中青年学者外国语言文学学术前沿研究丛书

第二语言加工及R语言应用

吴诗玉 著

SECOND LANGUAGE PROCESSING AND R IN APPLICATION

外语教学与研究出版社
FOREIGN LANGUAGE TEACHING AND RESEARCH PRESS

中青年学者外国语言文学学术前沿研究丛书
上海交通大学文理交叉专项基金（16JXRN09）

第二语言加工及R语言应用

吴诗玉 著

SECOND LANGUAGE PROCESSING AND R IN APPLICATION

外语教学与研究出版社
FOREIGN LANGUAGE TEACHING AND RESEARCH PRESS
北京 BEIJING

图书在版编目（CIP）数据

第二语言加工及R语言应用 / 吴诗玉著. -- 北京 : 外语教学与研究出版社, 2019.8
（中青年学者外国语言文学学术前沿研究丛书）
ISBN 978-7-5213-1176-1

Ⅰ. ①第… Ⅱ. ①吴… Ⅲ. ①第二语言－研究 Ⅳ. ①H003

中国版本图书馆 CIP 数据核字（2019）第 196218 号

出 版 人　徐建忠
责任编辑　孔乃卓
责任校对　付分钗
封面设计　高　蕾
出版发行　外语教学与研究出版社
社　　址　北京市西三环北路 19 号（100089）
网　　址　http://www.fltrp.com
印　　刷　北京九州迅驰传媒文化有限公司
开　　本　650×980　1/16
印　　张　20.5
版　　次　2019 年 9 月第 1 版　2019 年 9 月第 1 次印刷
书　　号　ISBN 978-7-5213-1176-1
定　　价　79.90 元

购书咨询：（010）88819926　电子邮箱：club@fltrp.com
外研书店：https://waiyants.tmall.com
凡印刷、装订质量问题，请联系我社印制部
联系电话：（010）61207896　电子邮箱：zhijian@fltrp.com
凡侵权、盗版书籍线索，请联系我社法律事务部
举报电话：（010）88817519　电子邮箱：banquan@fltrp.com
物料号：311760001

丛书序

进入新世纪以来，我国外语学科的学术研究呈现繁荣的局面，不仅学界专家、尤其是中青年学者的研究热情持续高涨，而且申报项目、出版论著或者发表论文的数量和质量也大幅度提高。以全国哲学社会科学规划办公室提供的数据为例，2014年国家社科基金年度项目和青年项目申报总数为28,186项，其中30-45岁这一年龄段共有17,729人申报，占总申报量的62.9%；经过评审，共有2,395人立项，占总立项数的62.7%。这表明包括外语学科在内的中青年学者已成为学术研究的中坚力量。与此同时，外语学科的研究者除了在国内外重要学术刊物发表高水平的论文之外，还不断开拓论著发表渠道，通过创办学术集刊等形式表达和传播学术思想。据统计，近年来，各高校创办的外语集刊达几十种之多，在国外刊物上发表论文的数量也逐年增加。形成这种局面的原因多种多样，但主要归功于研究者们探索新知的内在需求和国家激励学术创新的外部环境。

然而，由于种种原因，中青年学者仍然面临着专著出版难的问题。尽管他们大多在国内外接受过系统的学术训练，思想活跃、心智敏锐，能够迅速抓住学术前沿话题，撰写的专著具有较高的学术价值，但常常陷入出版难的窘境，即使几经周折得以出版，也是“零散型”的著作，淹没在书海中而难以产生较大影响，更无法形成“集约型”效应。从某种意义上来说，这是极大的智力资源浪费，也在一定程度上挫伤了中青年学者学术研究的积极性。

中青年学者是中国学术发展的希望。为了保护中青年学者的学术热情，推动他们在学术前沿勇于探索，多出成果，外语教学与研究出版社秉承“开放性、学术性、原创性、前沿性”的原则，设立“中青年学者外国语言文学学术前沿研究丛书”出版项目。“开放性”是指对全国各高等学校和科研院所开放，凡是希望通过这一项目出版外国语言文学类各领域研究成果的中青年学者，皆可申请加入；“学术性”是指研究成果具有深刻的学术思想和观点，能够反映外国语言文学各领域理论和实践的本质和规律；“原创性”是指发前人所未发、想前人所未想，在研究内容和形式上有创见和创新；“前沿性”是指研究成果反映外国语言文学各领域的最新发展动态，具有前瞻性。近年来，随着计算机技术的广泛应用

