基于大数据的语言治理研究：内涵、方法与应用*

郭书谏， 李晓阳
（同济大学 语言规划与全球治理研究中心，上海 200092）

［摘 要］ 以ChatGPT为代表的AI语言大模型和大数据的迅速发展，给语言治理带来了新的挑战，同时也为中国自主的语言治理理论体系建构提供了重要契机。本文回顾了语言规划理论发展史，评述国外语言规划理论的局限性和未来“数据转向”的可能性，论证了大数据和语言治理的关系，界定了基于大数据语言治理的内涵和方法论依据，探讨了语言大数据在不同领域语言治理的应用前景，提出在构建中国语言政策与规划知识体系的过程中，应充分运用大数据优势，避免国外语言治理理论的窠臼，推进语言治理研究的数据转向，充分阐释中国的语言治理和社会治理成效，深刻体现中国式现代化的科学内涵。

［关键词］ 大数据；语言教学；语言治理；人工智能；ChatGPT
［中图分类号］H0－0 ［文献标识码］A ［文章编号］1000－5110(2024)01－0046－08

引 言

国外语言规划实践可以追溯到20世纪初，民族国家开始认识到语言作为社会凝聚力和国家建设纽带的重要性。早期政策旨在推广标准化的国家语言，以法国围绕法语的语言政策①以及许多欧洲宗主国在殖民地推行的语言标准化政策为代表。20世纪50年代，埃纳尔·豪根（Einar Haugen）提出的语言规划理论倡导官方标准化语言形式的重要性和方法论。② 但随着20世纪70年代世界各国濒危语言问题日益凸显，以约书亚·费什曼（Joshua Fishman）为代表的学者提出“扭转语言转用”的概念体系，③如何在语言规划中更加注重社会公平成为重要的考虑因素。

20世纪90年代以来，国外语言规划的批判性方法挑战了当时语言政策的主导叙述和意识形态。这一视角强调语言规划决策的社会、政治和经济影响。伯纳德·斯波尔斯基（Bernard Spolsky）④和埃拉娜·肖哈米（Elana Shohamy）⑤等学者对语言政策中固有的不平等权力关系进行了批判性研究，并

* ［作者简介］郭书谏，男，湖北宜昌人，同济大学助理教授，博士，研究方向为语言数据、AIGC；李晓阳，女，山东潍坊人，同济大学在读博士研究生，研究方向为语言规划。

［基金项目］国家社会科学基金项目“全球治理视域下国家语言能力评价指标体系研究”（20BYY061）；国家语委重点项目“语言安全关键问题研究”（ZDI145-11）；教育部人文社会科学研究青年基金“人工智能生成语言的治理策略和路径研究”（23YJC740016）。

① Spolsky B. Language policy in French colonies and after independence[J]. Current Issues in Language Planning, 2018:1～85.

② Haugen E. Dialect, Language, Nation[J]. American Anthropologist,1966,(68).

③ Fishman J. Reversing language shift: Theoretical and empirical foundations of assistance to threatened languages[M]. Multilingual Matters, 1991:381～420.

④ Spolsky B. Language policy[M]. Cambridge University Press, 2004:113～132.

⑤ Shohamy E. Language policy: Hidden agendas and new approaches[M]. Routledge, 2006:137～166.

倡导更具包容性和互动性的语言规划过程。批判的语言规划理论还强调多语和少数族裔的语言权利，如詹姆斯·托勒弗森(James Tollefson)[①]等学者关于“语言权”概念的提出。

国外语言规划理论 70 余年的发展体现了从宏观到微观的脉络，微观语言规划和宏观语言规划既相互独立，又相互补充，相辅相成[②]，涵盖了广泛的主题和多元的主体，从早期的国家政府的语言规划，到当前“剥洋葱”式的多层级、不同族群、不同社区、不同领域的语言规划。如南希·霍恩伯格(Nancy H. Hornberger)等广泛探讨了教育领域的语言政策，[③]进一步充实了罗伯特·库伯(Robert Cooper)提出的语言教育规划的理论框架[④]。语言教育规划(Language Education Planning)作为语言治理的重要范畴，其传统研究涵盖了课程开发、教师培训、评估以及国家和机构层面的语言教育政策的设置和实施过程等主题。

