语言数据科学

陈居强简介

澳大利亚西悉尼大学，MARCS脑、行为与发展研究所，博士候选人。南京大学英语语言文学硕士。中国英汉比较研究会语音教学研究专业委员会理事，International Speech Communication Association会员，Asia-Pacific Signal and Information Processing Association (APSIPA) 会员 。学术论文发表语音学顶级刊物Journal of Phonetics，以及International Congress of Phonetic Sciences (2019)， Interspeech (2019)， APSIPA (2019)等国际重要语音学会议上。个人主页：<https://juqiangj.github.io/index>。研究兴趣为：语音学、心理语言学、语料库语言学、数据科学、数字人文。

内容简介

本书介绍数据科学视域下数据分析原理与流程，以及基于R语言的编程操作。内容方面，用语言研究实例重点介绍语言实验数据、语料库文本数据和语音库声学数据的分析，包含了语言研究的各种数据类型。针对每一种数据类型，深入浅出地介绍数据导入，数据清洗与整理，探索性数据分析、数据可视化和数据建模（统计建模和机器学习建模）的各个步骤以及R语言实现过程。

读者对象

本书的目标人群是从事语言研究的研究者，老师和同学。既可以单独作为教材用于语言学和应用语言学专业的方法论课程，也可以用于实验语言学、心理语言学、语料库语言学和语音学等课程讲解数据分析之用。同时对于语言研究感兴趣的人士也可以自修。

该选题的特色技术背景和必要性：

当今社会，数据已经成为与“劳动、资本、土地、知识、技术、管理”并列的第七大生产要素，习近平总书记提出“要构建以数据为关键要素的数字经济”。语言数据既是一种重要资源，同时也具有一定处理分析难度。语言研究中，既要处理结构性数据（structured data）如实验数据，也要处理非结构性数据（unstructured data），如文本数据。在大数据时代，语言研究者需要了解和掌握，如何科学高效地处理数据，使用统计和机器学习模型为数据建模，以及有效地进行数据可视化和数据展示。文科背景的语言研究者对于数据科学了解不多，对于基本的计算机编程技能掌握有限。需要一本紧密结合语言研究实例，深入浅出讲解编程步骤，可以直接上手操作的指南性入门教材。本书以此为目标撰写。

列出已出版的同类书的出版社、印数、定价、作者及出版时间

R语言数据可视化之美，电子工业出版社，159元，张杰，2019年

数据科学导论——R与Python实现，高等教育出版社，48元，吴喜之、刘苗，2019年

语料库与Python应用，上海交通大学出版社，58元，管新潮，2018年

基于Python的语料库数据处理，科学出版社，68元，雷蕾，2020年

外国语言学应用语言学统计方法，西北农林科技大学出版社，25元，马广惠，2003年

本书与同类书比较

国内市面上介绍数据科学的中文著作，如《数据科学导论——R与Python实现》，很少从语言研究的角度来讲解，大多案例都是商业、医疗、工程等内容，虽然能对语言研究者有所启发，但是内容关联度很低，难以直接运用。

一些以语言学和应用语言学方法论为主题，结合R语言讲解数据分析的书籍，往往只介绍各种统计模型，然而却没有系统讲解数据分析的各个过程。语言数据的预处理难度大，如何高效地将原始数据清洗、整理、加工成为可以建模的清洁数据，并且将这些过程编写成自定义函数、代码块，从而保持可重复性和自动化，并且方便日后调取使用，迭代更新。这些问题往往被忽视。换句话说，现有书籍是统计学在语言研究中的应用指南，但是没有站在数据科学的视角全盘考虑语言数据分析和处理。

另一些书，虽然介绍了数据处理的各个过程，但是只介绍了单一类型的数据，主要是语料库文本数据，如《语料库与Python应用》等。语言研究的其他数据类型，如实验数据和语音数据的分析有各自的特点，也需要得到重视。

编著计划：（计划何时写完此书，有无合作者，拟请谁对本稿进行审阅）

计划2021年下半年，无合作者

章节目录和编著此书的重要参考资料

本书的结构

数据科学引论

语言研究中的数据科学

R语言入门

语言实验数据预处理

语料库构建与数据预处理

语音库构建与数据预处理

探索性数据分析与数据可视化

统计数据建模入门

机器学习建模入门

参考文献

Cady, F. (2017). The data science handbook. Wiley.

Leek, J. T., & Peng, R. D. (2015). What is the question? Science, 347(6228), 1314–1315. https://doi.org/10.1126/science.aaa6146

Ozdemir, S. (2016). Principles of data science. Packt Publishing.

Peng, R. D., & Matsui, E. (2015). The art of data science. Leanpub.

Wickham, H., & Grolemund, G. (2017). R for data science. O’REILLY.

科研成果

|  |  |
| --- | --- |
| 2020 | **Chen, J.,** Best, C. T., & Antoniou, M. Native phonological and phonetic influences in perceptual assimilation of monosyllabic Thai lexical tones by Mandarin and Vietnamese listeners. *Journal of Phonetics*, *83*, 101013. （SSCI检索） |
| 2020 | Wang, F. (Robin), Humblé, P., & **Chen, J**. Towards a Socio-Cultural Account of Literary Canon’s Retranslation and Reinterpretation: The Case of The Journey to the West. *Critical Arts*, 0(0), 1–15. <https://doi.org/10.1080/02560046.2020.1753796>.（SSCI、A&HCI检索） |
| 2019 | **Chen, J.,** Best, C., & Antoniou, M. Cognitive Factors in Thai-Naïve Mandarin Speakers’ Imitation of Thai Lexical Tones. *Proc. Interspeech 2019*, 2653–2657. <https://doi.org/10.21437/Interspeech.2019-1403>. （EI检索） |
| 2019 | **Chen, J.,** & He, X. Computational perception of information foci produced by Chinese English learners and American English speakers. *Proceedings of APSIPA Annual Summit and Conference 2019, 1780–1785.* （EI检索） |
| 2019 | **Chen, J.,** Best, C., Antoniou, M., & Kasisopa, B. Cognitive factors in perception of Thai tones by naïve Mandarin listeners. In S. Calhoun, P. Escudero, M. Tabain, & P. Warren (Eds.), *Proceedings of the 19th ICPhS*, (pp. 1684–1688).（同行评审、全文发表） |
| 2019 | Carne, M., **Chen, J.,** Luk, E., Strangways, S., Stockigt, C., Maihammer, R., & Harvey, M. Rhotic contrasts in Arabana. In S. Calhoun, P. Escudero, M. Tabain, & P. Warren (Eds.), *Proceedings of the 19th ICPhS,* (pp. 1278–1282). （同行评审、全文发表） |
| 2018 | **Chen, J.,** Best, C. T., Antoniou, M., & Kasisopa, B. Cross-language categorisation of monosyllabic Thai tones by Mandarin and Vietnamese speakers: L1 phonological and phonetic influences. In J. Epps, J. Wolfe, J. Smith, & C. Jones (Eds.), *Proceedings of the Seventeenth Australasian International Conference on Speech Science and Technology* (pp. 168–172). Sydney, Australia. （同行评审、全文发表） |

A person posing for the camera

Description automatically generated