2 语言研究中的数据科学

语言研究中的数据与特点

语言可以包括口语和书面语。因此，语言数据既包括文本数据（含口语转录的文本），也包括语音数据。此外，语言使用者的社会人口学数据也是语言研究的重要组成部分。在实验语言学或心理语言学研究中，数据还包括语言使用者加工语言的行为数据和神经活动数据。语言研究涉及的数据类型十分丰富，同时也为数据处理提出了巨大的挑战。

从数据结构来看，语言数据可以分为结构型数据和非结构型数据：

结构化数据是指数据以表格形式呈现，其中每一行是一系列观测值，每一列记录一个属性特征。例如，在应用语言学中，学生成绩和相关的教学实验条件就是典型的结构化数据。每一行是一个学生的个人信息（如学号、性别）以及成绩。每一列是一个变量，比如组别可以包含实验组或者控制组。

|  |  |  |  |
| --- | --- | --- | --- |
| 学号 | 性别 | 组别 | 成绩 |
| 18301 | 男 | 实验组 | 93 |
| 18301 | 女 | 控制组 | 98 |
| 18301 | 男 | 控制组 | 87 |
| 18301 | 女 | 实验组 | 75 |
| 18301 | 女 | 实验组 | 96 |
| 18301 | 男 | 控制组 | 68 |
| … | … | … | … |

在语言研究中，实验数据通常是结构性的。此外，语音数据及标注，还有相应的声学指标，如时长，元音的共振峰，响音部分的基频等一般也是以结构性数据形式存在。通常，对于语音数据，我们需要通过转写（语音转写、文字转写）得到对应的文本信息，然后把转写的内容和语音材料进行时间上的对齐（alignment）。尽管目前在工业应用上，机器学习和深度学习可以极大的提升数据处理的效率，在语言研究中，由于对精度的要求，人工标注和校对仍然是一项非常耗费人力和时间的工作。当然，如果不需要对语音做声学分析，不是所有的语音材料都需要时间的对齐。大部分的口语语料库是将语音材料转写并填加备注信息。结构化数据是较为容易存储，处理的。

非结构性数据是指不遵循某种标准组织结构的数据。非结构化数据本身是一个整体，没有表格化的结构。文本数据比如博客，文章，小说等本身是非结构化的，不能直接用数据科学中的数据探索分析方法或者数据建模和数据可视化。文本数据可以通过分词，词频统计，特征提取等方法转化为结构性数据。只有当转化为结构性数据之后，进一步加入元数据，我们才能够来讨论变量之间的关系。在语言研究中，语料库数据通常是非结构性的，对语料库进行量化分析需要我们提取相应词频等信息并转化成为结构性的数据。语料库语言学，量化语言学和人工智能中的自然语言处理等学科在文本处理方面探索发展了一系列方法。

语言数据还可以分为质性数据和量化数据。质性数据是指不能够用数字来表示，也不能够进行简单的数学计算，而仅仅用来描述某个事物的范畴属性。比如，语音学中一个音是元音，还是辅音。一个词的词性，名词，动词。一句话所具有的语用功能。量化数据是可以用数字来描述并且可以进行简单的数学运算。进一步可以分为离散数据（Discrete）和连续数据（ Continuous）。离散数据只能用某一些数值代表，比如参与实验的人数，实验中刺激的个数。连续数据（ Continuous）可以有无限的数据数值代表，比如实验任务的反应时。

语言数据可以根据其描述性统计特性分为四种不同层次的数据，分别为定类数据、定序数据、定距数据及定比数据。定位数据和定序数据为类别性数据，定距数据和定比数据为数值型数据。

定类数据描述一个类别，比如性别、国籍。单词本身也是一种定类数据。对于定类数据，我们可以描述其频数（率），或者一组数据中的众数[[1]](#footnote-1)（出现频数最高的）。

定序数据描述一个有顺序类别。比如，调查问卷中，喜欢、一般、厌恶，这3个选项就构成了有顺序数据。定序数据可以按数据顺序排列，因此除了众数之外，我们还可以描述其中位数，也就是将一组数据按一定顺序排列，其中处于中间位置的数据。

定距数据是一种数值性数据，比如气温。定距数据可以进行加减。比如，在实验中一个人的反应时长，或者某一个人发音元音的第二共振峰的数值等等。

定比数据和定距数据的区别在于，定比数据有一个零点。比如，年龄、体重、身高都属于定比数据，不存在负值。定比数据除了可以加减运算以外，还可以乘除运算。我们可以说一个人的年龄是另外一个人的两倍。定比数据和定距数据的区分有一定的数据科学的意义，但是通常在语言研究当中我们不严格区分定比数据和定距数据。