和全球化进程的进一步加快，外国语言文学学术研究的内容和形式发生了变化，并呈现交叉融合的趋势。因此，我们鼓励中青年学者以问题为导向，运用新技术、新方法和新视角，对语言学、文学、翻译学、外语教学等诸多方面进行跨学科研究，力争推出一批集学术性、原创性和前瞻性为一体的最新研究成果，构建具有中国特色的外国语言文学学术话语体系。

首批入选“中青年学者外国语言文学学术前沿研究丛书”的专著共有十本：《教育语言学——一个社会符号的模式》、《文学交际中的读者：叙事虚构作品解读的自由与局限》、《翻译诗学》、《中国晚期二语习得者英语被动句加工的认知神经机制：行为与脑科学的整合研究》、《萧乾文学翻译思想研究》、《詹姆斯·马丁访谈录》、《英汉语篇跨文化修辞研究》、《莎士比亚戏剧中强调语的语用分析》、《基于语料库的中国英语学习者强化语型式和意义研究》和《国家形象与外宣翻译策略研究》，涵盖语言学、文学、翻译、外语教学等不同领域。

本中青年学者外国语言文学学术前沿研究丛书注重质量和创新。首批入选的专著均经过专家审读和评阅。为确保丛书的学术品质，外语教学与研究出版社还将聘请资深专家担任系列丛书的顾问，并成立由外国语言文学学科领域知名学者组成的编委会，对申请出版的学术专著进行评议和遴选。本丛书项目将按照相关质量标准，组织专家对申报项目进行评审，入选者列入出版计划。

我们秉承严谨治学的态度，希望通过学术出版为学术传承与创新提供平台，但由于水平有限，书中恐有疏漏之处，恳请诸位专家和学者不吝指教。

外语教学与研究出版社
高等英语教育出版分社
2015年9月

前　言

本书分上下两篇。上篇集中介绍R语言基础知识以及使用R进行统计建模的过程；下篇将以具体的语言研究的数据为实例，尤其是在第二语言加工这一主题下，展示R语言的具体应用，将重点介绍我们前期所开展的第二语言加工的多个实验，并展示使用R进行数据分析的详细过程。

上篇，R语言基础及使用R进行统计建模部分，一共包括六章。其中，前面三章是基础，对R语言基础很薄弱的读者来说，阅读和学习这三章，可以比较快地掌握一些重要的基础知识，并快速上手。概括起来，这些基础知识主要包括三方面的内容：（1）数据操作和管理；（2）描述统计，即对数进行探索和挖掘；（3）作图，即数据可视化，通过图形来挖掘数据。第四章是知识的核心，它比较详细地介绍了使用R进行统计建模的思路、方法和思想。第五章有点离题，介绍了传统的重复测量的t检验和方差分析（ANOVA）方法。为什么说有点离题呢？因为重复测量的统计方法完全可以去掉，用第六章的混合效应模型的知识来解决。

下篇，第二语言加工主题下的R语言应用，将包括三章，每章介绍一个主题，涵盖第二语言语音加工、词汇加工（两章）和句子加工。每章会先介绍研究背景、动机、意义，以及具体的研究问题，然后介绍数据管理和操作的过程，最后介绍统计建模的过程以及数据的解释。通过这些具体的实例，大家可以把上篇学到的知识付诸实践。由于每个研究我们都提供了原始数据，大家可以通过自己操作、学习和体会，进一步提升自己的能力。

由于时间仓促，再加上水平有限，错误之处在所难免，恳请广大读者批评、指正！

使用说明

本书附带有每个章节相应的数据和操作代码，请到相关网页下载，网址为：http://heep.unipus.cn。数据按章节分类，但是代码以一个总文件的方式呈现，按章节排列，文件名为L2ProcessingR.zip。使用R语言来分析数据首先要学会的是如何把使用别的软件（如excel等）准备好的数据导入到R（或Rstudio）。本书作者在写作时所设想的阅读对象是从入门逐渐到高级阶段的读者，为了方便理解，在介绍如何导入数据时都使用了绝对路径，即指定在电脑的哪个盘，在哪个目录下，以及文件名（包括扩展名），比如：E:/bookR/MAdata.csv。初学者完全可以按相同的路径，把自己的数据和代码下载到相同的路径指向的文件夹。但是在大部分情况下，一般不会在代码里提供绝对路径，因为这并不利于代码的共享，也不利于科研合作，因为别的研究者未必在电脑上也建立了同样的路径，而直接提供一个路径，未经同意改变别人电脑的设置也不符合社会行为规范。

