一新兴业态的核心要素,其本质可以理解为基于语言符号体系形成的信息资源(李宇明 等, 2022)。在整个数据生态系统中,语言数据占比最大,已成为数字化与智能化处理的重点内容。在当今信息社会,语言数据已被视为重要的信息资产和生产要素,其价值地位如同“土地之于农民,机器之于工人”(李宇明, 2020b)。语言数据的经济价值主要表现为:(1)为数字技术发展提供基础支撑,促进数字技术资本的形成与积累;(2)作为信息分析和知识生产的关键要素,有效促进资源优化配置和生产效率提升;(3)催生新语言职业与语言产业,推动社会分工和产业结构的优化调整(王海兰, 2022)。总之,语言数据已经成为语言要素参与社会生产和国民经济发展的一种新形式,催生了一种新的产业生态(姜国权 等, 2024)。

作为知识经济的一部分,语言产业以新经济为主要特征,是国家语言经济的重要支柱,对推进语言生活进步和社会进步意义重大。根据陈鹏(2016)的粗略估算,我国 2010 年语言产

业合计产值 1 920 亿,占当年全国 GDP 的 0. 47%,占当年第三产业增加值的 1. 11%。2019 年,我国语言产业全行业产值达 1 万亿元(李艳 等, 2020)。国外,瑞士语言产业对该国 GDP 的贡献高达 10%(李宇明, 2020a)。

从以上概述中可以看出,语言产业是一个庞大复杂、潜力巨大、前景广阔的产业新领域,我国的语言产业也已获得了相当的发展,“语言数据科学”的创生已经具备了坚实的产业基础和无限广阔的产业应用空间。

2.2 语言学术条件

数据是近代科学研究范式形成的基础。长期以来,在语言学中,也形成了一种以语言小数据为基础的语言研究统计范式,可称为“语言小数据统计范式”(姜孟, 2023a),这一范式广泛应用于理论语言学、计量语言学、心理语言学、神经语言学等众多语言学分支领域。它具有以下四个方面的明显特征:

第一,从数据对象看,语言小数据统计范式所涉及的是一种“以统定计”的小样本数据,即针对特定研究目标而采集的样本数据。此类数据具有目标明确、针对性强、规模有限、结构化主导、精准度高等特点。第二,从统计方法看,主要运用描述性统计方法与推断性统计方法,前者包括标准差、方差、中位数等,后者涵盖 t 检验、卡方检验、参数估计、相关分析等。第三,从理论基础来看,语言小数据统计范式基于“抽样理论”与“分布理论”,通过对研究对象抽样获取数据,对“样本”数据进行分析描述,以推断总体特征与本质规律,通常采用“总体”“样本”“统计量”“显著性”“置信水平”和“小概率事件”等核心概念,遵循“分布理论—概率保证—总体推断”的逻辑路径(李金昌, 2014)。此外,语言小数据统计范式还强调理论驱动的“因果思维”模式,侧重于语言现象背后的因果关系,通过对所收集的样本数据钻深弄透、深入解析,以“以小见大”的方式,不仅探究多种语言现象之间的关联方式(是什么),还揭示其为何如此关联(为什么)。

然而,随着人工智能与大数据时代的到来,语言活动数据也日益呈现出前述大数据泛在的“4V”(Volume、Velocity、Variety 和 Veracity)特征。语言大数据的涌现使得传统小数据统计研究范式面临前所有未有的挑战,从数据处理平台、系统到统计的理论、方法与技术,以及成果应用效果都面临极大的危机。其一,从所涉平台、系统来看,当数据规模达到 GB(10^9)或 TB(10^{12})字节级别时,传统的小数据收集、整理、加工、存储、管理技术已无法或无法在可容忍的时间内应对海量的大数据,使得前述 Spark、Hadoop、Lambda 等大数据系统或平台已经成为必需的选择。其二,从方法技术来看,传统小数据统计方法与技术已无法满足大数据分析需求。大数据的多源性、异构性、海量性使得数据处理极为复杂,须借助数据挖掘技术。除运用描述统计、概率论、时间序列分析、因子分析等经典统计方法外,还需采用关联规则挖掘、决策树、神经网络、模糊算法等人工智能方法。第三,从理论基础来看,小数据统计研究主要采用