国外语言规划理论自 20 世纪 90 年代的“批判范式转向”(Critical Turn)以来，已有近 30 年的发展历史。尽管当时国外学者反思了语言规划在知识建构和权力关系等方面的问题并表达了对语言生态破坏的忧虑，但是迄今国外语言规划理论并未提出合理的解决路径。语言公平和语言权利仍然只是学术研究的“乌托邦”式蓝图，缺乏具体的、技术的实施路径和范式转换。与此同时，过去 30 年是人类技术进步的重要时期，国外语言规划从理论到实践均没能关注到大数据和人工智能对人类语言的关键影响，语言规划理论存在“数据转向”的可能性，[⑤]因此有必要阐释清楚语言治理和大数据的关系。在建构语言治理的中国本土理论过程中，语言大数据应该成为重要的资源和范式。

国内的语言规划研究注重对语言资源的保护、开发和利用。国家通用语在促进国家发展、民族团结和文化认同方面发挥着重要作用，因此要完善法律法规做好国家通用语的推广。[⑥] 对濒危语言的定义、现状、意义和策略等方面的研究也为濒危语言保护提供了理论支持和实践指导。[⑦] 相较于国外“微观化”和“批判化”研究的抽象探讨，国内语言规划研究推动了语言文字事业的高质量发展。在人工智能时代，语言治理研究如何充分运用语言大数据资源，提升国家语言能力建设，服务于国家治理体系和治理能力现代化，是本研究重点探讨的问题。

一、基于大数据的语言治理的方法论

基于大数据的语言治理从理论层面来看，具有方法论层面的可行性。在科学研究中，大数据在多学科中日益发挥更加重要的作用；在语言学研究中，通过大数据方法研究语言演变、语言消亡和语言类型等也产生了一系列成果。未来语言治理研究应用大数据方法具有方法论的可行性。

(一)作为科学研究第四范式的大数据

大数据是近年来一系列重要科学进展的重要基础，以人工智能、机器学习乃至以 ChatGPT 为代表的语言大模型，都基于海量大数据的计算和建模。“随着大数据和深度学习的应用、计算能力的提升、网

① Tollefson J. W. Planning language, planning inequality: Language policy in the community[M]. Routledge, 2006:167.

② 张蔚磊，王辉.微观语言规划理论及其对我国外语教育规划的启示[J].外语研究，2022，(1).

③ Ricento T.K., Hornberger N.H. Unpeeling the Onion: Language Planning and Policy and the ELT Professional[J]. TESOL Quarterly, 1996,(3).

④ Cooper R.L. Language planning and social change[M]. Cambridge:Cambridge University Press, 1989:157～163.

⑤ 沈骑，刘思琪.数智时代语言规划研究的范式转换与方法创新[J].外语与外语教学，2022，(6).

⑥ 李宇明.中国语言规划论集[M].北京：商务印书馆，2019.

⑦ 曹志耘.中国语言资源保护工程的定位、目标与任务[J].语言文字应用，2015，(4).

络的发展,如今的人工智能研究不仅仅是信息学科的研究范畴,而是与网络科学、数据科学、语言学、心理学、神经科学等多学科紧密相关。"①

有学者将大数据视为科学研究的第四种范式,也就是所谓的数据密集型研究(data-intensive scientific discovery)②。科学研究的第一范式是实验,对关键因素进行描述和记录,形成相应理论;由于很多实验无法进行,理论推演成了第二范式,以相对论等物理学理论为代表,这些理论并非诞生于实验室,而是通过理论推演得来,后经观测证实;得益于计算机科学的发展,计算机模拟仿真的第三研究范式应运而生;由于数据的爆炸性增长,大数据研究被称为第四科学范式。

