大数据时代背景下的中国传统语言学研究机遇与挑战

陆 梅，杨子淇

武汉工程科技学院　湖北武汉

【摘要】数字化技术的迅猛推进使得海量语言数据的采集、存储与分析成为可能，这一趋势正在深刻改变语言学研究的方法论格局。中国传统语言学（即“小学”，涵盖文字、音韵、训诂之学）拥有两千余年的学术积淀，但在数据驱动的研究范式面前，其既有方法和话语体系面临着更新压力。本文从大数据的基本特征出发，结合索绪尔关于“语言”与“言语”的经典区分，讨论数据驱动方法为传统语言学带来的三重机遇：语言现象的量化感知与自动化处理、基于语料证据的语法合法性梯度判断、语言共性与多样性的经验性解释。与此同时，文章也分析了数据质量参差、计算资源门槛、田野调查被弱化以及个人隐私与伦理等现实挑战，并就传统语言学如何在保持自身学术品格的前提下吸收计算方法提出若干思考。

【关键词】大数据；中国传统语言学；小学；数据驱动；研究范式；机遇与挑战

【收稿日期】2026 年 9 月 29 日　　【出刊日期】2026 年 10 月 21 日　【DOI】10.12208/j.ssr.20260345

Opportunities and challenges for traditional Chinese linguistics research in the era of big data

Mei Lu, Ziqi Yang

Wuhan University of Engineering Science, Wuhan, Hubei

【**Abstract**】The rapid advancement of digital technologies has made possible the collection, storage, and analysis of massive linguistic data, a trend that is profoundly reshaping the methodological landscape of linguistic research. Traditional Chinese linguistics—known as *xiaoxue* (“elementary learning”), encompassing the studies of script/characters (*wenzi*), phonology (*yinyun*), and exegesis/semantics (*xungu*)—has accumulated an intellectual heritage of more than two millennia. Yet in the face of the data-driven research paradigm, its established methods and discursive systems are under pressure to renew themselves. Starting from the basic characteristics of big data, and drawing on Saussure’s classic distinction between *langue* and *parole*, this article discusses threefold opportunities that data-driven methods bring to traditional linguistics: the quantitative perception and automated processing of linguistic phenomena, corpus-evidence-based gradient judgments of grammatical well-formedness, and empirical explanations of linguistic universals and diversity. At the same time, the article examines practical challenges, including uneven data quality, the threshold of computational resources, the weakening of fieldwork, and issues of personal privacy and ethics. It concludes with reflections on how traditional linguistics can incorporate computational methods while preserving its own scholarly character.

【**Keywords**】Big data; Traditional Chinese linguistics; *Xiaoxue*; Data-driven; Research paradigm; Opportunities and challenges

1 引言

近十余年，“大数据”从一个信息技术领域的术语逐渐扩散为人文社会科学普遍关注的议题。数据采集手段的丰富、存储成本的下降以及计算能力的跃升，使得以往难以想象的大规模语言材料分析成为可能。这一变化对语言学的影响尤为直接——语言本身就是一种数据，语料的规模和多样性在很大程度上决定了语言研究能走多远。

中国的语言研究有自身的传统。二十世纪以前，中国并无现代学科意义上的“语言学”，文字学、音韵学、训诂学长期依附于经学与文学，却也在这一框架内积累了极为丰厚的成果。从《说文解字》以部首统摄九千余字的体例创辟，到《广韵》对中古音系的系统排比，再到清代干嘉学者在古音学上达到的精密程度，传统“小学”在汉字形、音、义的考索方面展现出令人惊叹的学术智慧。二十世纪初，西学东渐的风潮之下，尤其

是新文化运动的推动，一批受过现代语言学训练的学者建立起专门的研究机构，语言学自此成为中国现代学术体系中的独立学科[1,2]。今天所讨论的“中国传统语言学”

问题由此而来：在大数据已经深刻重塑众多学科研究方式的当下，这门古老的学问应当如何自处？它能否借助新的技术手段焕发活力，又会在这一过程中遭遇哪些困难？本文不打算给出面面俱到的答案，而是希望结合已有研究和笔者的观察，对其中若干关键问题做初步讨论。

2 大数据的基本特征及其语言学意涵

通常用“4V”来概括其特征：Volume（体量浩大）、Velocity（生成快速）、Variety（类型多样）、Value（价值密度低但潜在价值高）[3]。需要指出的是，“大”只是一个相对概念