基于抽样推断的逻辑框架,强调通过“样本”分析推断“总体”特征,两者概念相对清晰,但在大数据条件下,海量数据的可获得性使得这种推断逻辑发生根本性变化,“样本”与“总体”的界限趋于模糊。同时,“显著性检验”和“置信水平”等可靠性评价指标也失去了原有效力。一方面,海量数据导致传统统计量必然表现出“显著性”;另一方面,当大数据接近全域数据时,样本推断需求不复存在,“置信水平”便没有价值。统计假设检验的“小概率原理”同样不再适用,因为数据量大到一定规模后,小概率事件必然出现(中国人民大学统计学系数据挖掘中心, 2002)。此外,大数据强调对全域数据的直接分析,总体特征可通过数值计算获得,无需依赖分布理论进行统计推断。基于全域数据,研究者根据实际分布状况直接评估某类事件的发生概率,分析逻辑转变成“实际分布—总体特征—概率评估”,概率不再是先验假设,而是基于实际分布得出的判断(李金昌, 2014)。在技术、方法和理论的多重冲击下,小数据研究思维亟待革新。传统小数据统计范式通过样本数据,揭示语言现象背后的因果机理,侧重理论阐释,追求“纵深化”的价值发现层次。然而,大数据因其覆盖广泛、体量巨大、类型多样、变量繁杂等特征,其分析与处理需采取数据驱动的“相关思维”,让数据“说话”。研究者更侧重从海量数据中挖掘“是什么”(相关关系),而非“为什么”(因果关系),即可以不作出理论解释,追求一种“广”的价值发现层次(刘朝 等, 2021)。

因此,语言大数据给语言学带来了新问题与新挑战。语言研究迫切需要采取数据挖掘、知识发现等数据科学的思维、理念、路径、方法、技术与工具来进行方法论的革新,以促成语言统计研究的迭代变革。这为语言数据科学的创新提供了必要的学术应用条件。

3　语言数据科学的构建

基于数据科学的共性与内涵,结合语言学的学科特征和语言产业服务需求,笔者提出如下语言数据科学“学科图景”(见图 1)。

该学科图景主要聚焦以下三个核心问题:语言数据科学的学科属性是什么,用什么理论与技术工具来开展研究,主要承担哪些学科功能?

3.1 语言数据科学的学科属性

语言数据科学由语言产业和语言大数据所驱动,是语言学与数据科学的交叉学科,兼具语言学与数据科学两个母源学科的特征。

首先,从研究对象看,语言数据科学以语言大数据为主要研究对象(当然包括“语言小数据”),致力于语言数据的加工、管理、计算、产品开发等活动(李宇明,2019)。此处所说的“语言大数据”既包括学术研究导向的大数据,如李宇明等(2022)分类中的“话语数据/言语数据”或王春辉(2022)分类中的“语言功能数据”,也包括非学术研究导向的大数据,如“语言学科数据、语言衍生数据、人工语言数据”(李宇明 等, 2022),或“语言结构数据、

语言社会数据”(王春辉, 2022),或“政治语言数据、军事语言数据、经济语言数据、文化语言数据、社会语言数据、科技语言数据、网络语言数据、资源语言数据、海外利益语言数据、生物语言数据、极地语言数据、深海语言数据”等(王春辉, 2022)。其中,产业导向的语言大数据,即语言产业大数据,尤其重要。限于当前的研究现状,难以分门别类、细述这些产业数据,但大致可以说,前述语言能力产业、语言内容产业与语言处理产业所生发的数据,都属于语言产业数据(贺宏志 等, 2012; 陈鹏, 2012)。