前 3 种范式的语言学研究当前比较常见,例如以实验语音学为代表的实验范式,以句法学为代表的理论推演范式,以及通过贝叶斯仿真方法研究汉藏语发源地。③ 大数据作为第四种范式为语言学研究带来了新的方法,推动了语言学研究的创新和进步。通过大数据的应用,我们可以更好地理解语言和语言生活的复杂性,探索语言与人类社会、经济、文化等方面的关系,为数据驱动的跨学科创新提供更广阔的空间。

(二)语言学研究与大数据

语言学领域当前关涉大数据的研究大致可以分为 3 类。第一类研究探讨大数据对语言学研究的方法论意义。文旭认为"大数据时代三大转变在认知语言学研究上会大大地改变我们的本体论、认识论和方法论"。④ 梁茂成认为"以深度学习为代表的大数据方法将突破语料库容量扩大带来的方法瓶颈。"⑤第二类研究关注到自然语言处理(NLP)中大数据的作用,进而从语言学理论出发,指出具有社会场景化的语言大数据对 NLP 的应用前景。⑥ 第三类研究面向具体的应用语言学问题开展,如相关研究通过大数据探讨"一带一路"沿线国家对中文学习的关注度。⑦

作为一项概念的提出,语言大数据学理上存在几个基本内涵问题尚待解决。首先,大数据和传统的语言研究中的语料数据存在哪些区别,只是单纯数据规模的扩大吗? 第二,学界缺乏语言和大数据之间的关系论述,大数据作为一种新的方法和范式,和传统意义上收集语料建立语料库开展研究,方法上的创新体现在何处?

大数据概念的提出最早源自数据科学,其主要特性概括为 4V,即规模大(volume)、种类多(variety)、价值低(value)和存取速度快(velocity)。但在各类语言研究的文献中,学者们更多关注数据的规模,而忽视了大数据的首要特征在于数据维度高。"维度"(dimension)指的是数据的特征值,"诸多领域产生了大量的高维数据,例如基因数据、天体物理数据、图像数据,等等。这些数据有一个共同的特点是样本的维数(特征)远远大于样本个数,即特征要素和样本量可能都趋于无穷大的增长"。⑧ 由于数据规模大且类型复杂,在这种情况下传统的统计学和线性分析方法(如相关系数、回归等)难以刻画数据内部的复杂关系,因此产生了一系列新的算法来处理高维度的大规模数据。

---

① 蔡三发,王倩,沈阳.人工智能赋能:高校学科建设的创新与发展——访中国工程院院士陈杰教授[J].电化教育研究,2020,(2).

② Hey A. J., Tansley S., Tolle K. M. The fourth paradigm: data-intensive scientific discovery[M]. Redmond:Microsoft research Redmond, 2009:1.

③ Zhang M., Yan S., Pan W., etc.Phylogenetic evidence for Sino-Tibetan origin in northern China in the Late Neolithic[J].Nature,2019,(569).

④ 文旭.大数据时代的认知语言学[C].第四届全国认知语言学与二语习得学术研讨会论文集,2014:1.

⑤ 梁茂成.大数据时代的语料库语言学研究探索[J].中国外语,2021,(1).

⑥ 徐大明.语言学理论对自然语言处理的影响和作用[J].云南师范大学学报(哲学社会科学版),2017,(3).

⑦ 何山华,杨晓春.基于大数据的"一带一路"沿线国家中文学习关注度研究[J].云南师范大学学报(哲学社会科学版),2022,(5).

⑧ 梁吉业,冯晨娇,宋鹏.大数据相关分析综述[J].计算机学报,2016,(1).