一般来说，读者可以采用三种方法来把外部数据导入到R。第一种方法就是上面说的提供绝对路径，告诉R数据文件准确的存放路径，R会按照这个地址把它读入。第二种方法就是提供一个交互页面，通过这个交互页面，读者自己找到数据文件存放的位置，比如：read.csv（file.choose（），header = TRUE）。第三种方法，是最为推荐、也是高级使用者使用最多的方法，即以创建项目的方式来读取数据。这种做法的理念就是，我们每一次数据分析都是基于一个独立的项目（project），因此，我们会习惯性地把跟这个项目相关的所有的东西存在同一个文件夹，比如项目分析的代码、数据文件、项目生成的结果、图形以及分析历史等等，这样，以后每次要回顾这个项目时，就非常方便，并且可以获得所有的记录。这么做的方法是从Rstudio菜单中的File菜单下找到New Project，然后创建New Directory。把代码，数据文件等都一并存入这个New Directory。进行这样操作后，再读入数据时就不再需要提供详细路径了，只要提供数据文件名就行了，比如：read_csv（'MAdata.csv'），就能读入相应的数据（也可参看书的1.1.2小节，设置工作目录）。这么做也容易跟别的研究者共享分析代码。

对初学者来说，上述内容可能需要经过一段时间学习后再来读才能读懂。如果是那样的话，读者可以先按书上提供的绝对路径读入数据。您要做的是创造相同名字的文件夹，然后把数据放入这个文件夹。

数据和代码文件从上面的网站下载后，需要进行解压缩，解压缩的密码是：A20190416L。

目 录

上篇 使用R语言进行统计建模

上篇　使用R语言进行统计建模

第一章
R语言数据科学

1.1 基础操作

1.1.1 简介及安装

就像任何别的学科一样，语言科学研究的进步也必须依赖于合理的研究设计，科学的数据分析和对研究结果清楚、透明的报道。为了在实验过程中去除“噪音”，第二语言加工研究经常需要设计大量的实验刺激材料，针对不同的被试群体开展实验，由此获得大量的数据，从而对提出的研究假设进行科学的论证。但是，如何才能高效、科学地整理实验数据，再在此基础上构建简洁、准确的统计模型，并对模型进行解释呢？R语言给我们提供了简单便捷的答案。掌握R语言将使我们很好地“武装”起来，从而能更加“气定神闲”地面对许多科学研究的问题。这不仅让我们具备极其重要的数据管理、分析和解读的能力，而且还会让我们具备科学的思维和视角，去提出问题、分析问题和解决问题，或者去思考、验证和解决一些科学难题。

那么什么是R语言？这几乎是所有的相关书籍一开始都会介绍的问题。总结起来，大都是这样说的：“R是一种为统计计算和绘图而生的语言和环境，它是一套开源的数据分析解决方案，由一个庞大且活跃的全球性研究型社区维护”，“直到大数据的爆发，R语言变成了炙手可热的数据分析的利器”。正是因为这个原因，有学者呼吁让R成为应用语言学研究者的学术通用语（Lingua Franca）（Mizumoto & Plonsky 2016），这个呼吁跟最近二语习得研究领域有学者提出要尝试使用多元回归来代替方差分析的思想遥相呼应（见Plonksy & Oswald 2017），这一方面表明了语言

研究者更加关注科学的发展；另一方面也说明二语研究在方法方面不断取得进步。

Mizumoto & Plonsky（2016）在阐述让R成为应用语言学研究的学术通用语时总结了R在四个方面的优势，包括：(1) 数据分析的可重复性。研究者只需要分享他的代码就可以共享他整个数据分析的过程，极大地便利了研究的可重复性。同时，这也有利于不同的研究者，包括距离遥远的研究者开展科研合作。(2) 使用R是站在巨人肩膀上的工作。这是因为R社区有许多“聪明的人”开发了无数的可供直接利用的包（packages），这些包让极其复杂的运算和统计分析变得简单。(3) 极强的数据可视化能力。在二语研究中，有时需要处理大量的反应时数据或者频数数据，数据可视化就显得非常重要。通过可视化一方面可以清楚地看出数据的分布规律，观察可能出现的异常值，离群点，等等；另一方面，还可以帮助解读数据，尤其是解读多个变量的交互效应，被认为“困扰着世界上许多聪明的大脑”的一件极其复杂的事情（Field et al. 2012）。(4) R是一门编程语言，因此具有极强的灵活性（flexibility）和无限的才华（versatility）。