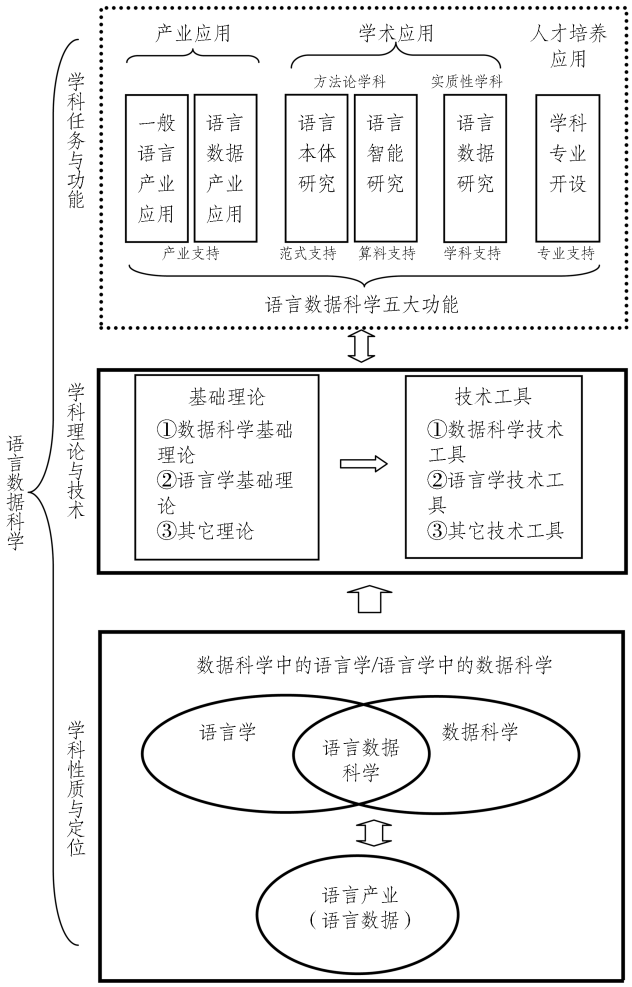

图 1　语言数据科学“学科图谱”

从与语言学的关系来看,语言数据科学是一个新兴的语言学分支学科,致力于用“数据的方法”来研究语言,属于“语言学(科学)中的数据科学”。在语言学的范畴内,采用或主要采用“数据的方法”来开展研究的分支学科早已有之,如应用语言学、语料库语言学、计量语言学、心理语言学、实验句法学(周统权 等,2025)等,它们实质为前文所述的“语言小数据统计范式”[①](姜孟, 2023a)。与此不同,语言数据科学面对的主要是语言大数据,所采

① 笔者已专文讨论“语言研究中的小数据统计范式”。

取的数据分析处理方法主要是当今数据科学中的“数据挖掘”“知识发现”等方法与技术,如规则归纳、决策树方法、人工神经网络、遗传算法、模糊技术、粗糙集方法以及可视化技术等。这些方法与技术主要基于统计学与词向量方法,语言数据科学取向下的语言研究实质上为一种“语言智能统计范式”[①](姜孟, 2023b),区别于上述语言小数据统计范式。“语言学中的数据科学”体现了语言数据科学作为“方法论科学”的学科特征。

再从与数据科学的关系来看,语言数据科学无疑也是数据科学的一个分支,它用“科学的方法”来研究“语言数据”,语言学是其“实务领域”,语言学知识构成其“领域实务知识基础”(Conway, 2011),因此,语言数据科学在此意义上又属于“专业数据科学”(与“专业中的数据科学”相对),即“语言学数据科学”(与“语言学中的数据科学”相对),体现了语言数据科学作为“实质性科学”的学科特征。