表 1　大数据方法和统计学方法在语言学研究中的不同

差异项	大数据方法	统计学方法
数据结构	特征值较多（维度较高）	特征值较少（维度较低）
数学方法	关联分析、聚类分析、神经网络	假设检验、回归分析
理论关联	由数据推导结论	由理论提出假设加以验证
变量关系	通常是非线性关系	通常是线性关系
潜在问题	过拟合或者欠拟合	弃真错误或取伪错误

因此大数据不单指语料规模的扩大，同时也是数据结构和方法的不同，可以认为是一种研究范式的转变。① 如表 1 所示。大数据的数学方法在过去几十年间经历了长足的发展，产生了包括关联分析、聚类分析、神经网络等各种类型的方法，②其中每一类又有不同的具体算法，如聚类算法常见的就包括决策树、神经网络、贝叶斯分类器等，通过这些算法从大规模、高维度的复杂数据中训练人工智能模型，发现有价值的信息③。相关实证研究证明，大数据范式对于语言本体和应用研究具有可能性和适用性。④

二、基于大数据的语言治理研究内涵

基于大数据的语言治理研究具有三重内涵：新的领域（语言数据参与社会治理）、新的方法（以关联分析、聚类分析、神经网络等为代表的大数据算法）和新的对象（人工智能生成语言）。语言治理的“数据转向”具备十分丰富的理论内涵和实践前景。

（一）新的领域：语言数据参与社会治理

语言数据可以作为大数据集合中的一个或多个维度，致力于解决社会治理的具体问题。如在中国家庭追踪调查⑤中，有十余个语言相关的维度，这些维度可以参与到特定问题的大数据研究中。从应用角度来看，既然语言数据的变量和其他变量能够构成相关性，语言数据参与到社会治理就存在可能性。比如在国家安全理论和实践中，研究者可以利用深度学习，建立国家安全和特定语言大数据特征值之间的关系。⑥ 因此，当语言数据被纳入社会经济大数据的挖掘过程，凭借其丰富的信息特征，应用前景广阔，比如可以通过社交媒体大数据的信息抽取，进行反恐预防和舆情分析。⑦ 一些研究也通过语言数据和医疗数据关联，探索老年人阿尔兹海默症和语言使用的相关提示表征。⑧ 可以预见，未来语言数据能够与特定社会经济治理实践结合起来，共同服务于心理干预⑨、社会治安、民生问题、反恐反诈⑩等领域。

① 刘海涛，郑国锋.大数据时代语言学理论研究的路径与意义[J].当代外语研究，2021，(2).

② 何清，李宁，罗文娟等.大数据下的机器学习算法综述[J].模式识别与人工智能，2014，(4).

③ 周志华.机器学习[M].北京：清华大学出版社，2016：73～246.

④ Abrams D. M.，Strogatz S. H. Modelling the dynamics of language death[J]. Nature，2003，(424)；Futrell R.，Mahowald K.，Gibson E. Large-scale evidence of dependency length minimization in 37 languages[J]. Proceedings of the National Academy of Sciences，2015，(33).

⑤ Xie Y.，Hu J.An introduction to the China family panel studies (CFPS)[J].Chinese sociological review，2014，(1).

⑥ 郭璇，吴文辉，肖治庭，等.基于深度学习和公开来源信息的反恐情报挖掘[J].情报理论与实践，2017，(9).

⑦ 秦颖.中外语言技术开发应用现状与展望[J].云南师范大学学报（哲学社会科学版），2016，(2).

⑧ Vigo I.，Coelho L.，Reis S. Speech-and Language-Based Classification of Alzheimer's Disease：A Systematic Review[J]. Bioengineering (Basel). 2022，(1).

⑨ Coppersmith G.，Leary R.，Crutchley P.，etc. Natural language processing of social media as screening for suicide risk[J]. Biomedical informatics insights，2018，(10).

⑩ Pelzer R. Policing of terrorism using data from social media[J]. European Journal for Security Research，2018，(2).