这一思路与语言学的根本问题有着深层关联。索绪尔在《普通语言学教程》中区分了“语言”（langue）和“言语”（parole）：语言是言语活动的社会规约层面，是一种平均数；言语则是个人在具体情境中的实际表达。索绪尔本人认为语言学应当优先研究“语言”，即那个抽象的规约系统。但若进一步追问：这个“平均数”究竟如何计算？怎样才能让抽象的语言系统逼近真实的人类语言能力？仅仅依靠逻辑代数的形式化描写恐怕不够。大数据方法恰恰提供了一种“计算平均数”的可能路径——通过对海量言语样本的统计分析，归纳出语言系统的概率性规律。反过来看，任何完全脱离使用者、脱离真实语料的过度形式化，都可能错失语言运作的实际机制。这或许是当代语言学在数智时代需要正视的一个方法论问题。

对中国传统语言学而言，上述意涵尤为值得关注。传统小学在方法上长于个案考据和系统排比，前者如王念孙《读书杂志》对具体文句的校释，后者如陈澧《切韵考》用系联法归纳《广韵》声类韵类——这些工作本质上也是在“数据”中发现规律，只不过受限于人力和材料，所能处理的数据规模有限。如果能将传统学者的学术敏感与大数据的处理能力结合起来，其前景值得期待。

3 大数据为中国传统语言学研究带来的机遇

“能用数据解决哪些语言学问题”，而非仅仅追求数据规模[3]。这一判断对传统语言学同样适用。

3.1 语言现象的量化感知与自动化处理

数据驱动方法的一个直接优势，是为研究者提供了感知语言现象的量化维度。借助机器学习和自然语言处理技术，自动分词、词性标注、句法分析等以往需要大量人工的工作，如今可以在短时间内对海量文本完成。这对于传统语言学研究的意义不可低估。以古汉语语料为例，从上古的金文、《尚书》《左传》，到中古的佛经译文、唐宋诗文，再到近代的白话小说，汉语留下了世界上最为庞大且连续的文献记录。然而，这些材料的语言面貌长期以来主要靠学者凭语感和手工统计来把握。如果能够建设高质量的历时语料库，对不同时期的词汇频率、句式分布、语法标记消长等做精细化的计量分析，许多此前只能定性描述的问题就可能获得定量证据的支撑。

事实上，这方面已有一些探索。例如，基于先秦两汉传世文献与出土文献的语料库，研究者可以统计某一虚词在不同文体、不同时期的出现频率和句法分布，从而更准确地描写语法化的进程；利用音韵学研究中积累的反切数据和韵书材料，结合统计模型，也可以对中古音系的声韵配合关系做更系统的检验。这些工作并不是要取代学者的学术判断，而是为判断提供更扎实的经验基础。

3.2 语法合法性的梯度判断与语言真实性的回归

有些结构在书面语中罕见，在口语中却相当自然[4]。如果有足够规模的真实语料作为参照以对不同说法的可接受程度做更细致的区分，而不是简单地一分为二。

“语言学研究语言，而语言是民众实际所讲的语言。”[5]这句话看似平实

3.3 语言共性与多样性的经验性解释

王士元将语言视为一种“复杂适应系统”，强调语言演化受到多种因素的交互影响[6]。这一视角在大数据条件下获得了新的研究可能。

但语料证据显示，三层以上的嵌套递归在实际话语中极为罕见——人毕竟不是机器，认知加工能力是有约束的[3]。这类发现单靠内省难以获得

4 大数据时代传统语言学面临的挑战

机遇与挑战往往是一体两面。大数据在为语言研究开辟新路径的同时，也带来了若干不容回避的问题。

4.1 数据质量与代表性问题

大数据不等于好数据。对于以传统语言文字为对象的研究而言，材料的可靠性和精准度直接关系到结论的可信度。目前网络上可获取的古汉语语料资源虽然不少，但质量参差不齐：有的文本未经校勘，讹误衍脱随处可见；有的标注体系不统一，分词和词性标注错误率较高；还有的在数字化过程中出现了繁简转换错误、异体字处理失当等问题。如果研究者不加甄别地使

用这些数据，再精密的统计模型也只会产出不可靠的结论。此外，“海量”本身也可能造成一种认知偏差——当数据量足够大时，统计上的显著效应很容易获得，但效应的实质意义未必有多大。如何区分统计显著性与学术重要性，是研究者需要警惕的。