与此同时,语言数据科学还是“语言产业”驱动的新学科。如前所述,数据科学兼具“学术科研功能”与“产业服务功能”。语言数据科学也如此,它不仅以对“大、小数据”的分析处理结果来支持服务于语言学术研究,也以其为手段来支持并服务于各业态、各类别的语言产业发展与升级。因此,语言产业是它的一个特殊的“实务领域”,语言产业知识也是它的一种特殊的“领域实务知识基础”。语言产业属于非语言本体范畴的内容,而语言产业知识不同于语言本体结构知识与功能知识(即语言学科知识),它是一种经济、社会形态的非语言知识,超越了传统的语言学研究范畴。换言之,语言产业为语言数据科学的孕育与兴起,提供了动因,创造了产业应用空间,语言数据科学也因此超越了学术范畴,而较多涉入了经济产业范畴,因而获得了学术支持与产业服务双重功能。这也从一个侧面回答了为何有必要新建语言数据科学的问题;倘若只是面对语言大数据的挑战,那只需将现有的语言“小数据统计范式”拓展至语言“智能统计范式”,两者结合互补,可很好地利用大、小两种数据,充分为语言研究服务。正是由于还需为社会经济范畴的语言产业提供支持与服务,因而才必须建立一门能超越语言本体研究范畴的学科,既充分包纳语言产业领域实务知识,又具有一般数据科学特征性质,这就非“语言数据科学”莫属了。

3.2 语言数据科学的研究工具

语言数据科学作为语言学和数据科学交叉形成的新学科,自然传承了两个母源学科的研究工具,同时也将在实践中形成自己独有的研究工具。其研究工具也分为“基础理论”与“技术工具”两部分。“基础理论”部分由数据科学的基础理论、语言学的基础理论和将来在实践中形成的“其他”属于本学科的特异性理论构成。数据科学的基础理论包括统计理论、机器学习理论以及云计算理论、数据感知(可视化)理论、数据认知(故事化)理论、规整数据理论、数据管理理论、数据治理理论、大数据生态图谱、数据伦理与安全等。语言学

① 对“语言智能统计范式”,作者也已撰文专门讨论。

的基础理论则既包括传统的语音学、语义学、句法学、语用学等语言本体相关的理论,也包括计算语言学、计量语言学、语言经济学等多个交叉学科的理论。“其他”的学科特异性理论指在学科发展和解决实际领域实务(如语言产业实务)问题的过程中,提出和构建的新理论、新模型或新假设。“技术工具”部分也类似。它首先包括了数据科学的一般方法与技术工具,如 Hadoop/Spark 生态系统、Lamda 可靠性—实时性架构、Tableau 可视化工具、NewSQL 非结构化数据实时处理技术、Kaggle 数据科学专业平台、Anaconda 数据科学集成开发平台、DSX/Watson Studio 数据—模型全套分析工具等。“技术工具”不仅包括了语言学本体研究和交叉边缘学科研究所采用和发展起来的各种方法与技术工具,还包括了未来在学科发展和解决实际领域实务过程中提出和创建的“其他”方法与技术工具,这些“学科理论与技术”构成了语言数据科学的硬核性“底盘”,是学科自立的基础。学科的基本运行和功能的发挥都以此为前提。

3.3 语言数据科学的学科功能

笔者认为,语言数据科学共有“三大应用领域”和“五大学科功能”。这三大应用领域是“产业应用”“学术应用”和“人才培养应用”。“产业应用”是语言数据科学产业支持功能的体现,也是语言数据科学与一般数据科学在“领域实践”方面的共性特征,也区别于以学术科研为导向的传统语言学及其分支学科。语言数据科学的“产业应用”分为“一般语言产业应用”和“语言数据产业应用”两部分。前者是指该学科在语言产业各领域的应用,如贺宏志等(2012)学者等所界定的语言能力产业、语言内容产业、语言处理产业等领域的应用;后者则指该学科在以“语言数据产品”的消费与服务为基础,所形成的语言产业领域的应用(李宇明, 2020a)。语言数据科学在语言产业领域的应用表现为对语言产业大数据分析处理的结果,为这些产业领域的发展提供决策咨询,促进其赢得先机,提高运行效率,取得更大效益,获得更大发展。