使用R与使用其他一些商业软件如SPSS的一个很大区别就在于我们需要使用键盘输入代码，因为R的特点就在于你必须告诉它干什么，它才能干什么。这似乎会让我们觉得这东西有点“傻”，这取决于您怎么看，如果习惯了傻瓜式的一键操作，可能是会有这种感觉，但是反过来想，“告诉它干什么，它才能干什么”不正是它的强大之处吗？R的才华横溢之处正是来自于这里，对这一点，我相信后面大家会有更多的体会。但是无论如何，大家都得从输入开始，而且R还非常挑剔，输入时稍有错误，就无法运行。一开始大家不免觉得厌烦，甚至沮丧，甚至“绝望”。我给大家的建议就是一定要坚持下去，因为这东西怪就怪在你越使用它，你就会越喜欢它，而操控的感觉是一种真正的自由！这一点是很多R使用者共同的感受。

先从软件安装入手说起吧。大家可以到CRAN（Comprehensive R Archive Newwork）下载R（建议安装全英文版），网址是：https://cloud.r-project.org。尽管CRAN由分布在世界各地的很多镜像服务器组成，用于分发R和R包。但建议不要选择离你近的服务器，而应该使用云镜像，因为它会自动找出离得近的服务器。打开网站后，根据自己的计算机系统以及配置选择要安装的R。安装好R后，建议大家都安装Rstudio，它

是用于R编程的一种集成开发环境，可以从这个网址下载安装：http://www.rstudio.com/download。Rstudio安装好后会自动跟已经安装好了的R关联起来，以后你只要打开Rstudio，在那里进行所有的操作就行了。

把Rstudio安装好后，打开Rstudio，点击菜单中的File，然后再点击New File，选择R Script就会显示如下图所示的页面（也可以使用快捷键：Shift + Ctrl + N）：

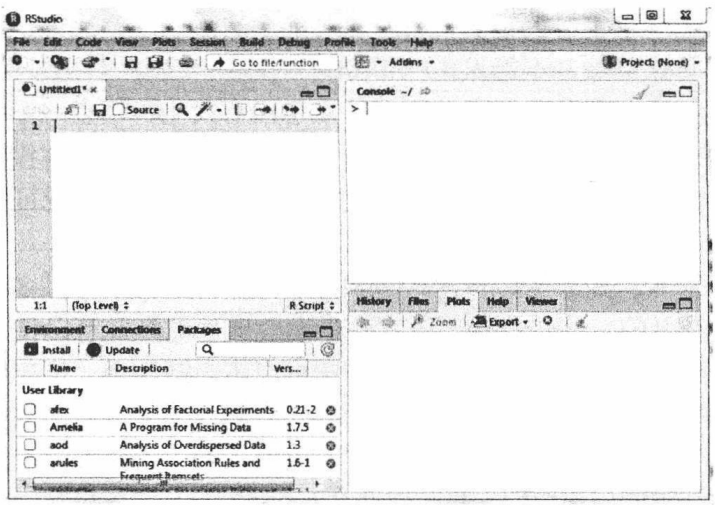

图1.1　Rstudio的工作区

当然，您的页面可能会有些不同，但都可以使用Tools菜单下面的global options自行设置。在global options菜单下，可以根据自己的偏好，设置Rstudio的页面网格，包括字体大小、背景颜色和界面风格，等等。Rstudio里而有三个关键区域（它们的位置布局通过global options调整）：左上角，那里是代码编辑区，在那里输入代码，按“Ctrl + ENTER”键后就可以执行代码，执行的结果显示在右上角称作console的地方，即控制台。右下角可以显示通过代码编辑区生成的图片（plots）。

R和Rstudio都会经常更新。按理应该经常保持更新，Rstudio的更新比较简单，但是更新R却比较麻烦，因为我们在使用了R一段时间以后，已经在上面安装好了很多的包，如果重新下载安装新的R的话，之前安装好的所有包必须全部重新安装，这是一件比较麻烦的事情。但是如果不经常更新，就无法使用一些新的功能，大家只能权衡利弊了。

软件安装好了，就可以开始正式的工作了。比如，在代码编辑区输