(二)新的方法:以关联分析、聚类分析、神经网络等为代表的大数据算法

以关联分析、聚类分析、神经网络等为代表的大数据算法,可以为语言治理研究带来新的方法。语言数据自身具备大规模、高维度的特征。语言本体研究划分为语音、词汇、句法、语义等各个层面,每个层面具备不同的特征值。以语义为例,以 Word2vec 为代表的词向量算法将语料转换为高维的数学向量,①自然语言的语义空间可以转化为可计算的高维向量空间,不同语种的数据可以进行语义相似度计算和聚类分析。

表 2 大数据语言治理的方法及其优势对比

语言治理问题	传统方法	大数据方法	优势
濒危语言保护	编撰方言志、设立语料库	训练濒危语种的语言模型	具备一定生成能力和对话能力
语言内容治理	开展培训课程和宣传活动	监测互联网语言内容,及时识别和过滤	干预更精准更及时
城市语言管理和服务	公共场所设立翻译岗位,志愿者服务	信息流再造,多语种人机交互	语言交互数据全流程可追踪复盘

作为大数据方法,对传统意义上的语言治理问题,具有方法层面的先进性,如表 2 所示。长期以来濒危语言保护都是语言治理的关键议题之一,早期以编撰方言志为主,随着语料库技术的成熟,不少濒危语言得以建库存档。但传统方法只能将濒危语言存档成为"博物馆语言",基于大数据算法训练的语言模型能够具备一定的生成能力。在语言内容治理领域,过去通常以培训课程和宣传活动为主,基于大数据的监测算法对于互联网语言内容,能够精准捕捉、及时识别和过滤,从而实现及时干预和纠正。在城市语言管理和服务过程中,相较于传统人工方法,大数据方法基于信息流再造和人机交互,能够将语言交互全流程追踪复盘,更好地服务城市管理和决策。

传统的语言治理活动大多是分散的、基于人工的、不留痕的过程,相比而言大数据语言治理方法,核心价值在于能够将语言治理"数字化",进一步产生语言治理的过程数据,实现语言数据服务国家治理和城市治理的协同效应。例如,当城市语言服务数据显示:阿拉伯语在机场、政务场所等场域的翻译服务和终端交互频次大量增加时,这一数据变化提示来自阿语地区的来访者大量增加,中阿经贸往来在未来一段时间将持续增长,从而为国家的外汇金融政策提供先行指标。

(三)新的对象:人工智能生成语言

语言不再是人类独有的产物,也可能是大数据建模后人工智能生成的产物(以 ChatGPT 为代表)。在数智时代,语言规划的主体不仅是对人的语言使用和行为进行规划,未来也可能需要对人工智能生成语言内容(AIGC)进行规划,使其符合社会伦理和价值观。语言治理主体的拓展,反映了人工智能时代语言治理研究的复杂性和艰巨性。传统的治理路径,如语言教育规划、话语规划、舆情规划可能"失灵"。如 ChatGPT-4 在中文处理方面存在预数据质量和数量不足,新知识缺乏以及中文对话系统局限等问题。② 传统意义上对人的语言规划可以通过教学、教材、话语引导等方式进行,但对于人工智能的"黑箱",可解释性尚待讨论,如何进行 AI 语言的规划仍然是一个较为复杂的问题。

自 20 世纪 90 年代以来,国外语言治理从理论到实践逐步展现出"批判转向"的态势,语言的教育公平、性别公平和南北公平等成为关键议题。语言治理的目标从二战后致力于语言使用的规范化,逐步向着促进社会公平的方向发展。相关语言治理理论日益"情境化"和"微观化",忽视了新技术发展带来的变化。然而缺乏语言普查大数据的支持和社会经济大数据的联动,微观层面的语言治理难以实现,国外

① Mikolov T., Chen K., Corrado G., etc.Efficient estimation of word representations in vector space[J].arXiv preprint arXiv:1301,3781,2013.

② 袁義,吴应辉.ChatGPT Plus 给国际中文教育带来的机遇、风险及应对策略[J].云南师范大学学报(对外汉语教学与研究版),2023,(3).