“学术应用”是语言数据科学的又一重要而基本的功能,体现在范式支持、算料支持和学科支持三方面的功能上。“范式支持”主要针对语言本体研究,是语言数据科学为语言本体研究所提供的方法论新选择,也即吉姆·格雷(Jim Gray,2009)所称的“第四范式”“数据密集型科学范式”或“e 范式”新选择。本质上,它也就是上述“语言智能统计范式”的选择。该范式是大数据驱动,研究取向以数据挖掘的理念和方法技术为主,实质上是大数据时代对建立在抽样理论基础上的语言小数据统计研究范式的超越与革新,是语言学研究方法论的拓展与升级。

“算料支持”主要针对语言智能研究。“语言智能”就其本义而言可区分为“语言自然智能”和“语言人工智能”。前者是指人所具有的、建立在生物(遗传)基础之上的习得语言、理解语言和使用语言的高级认知机能,后者则是指对人的语言自然智能的模仿和机器

实现。语言数据科学将提供多源头、多领域、多场景的海量语言数据,为以深度人工神经网络为代表的当今主流语言智能技术提供有力的“算料”支撑,同时提升数据质量,提高算法效率,从而实现语言智能研究技术上的突破。徐宗本(2021：1967)指出：“当前人工智能技术和发展主要是靠‘算例(似应为‘算料’)、算法、算力’所驱动的, 其基础是数据, 其核心是算法。”语言数据科学为语言本体研究提供方法论支持,为语言智能研究提供的“范式”支持和“算料”支持,体现的都是该学科作为“方法论科学”的属性特征,因为它们都是把语言数据作为“工具性对象”来研究,整个语言数据科学也仅是作为一种学术方法论手段,服务于语言学及其分支领域。语言数据科学在此意义上,相当于上述“语言学中(即“专业中”)的数据科学”[与下文“语言学(即“专业”)数据科学”相对](朝乐门 等, 2021：12)。

“学科支持”功能针对“语言数据研究”本身。语言数据科学在服务于语言产业发展、语言本体研究和语言智能研究的同时,还将采取“科学的方法”来研究“数据”本身,以期能开发出语言数据分析中通用的方法、技术和工具,尤其是开发通用的语言大数据分析算法和工具。它把语言数据作为“本体性研究对象”,相当于“语言学(/专业)数据科学”[与“语言学(/专业)中的语言数据科学”相对](朝乐门 等, 2021)或“语言大数据分析学”(与“语言大数据分析”相对)(朝乐门 等, 2018),体现了语言数据科学的“实质性科学”的属性与特征。

“人才培养应用”方面,语言数据科学作为一个学科,需要设置自己的学科专业,开设相关课程,建立自己的本硕博人才培养体系,为语言学本体研究、语言智能研究、语言数据研究以及语言产业研究与发展培养合格的“语言数据科学家”或“语言数据工程师”。当今,数字技术的发展创造了海量语言数据,对语言数据的采集、清洗、标注、分析、销售等都需要专门的人才。随着数字经济的发展,将产生大量与语言数据相关的专门职业,语言数据科学专门人才将变得十分紧缺与紧俏。这一功能可以称为语言数据科学的“专业支持”功能。语言数据科学人才培养的挑战在于,正确分析语言数据科学家或语言数据工程师的岗位职责、用人需求、素质与能力要求、语言数据科学项目管理以及语言数据科学家的职业规划等。

综上所述,语言大数据时代不仅需要为语言研究提供“数据密集型科学范式”或“智能统计范式”的方法论新选择,以超越传统的语言“小数据统计范式”,更需要为语言智能研究、语言数据研究,以及“新经济”形态的语言产业发展,尤其是语言数据产业的发展,提供“数据挖掘”“知识发现”等新理念、新方法与新技术方面的支持与服务。换言之,创生语言数据科学至少有五个方面的必要性:一是语言大数据统计研究的需要,二是大数据驱动的语言智能研究的需要,三是语言产业大发展的需要,四是语言产业大数据本体研究的需要,