更深一层的问题在于数据的代表性。现存古汉语语料以传世文献为主，而出土文献、民间文书、口传材料等在数字化程度上相对滞后。如果研究结论主要建立在经过历史筛选的书面文本之上，那么我们所描绘的“古代语言面貌”很可能是不完整的，甚至带有某种精英偏向。这一问题并非大数据时代才有，但数据规模的扩放大可能会让这种偏向以更隐蔽的方式存在。

4.2 计算资源与技术门槛

但要真正做好一个计算语言学项目，仍然需要相当的技术积累[7]。民办高校和地方院校在硬件条件和经费支持上相对有限

另一个值得注意的倾向是：在数据库和检索工具日益便利的情况下，部分研究者容易产生“捷径依赖”，习惯于在屏幕前做检索和统计，而忽视了田野调查和文献查阅的基本功。以方言学为例，语音数据库可以提供大量录音和标注材料，但一个音位的实际音值、一个变调规则的适用条件，往往需要调查者在现场反复听辨、与发音人深入交流才能准确把握。语言学本质上是一门经验科学，数据调查始终是必要环节——它是有目的、有假设的搜集和研究活动，需要明确调查对象、调查方式和筛选标准。数据库是辅助手段，不能替代脚踏实地的田野工作。

4.3 隐私保护与学术伦理

语言数据常常涉及个人身份、社交关系、观点态度等敏感信息。在大规模采集和分析语料的过程中，如何保护被调查者的隐私、防止数据滥用，是一个必须正视的伦理问题。尤其在方言调查和口语语料库建设中，发音人的个人信息、家庭背景、生活经历等都可能被录音和记录。如果这些材料在数字化存储和共享过程中管理不当，就可能造成隐私泄露。学界有必要在数据采集知情同意、脱敏处理标准、使用权限分级等方面建立清晰的规范，让技术进步不以牺牲个人权益为代价。

4.4 方法借鉴与学术主体性的平衡

有什么工具就做什么题目，而不是从真正的学术问题出发选择方法[8]。

这些能力不会因为技术手段的更新而过时，反而应该在新技术的加持下得到更好的发挥[9]。理想的状态是“以我为主、为我所用”

5 结语

大数据时代的到来，对中国传统语言学而言既是契机也是考验。从积极面看，量化手段和自动化处理拓展了研究的边界，使以往难以企及的大规模语料分析成为可能；语料证据的引入让语法判断更加精细，也为语言共性和多样性的解释提供了新的经验基础。从困难面看，数据质量参差不齐、计算资源分配不均、田野调查传统有被弱化的风险，隐私伦理和学术主体性问题也需要认真对待。

走出一条既具本土特色又有国际视野的创新之路[1]。

参考文献

[1] 刘海涛. 中国语言学建设两大要务: 成果国际化和方法科学化[J]. 语言战略研究, 2018(1):37-50.

[2] 李宇明. 语言学是一个学科群[J]. 语言战略研究, 2018(1): 15-24.

[3] 刘海涛, 林燕妮. 大数据时代语言研究的方法和趋向[J]. 新疆师范大学学报(哲学社会科学版), 2018(1):72-83.

[4] 威廉 •冯 •洪堡特. 论人类语言结构的差异及其对人类精神发展的影响[M]. 姚小平, 译. 北京: 商务印书馆, 1999.

[5] 伯纳德 • 科姆里. 语言共性和语言类型[M]. 沈家煊, 罗天华, 译. 北京: 北京大学出版社, 2010.

[6] 王士元. 语言是一个复杂适应系统[J]. 清华大学学报(哲学社会科学版), 2006(6):5-13.

[7] 郑国锋, John Whitman. 语言学、语言学流派、语言学教育:康奈尔大学语言学系主任约翰 •惠特曼教授访谈录[J]. 外国语, 2020(5): 121-125.

[8] 徐烈炯. 生成语法理论[M]. 上海: 上海外语教育出版社, 1988.

[9] 维特根斯坦. 哲学研究[M]. 陈嘉映, 译. 上海: 上海人民出版社, 2005.

版权声明：©2026 作者与开放获取期刊研究中心（OAJRC）所有。本文章按照知识共享署名许可条款发表。
http://creativecommons.org/licenses/by/4.0/

OPEN ACCESS