理论诉求也只能沦为“乌托邦”式的蓝图。大数据时代，未来研究应该具备问题导向和实用主义。回归费什曼的传统，在他提出的“分级代际传递严重度”（Graded Intergenerational Disruption Scale，简称GIDS）[①]中，通过调查详细量化了不同语言的濒危程度。基于大规模数据的量化评估，我们能够更为有效地监测、复现和评估具体问题，推动语言大数据在社会治理中发挥积极作用。

三、基于大数据的语言治理应用领域

基于大数据的语言治理在国家政府、企业、语言教育教学和学术研究四大领域具有广阔的应用前景。核心在于发挥语言的数据资源价值，关键在于重视对语言大数据的生产、建设、开发、利用的全生命周期的治理，确保语言数据合理规范使用，发挥其社会价值，避免数据安全风险。

（一）国家和政府的语言治理

大数据时代，国家政府领域的语言治理具有以下双重特性。第一，政策导向上，不完全以国外语言规划特别是“批判的”研究范式为参照，亦非致力于满足每个社群的语言诉求，以实现“微观化”为研究目标。相反，我们以语言大数据乃至社会经济普查大数据为基础，宏观地、整体地、全面地把握国家语情，整合语言社会经济大数据，为制定语言治理决策提供科学依据。

第二，具体实践中将语言数据视为资源，通过立法框架，加强顶层设计，推行分类治理，建立权责机制，不断应用语言大数据提升语言治理乃至社会治理水平，促进语言智能的实现。过去 10 年大数据立法已取得了长足进步。《中华人民共和国数据安全法》规定，“数据安全，是指通过采取必要措施，确保数据处于有效保护和合法利用的状态。”该法律将数据的使用纳入了法律监管之中，未经授权通过爬虫收集用户社交媒体中的语言数据等被认定为非法行为。《中华人民共和国个人信息保护法》第 28 条将个人信息划分为敏感信息和非敏感信息，确立了不同的治理手段和标准。

总而言之，国家和政府的语言治理长期以来主要关注不同语种的和谐发展和语言失范的风险防范。随着大数据时代的到来，语言作为数据要素，应该更多地参与到社会治理的具体问题中，在社情民意、舆论建设、文化传播之中发挥更多作用。以上作用的发挥需要两大基础：一方面需要更加系统的语言大数据基础设施建设，让语言大数据能够同社会经济数据联动；另一方面要进一步加强大数据立法，特别是将语言数据纳入立法体系之中，规范语言数据的生产、使用和应用。

（二）企业等商业组织的语言治理

企业等商业组织是语言治理数据化的重要参与者和责任主体，由于大数据时代语言数据的生产主要以社交媒体应用、互联网信息服务和 AIGC 为主，企业应在国家数据治理的法治框架下，进一步完善具体的语言数据治理方案。与其它类型大数据相比，语言大数据大致上是由人通过计算机中介交流（以智能手机应用为代表）产生，负载了较多与个人有关的信息。基于语言大数据的挖掘，能够刻画用户画像和行为模式。在法治框架下，须明确哪些语言数据是私人信息应该受到隐私保护，哪些语言数据可以用于商业化开发和应用，未来应建立相关的行业标准以进一步加强规范。

以 ChatGPT 为代表的语言大模型未来将生成大量类似人类自然语言的内容。其内容是否包含虚假信息？是否符合社会文化价值观？是否能够规避历史虚无主义等价值问题？一方面要求各高科技企业在模型训练中，强化人工标注监督和对抗机器学习，让 AI 模型能够规避特定的话语输出；另一方面，AI 生成语言应该进行用户识别和版本限制，比如针对未成年人学生等群体，开发教育版或设置相应权限，以避免技术不成熟对语言教育过程带来的负面影响。

（三）教育教学中的语言治理

随着 ChatGPT 等生成式人工智能的发展，未来大数据、人工智能等应用在语言教育中将扮演重要角色。回顾语言教学技术的发展历程，大致可以分为 3 个时期。从 20 世纪八九十年代提出的计算机辅助教学，到 2000 年前后基于信息技术和课堂的教学模式，再到 2010 年前后以慕课为代表的翻转课堂教

① Fishman J. Reversing Language Shift：Theoretical and Empirical Foundations of Assistance to Threatened Languages (Multilingual Matters 76)[M]. Clevedon：Multilingual Matters. 1991：395.