五是语言数据科学家/工程师人才培养的需要。

4 结语

大数据时代,数据获得了多重功能,它既可以推进科学研究的发展,也可以推进科学技术的发展,还可以推进经济社会的发展(李宇明, 2020a)。这为语言数据科学的创生提供了时代机缘。一方面,传统的语言学研究需要借用大数据挖掘、大数据智能统计的方法与范式去革故鼎新。虽然人工智能已经成为当今社会的“大明星”,但目前的语言研究成果对人工智能研究的贡献甚微,语言学在人工智能的发展中“有被边缘化的倾向”,语言研究要树立起交叉融合和数字化两大理念和意识,逐步走上数字化之路(陆俭明,2020)。这反映了语言研究在认识论与方法论导向上与时俱进不够,与大数据技术、大数据智能技术以及整个人工智能技术隔阂脱节、缺乏互动互哺的现状。另一方面,语言产业作为新经济、知识经济的一部分,正在崛起,也迫切需要语言学打破传统研究疆界、拓展新领地、开辟新内涵、旧貌换新颜。语言数据科学既以数据为“工具性”研究对象,又能以其为“本体性”研究对象,兼具学术科研支持、产业行业支撑等多方面功能,有望当仁不让地承担这一重任。

当下,数据已成为新时代重要的生产要素,上升为国家的基础性战略资源,数据科学不论是在学术研究方面,还是在产业应用方面都获得了越来越重的分量。伴随着语言产业尤其是语言数据产业的培育与发展,语言数据科学专门人才的需求将变得十分突出,这给语言数据科学的创生,提供了学术动因之外的现实产业动因与需求。学界、产业界的当务之急是,采取有力措施加强语言数据的科学研究,发展语言数据产业与职业,遵从市场机制,推动建立和完善语言数据收集、加工、交换、贮存及产权、收益等相关的技术标准、法律法规和政策体系(李宇明 等, 2022),更好发挥其作为生产要素的经济功能和社会功能,以期为语言数据科学的建立进一步创造产业应用方面的现实条件;与此同时,加强语言大数据统计理论与实践研究,以期为语言数据科学的建立进一步创造学术应用方面的现实条件。

参考文献:

Arthur, C. 2013. Tech Giants May Be Huge, but Nothing Matches Big Data [N]. *The Guardian*, 2013-02-13.

Aslett, M. 2011. How Will the Database Incumbents Respond to NoSQL and NewSQL? [R]. The 451 Group.

Baker, M. 2015. Data Science: Industry Allure [J]. *Nature* (7546): 253-255.

Blum, A., J. Hopcroft & R. Kannan. 2020. *Foundations of Data Science* [M]. Cambridge: Cambridge University Press.

Brewer, E. 2012. CAP Twelve Years Later: How the “Rules” Have Changed [J]. *Computer* (2): 23-29.

Cleveland, W. S, 2001. Data Science: An Action Plan for Expanding the Technical Areas of the Field of Statistics [J]. *International Statistical Review* (1): 21-26.

Conway, D. 2011. Data Science in the US Intelligence Community [J]. *IQT Quarterly* (4): 24-27.

Cukier, K. 2010. Data, Data Everywhere: A Special Report on Managing Information [N]. *The Economist*, 2010-02-27.

Davenport, T. H. & D. J. Patil. 2012. Data Scientist: The Sexiest Job of the 21st Century [J]. *Harvard Business Review* (10):

70-76.

Donoho, D. 2017. 50 Years of Data Science [J]. *Journal of Computational and Graphical Statistics* (4): 745-766.

Ghemawat, S., H. Gobioff & S. T. Leung. 2003. The Google File System [G]//*Proceedings of the* 19th *ACM Symposium on Operating Systems Principles* (*SOSP*' 03). Bolton Landing, New York: ACM, 29-43.

Gray, J. 2009. Jim Gray on eScience: A Transformed Scientific Method [G] // T. Hey, S. Tansley & K. Tolle. *The Fourth Paradigm: Data-Intensive Scientific Discovery*. Redmond, WA: Microsoft Research, xvii-xxxi.

Hinton, G. E., Osindero,S. & Y. W. Teh. 2006. A Fast Learning Algorithm for Deep Belief Nets [J]. *Neural Computation* (7): 1527-1554.

Jordan, M. I. & T. M. Mitchell. 2015. Machine Learning: Trends, Perspectives, and Prospects [J]. *Science* (6245): 255-260.

Kelleher, J. D. & B. Tierney. 2018. *Data Science* [M]. Cambridge, MA: MIT Press.

Knaflic C. N. 2015. *Storytelling with Data: A Data Visualization Guide for Business Professionals* [M]. New York: John Wiley & Sons Inc.

Kubben, P., Dumontier, M. & A. Dekker. 2019. *Fundamentals of Clinical Data Science* [G]. Cham: Springer Nature Switzerland AG.

Laney, D. 2001. 3D Data Management: Controlling Data Volume, Velocity, and Variety [J]. *META Group Research Note* (70): 1.

Marz, N. & J. Warren. 2015. *Big Data: Principles and Best Practices of Scalable Realtime Data Systems* [M]. Shelter Island: Manning Publications.

Mattmann, C. A. 2013. A Vision for Data Science [J]. *Nature* (7433): 473-475.

Mayer-Schönberger, V. & K. Cukier. 2013. *Big Data: A Revolution That Will Transform How We Live, Work, and Think* [M]. Boston, MA: Houghton Mifflin Harcourt.

National Science Board. 2005. *Long-Lived Digital Data Collections: Enabling Research and Education in the* 21st *Century* [R]. Arlington, VA: National Science Foundation.

Naur P. 1974. *Concise Survey of Computer Methods* [M]. New York: Petrocelli Books.

O' Neil, C. & R. Schutt. 2013. *Doing Data Science: Straight Talk from the Frontline* [M]. Sebastopol: O'Reilly Media.

Patil D. J. 2011. *Building Data Science Teams* [M]. Sebastopol: O'Reilly Media.

Pritchett, D. 2008. BASE: An Acid Alternative [J]. *ACM Queue* (3): 48-55.

Provost, F. & T. Fawcett. 2013. Data Science and Its Relationship to Big Data and Data-Driven Decision Making [J]. *Big Data* (1): 51-59.

Saltz, J. S. & J. M. Stanton. 2017. *An Introduction to Data Science* [M]. Thousand Oaks: Sage Publications.

Wickham, H. 2014. Tidy Data [J]. *Journal of Statistical Software* (10): 1-23.

TalkingData. 2019. 数据科学实战指南 [M]. 北京: 电子工业出版社.

朝乐门. 2016. 数据科学 [M]. 北京:清华大学出版社.

朝乐门. 2019. 数据科学理论与实践(第二版)[M]. 北京:清华大学出版社.

朝乐门, 邢春晓, 张勇. 2018. 数据科学研究的现状与趋势[J]. 计算机科学(1): 1-13.

朝乐门, 张晨, 孙智中. 2021. 数据科学进展:核心理论与典型实践[J]. 中国图书馆学报(1): 77-93

陈鹏. 2012. 语言产业的基本概念及要素分析[J]. 语言文字应用(3): 16-24.

陈鹏. 2016. 语言产业经济贡献度研究的若干问题[J]. 语言文字应用(3): 86-93.

贺宏志, 陈鹏. 2012. 语言产业导论 [M]. 北京: 首都师范大学出版社.

胡开宝. 2022. 语言数据科学与应用学科:特征、领域与方法[J]. 外语界(3): 37-44.

姜国权, 刘雪鸥. 2024. 数字时代语言产业的演进与思考[J]. 语言战略研究(3): 29-37.

姜孟. 2023a. 语言研究中的小数据统计范式及其人工智能变革——“语言智能学科”方法论构建(之一)[J]. 英语研究(1): 140-160.