学模式。[①] 由于语言具备大数据的一系列特征,大数据或将成为语言教学技术的重要创新点,具有广阔的应用空间。

首先,在大数据和机器学习技术的加持下,语言教学或将从过去的"一对多"向"千人千面"转变。无论是电子教材还是慕课教学,都是一套标准化内容面向多个不同学生。教学过程中,学生产生的海量语音、语句等数据尚未内化成为教学过程的一部分。得益于大数据技术的迭代和机器学习的特性,未来大数据加持下的语言教学可以具备交互性,将不同学生产生的教学数据纳入大数据教学系统中,进行模型调参,形成"千人千面"的教学内容和教学过程。由于数据的可远程传输特性,让来自不同地区、不同教育资源、不同经济条件的学生都能够得到适合自己的学习支持,从而弥补学习资源的差距,促进机会公平。

第二,大数据或将改变语言教学的评价方式。长期以来,标准化考试都是评价语言教学成效的"金标准",过程性评价仅具备参考意义。在大数据和区块链技术的帮助下,未来学生的全过程语言学习的全部信息都可以进行安全存储和记录,考试成绩只是全过程数据中的一个部分,避免了学生评价对单一考试成绩的过度依赖,有助于更加全面系统地了解学生的学习发展轨迹和潜力,从而实现更公正的学习成果评估,促进机会公平。

第三,大数据将催生一系列新工具,提升语言教学的效率和质量。例如,传统教学过程中的听力材料几乎都由出版社通过人工录制获得,其缺点在于成本较高且发音较为单一,缺乏多样性。在大数据时代语音合成技术的广泛应用之下,听力材料能够直接通过语音合成技术产生不同国家和地区的听力素材。另外,大数据时代自然语言生成技术的发展也为写作教学提供了全新的工具,帮助学生更高效地进行素材收集和语言文字创作,也会带来人机交互共创的新型教学模式。[②] 以 ChatGPT 为代表的大模型,更为语言教学提供了广阔的应用空间和诸多挑战。

语言教育规划未来应该更加关注和适应新技术的发展趋势,主动拥抱大数据及相关技术的发展,提升大数据时代教师的技术素养,丰富课堂教学的技术方法,推动大数据和语言教学的深度融合,以期实现语言教育的重构。在大数据时代,过去象牙塔中的知识已经通过机器学习内化为 AI 系统的一部分,知识检索和内容创造最终导向为 AI 大模型的访问入口。AI 时代的语言教育规划仍要依托大数据方法和资源,只有充分积累语言教学全过程中的大数据资源,才能为教学、评估、教材等诸多领域提供 AI 建模和进一步拓展应用的基础。

值得注意的是,虽然学术讨论中对大数据和 AI 技术在语言教育规划的应用持积极态度,技术革新的前景也十分广阔,但是在实践中语言教育作为基础教育领域的重要组成部分,除经济性、公平性和效率性等考量之外,学生品格能力的培养和育人意义更是重中之重。百年大计,教育为本,技术的进步终归要服务于人的全面发展。[③] 相较于大数据技术在其他领域的飞速应用,在语言教育规划领域的应用须全面评估其综合影响。

(四)语言治理的学术研究

从学术研究的领域而言,语言治理的"数据转向"需整合语言社会调查和社会经济普查大数据,以实现更为全面、科学的语言治理。当前人口普查等社会经济类研究,往往忽视了语言社会使用的各项数据(如语言习得水平、习得年龄、语言使用现状、高频话语等),语言的社会调查通常局限于语音、词汇等本体。语言数据始终是社会经济大数据的"孤岛",难以实现数据间的有效联动。只有当语言充分成为社会经济大数据的一部分,才能更好地应对语言多样性和社会公平的挑战,推动语言治理进入真正"微观治理"的实践阶段,为不同语言社群提供公平的发展机会。

近 30 年的实践证明,国外语言规划理论的"批判转向"并未取得应有的成果。尽管语言的教育公平、性别公平和南北公平被反复倡导,但强势语言独大、语言生态破坏和语言资源不公平仍然呈现出日

① 胡加圣,陈坚林.外语教育技术学论纲[J].外语电化教学,2013,(2).