姜孟. 2023b. 语言智能统计范式前瞻——“语言智能学科”方法论构建(之二)[J]. 外语电化教学(6): 50-56.
李金昌. 2014. 大数据与统计新思维[J]. 统计研究(1): 10-17.
李艳, 贺宏志. 2020. 大力发展语言产业 服务国家语言战略[J]. 中国教育报(3): 13-14.
李宇明. 2019. 语言产业研究的若干问题[J]. 江苏师范大学学报(哲学社会科学版)(2): 12-19+123.
李宇明. 2020a. 数据时代与语言产业[J]. 山东师范大学学报(社会科学版)(5): 87-98.
李宇明. 2020b. 语言数据是信息时代的生产要素[N]. 光明日报, 2020-07-04.
李宇明, 王春辉. 2022. 从数据到语言数据(主持人语)[J]. 语言战略研究(4): 13-14.
刘朝, 马超群. 2021. 大数据与小数据深度融合的价值与路径[N]. 人民论坛(Z1): 30-33.
陆俭明. 2020. 顺应科技发展的大趋势:语言研究必须逐步走上数字化之路[J]. 外国语(4): 2-11.
王春辉. 2022. 数字时代语言伦理的新形态和新表现[J]. 社会科学战线(12): 152-159.
王海兰. 2022. 试论语言数据的经济属性[J]. 语言战略研究(4): 26-34.
魏瑾瑞, 蒋萍. 2015. 数据科学的统计学内涵[J]. 统计研究(5): 3-9.
徐宗本. 2021. 人工智能的 10 个重大数理基础问题[J]. 中国科学: 信息科学(12): 1967-1978.
徐宗本, 唐年胜, 程学旗. 2021. 数据科学:它的内容、方法、意义与发展[M]. 北京:科学出版社.
叶鹰, 马费成. 2015. 数据科学兴起及其与信息科学的关联[J]. 情报学报(6): 575-580.
殷志平. 2021. 知识经济视角下语言产业的内涵和外延[J]. 语言战略研究(1): 68-76.
张清华, 高渝, 申秋萍. 2022. 数据科学:从数字世界到数智世界[J]. 数据采集与处理(3): 471-487.
赵国栋, 易欢欢, 糜万军, 鄂维. 2013. 大数据时代的历史机遇——产业变革与数据科学[M]. 北京:清华大学出版社.
中国人民大学统计学系数据挖掘. 2002. 统计学与数据挖掘[J]. 统计与信息论坛(1): 4-9.
周代数, 魏杉汀. 2024. 人工智能驱动的科学研究第五范式:演进、机制与影响[J]. 中国科技论坛(12): 97-107.
周统权,李雨璐. 2025. 句法的实验语言学研究:目标、方法与成就[J]. 外语跨学科研究(1):56-68.
朱扬勇, 熊赞. 2009. 数据学[M]. 上海:复旦大学出版社.

On the Landscape of Linguistic Data Science: Theoretical and Methodological Construction Series for the Newborn Discipline of Language Intelligence (Ⅳ)

JIANG Meng CAO Ting

Abstract: Data science, as an emerging interdisciplinary discipline spawned by big data, has become a hotspot in today's academic and industrial research. Domestic scholars have proposed the idea of establishing Linguistic Data Science, yet in-depth discussions remain lacking on issues such as its prerequisite requirements involved, disciplinary nature, research tools, and disciplinary functions. Based on the general functions and connotation of data science, the establishment of Linguistic Data Science requires two prerequisites: on one hand, a sufficiently mature language industry already in existence (i. e., Industrial Precondition); on the other hand, language big data posing challenges to linguistic research (i. e., Academic Prcondition). As an interdisciplinary field integrating linguistics and data science, Linguistic Data Science embodies dual attributes: methodological science ("data science in linguistics") and substantive science ("linguistic data science"). It naturally inherits theoretical and technical tools from its two parent disciplines while developing its own unique system of research tools. In the domains of industrial application, academic research, and talent cultivation, Linguistic Data Science will hopefully fulfill five key functions: industrial support, research paradigm support, computational material support, language data research support, and undergraduate major support.

Key words: data science; language industry; Linguistic Data Science; landscape of the discipline; data mining

责任编校:蒋勇军