② 袁羲,吴应辉.ChatGPT Plus 给国际中文教育带来的机遇、风险及应对策略[J].云南师范大学学报(对外汉语教学与研究版),2023,(3).

③ 张海波,杨兆山.ChatGPT 的教育挑战与应答[J].四川师范大学学报(社会科学版),2023,(4).

益严重的趋势。追根溯源，语言的层级化是全球化发展的一种自然演化结果。① 这种符合齐普夫定律(Zipf's Law)的自然秩序难以通过学术批判和倡议加以扭转。语言治理研究应该描述语言的自然演化秩序，发现客观规律，坚持描写主义的客观立场，而不应成为一种价值诉求。

大数据范式为建构中国本土的语言治理理论提供契机。中国具备充分的人口普查的社会经济数据资源，移动互联网覆盖率高，语言使用大数据非常丰富。在大数据基础上，避免国外语言治理理论的窠臼，积极探索中国的语言治理和社会治理方案，深刻体现中国式现代化的科学内涵。

结　论

自二战以来，语言规划理论发展迄今已有 70 余年历史，国外语言规划理论经历了数次转型，其中 20 世纪 90 年代以来的“批判转向”影响至今。尽管国外语言规划长期倡导语言公平并致力于推动社会公平，但目前仍然只是乌托邦式的蓝图，尚缺乏切实可行的解决途径和应用场景。另一方面，随着大数据和人工智能特别是 ChatGPT 为代表的生成式人工智能的发展，语言治理或可突破原有的理论路径，实现语言治理的“数据转向”。

大数据作为科学研究的一种新范式，在包括语言学研究的各个学科中日益展现出理论创新的重要价值，但在当前的语言治理研究中仍然处于探索阶段。基于大数据的语言治理研究具有三重内涵：新的领域、新的方法和新的研究对象。基于大数据的语言治理在国家政府、企业、语言教育教学和学术研究四大领域具有广阔的应用前景。未来应重视语言大数据的生产、建设、开发和利用的全生命周期的治理，使得语言数据合理规范使用，发挥其社会价值，避免数据安全风险。

语言治理的大数据研究范式或将成为建构中国本土语言治理理论的重要契机，避免国外语言治理理论的批判化和微观化窠臼；推进语言数据和社会经济大数据联动，扎根中国语言国情，面向中国式现代化进程中的语言发展和社会治理需求，促进语言数据参与到国家社情民意、舆论建设、文化传播等现实问题的治理过程中；将 AI 生成语言纳入语言治理的研究框架，加强教育全流程的语言数据应用和监督，以建构科学的语言治理体系，服务中国式现代化。

Big Data-based Language Governance Research: Implications, Paradigms and Application Prospects

GUO Shujian & LI Xiaoyang

(*Center for Language Planning and Global Governance, Tongji University, Shanghai* 200092, *China*)

Abstract: The rapid development of AI language models and big data, exemplified by ChatGPT, poses new challenges to language governance and provides a significant opportunity for the construction of an indigenous theoretical framework for language governance in China. This paper reviews the historical development of the language planning theory, points out the limitations of foreign language planning theories, and explores the possibilities of a future "data-oriented" approach. The paper expounds the relationship between big data and language governance, reveals the connotations and methodological foundations of language governance based on big data, discusses the prospects of applying the big data of a given language in language governance across various domains, and proposes leveraging the advantages of big data in the process of constructing the knowledge system for the Chinese language policy and planning. The paper advocates breaking free from the constraints of foreign language governance theories, advancing the data-oriented shift in the research on language governance, and fully illustrating the effectiveness of China's language governance and social governance, thereby profoundly manifesting the scientific implications of Chinese-style modernization.

Key words: big data; language teaching; language governance; artificial intelligence; ChatGPT

[责任编辑：　和智利]

① 郭书谏，沈骑.互联网空间的世界语言活力及其成因[J].语言文字应用，2019，(1).