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 数据层级 | 趋中性描述 | 变异性描述 | 数学运算 | 举例 |
| 定类 | 众数 | 不能 | 是否相等，集合关系 | 实验反应选项、词性、语音类别 |
| 定序 | 众数、中位数 | 不能 | 排序、比较 | 李克特量表问题数据 |
| 定距 | 众数、中位数、算数平均值 | 极差、方差、标准差 | 加减 | 反应时，声学数据 |
| 定比 | 众数、中位数、几何平均值 | 极差、方差、标准差 | 乘除 |  |

数据结构，类型，特点，决定了数据的预处理过程，数据探索性分析、建模及可视化十分重要。

语言研究者学习数据科学的优势与必要性

语言研究数据几乎包含了数据的各种类型，既有结构性数据也有非结构性数据。这对于语言研究者的数据处理能力提出了很高的要求。数据科学家大体上包含了三个能力模块：数学/统计学，专业知识，计算机编程。

语言研究者在数据处理方面的优势，最主要的体现在对于研究领域的知识背景熟悉。这对于数据预处理中的一些决策，如判断缺失值，偏异值以及采取相应的措施是十分重要的。此外，对于研究领域的背景知识的了解，对于探索性数据分析和数据可视化的思路上有很大的帮助。如果将语言研究中的数据处理任务外包给统计学或者计算机科学背景的专业人士，他们缺乏对于研究领域的知识，会阻碍数据的处理，探索性分析和建模。因此，新时代的语言研究者具备一定的数据科学知识与技能是十分重要的。

数学和统计学的能力是语言研究者，特别是文科背景的语言研究者担忧的方面。然而，数据科学是一个应用性科学，不同类型的数据处理对于统计能力的要求不同，并不一定都要求十分深厚的数学背景。比如，文本处理和一些语料库的基本数据统计，要求理解描述性统计的基本知识，结合不同的公式，理解其中区别就可以了。再例如一些推断性统计检验，需要理解一些基本的数据分布和概率假设。这些内容理解上并不难，具体的计算则可以通过计算机完成，并不需要非常强的人工计算能力。

第三个能力模块是电脑编程。所谓电脑编程就是用编程语言将自己需要做的事转化为指令传递给计算机。大部分文科背景的研究者都有学习外语的经历，电脑编程与之类似，需要学习编程语言的函数（相当于词汇）和语法格式。和学习自然语言一样，这是一个过程，并且其中会伴随一些困难，但这些都是暂时的，笔者认为在未来的十年里，在量化模式的语言研究中，通过编程的方式处理语言和语言应用中所产生的数据是大势所趋。

推荐书目

推荐资源：开源数据集

UCLA 语音学实验室档案

<http://archive.phonetics.ucla.edu/>

UCLA 语音学实验室档案。收录了世界各地的语言录音样本。这些数据集对于语音研究、音系学研究，都非常有帮助，涉及到了200个语言，并且配有相应的国际音标转写，以及田野调查的一些笔记。

电影电视字幕语料库（4680万字符，3350万单词）

词频是语言研究中最重要的变量。根据New、Brysbaert和其同事在英语、法语和荷兰语的工作，2010年，华东师大蔡清教授建立了一个基于电影和电视字幕语料库（4680万字符，3350万单词）的词频和字频数据库

网址：https://doi.org/10.1371/journal.pone.0010729.s002

参考文献：Cai, Q., & Brysbaert, M. (2010). SUBTLEX-CH: Chinese word and character frequencies based on film subtitles. PloS one, 5(6), e10729.

汉语词汇命名时间

研究者提供3314个传统汉字的心理语言学数据，收集了140名中国人的命名反应时间（RTs）。数据库中的词汇和语义变量包括频率、规则性、熟悉度、一致性、笔画数、同音异义度（homophone density）、语义歧义程度（semantic ambiguity rating）、语音组合能力（phonetic combinability）、语义组合能力（semantic combinability）以及由一个字符组成的双音节复合词的数量。

网址：http://ball.ling.sinica.edu.tw/namingdatabase/index.html

参考文献：Chang, Y. N., Hsu, C. H., Tsai, J. L., Chen, C. L., & Lee, C. Y. (2016). A psycholinguistic database for traditional Chinese character naming. Behavior Research Methods, 48(1), 112-122.

该词汇库总共有1100个中文单词，从效价、唤起度、具体性、熟悉度、形象度（imageability）和语境可用性评分。所有变量的评分均采用9分制李克特量表。

这一情绪词汇数据库将为利用汉语词汇进行情感研究提供有价值的信息来源，使研究人员能够使用高度受控的汉语语言刺激，更可靠地研究认知与情感之间的关系。

网址：https://static-content.springer.com/esm/art%3A10.3758%2Fs13428-016-0793-2/

MediaObjects/13428\_2016\_793\_MOESM2\_ESM.pdf

参考文献：Yao, Z., Wu, J., Zhang, Y., & Wang, Z. (2017). Norms of valence, arousal, concreteness, familiarity, imageability, and context availability for 1,100 Chinese words. Behavior research methods, 49(4), 1374-1385.

中文词汇库（Chinese lexical database, CLD）是一个大型的简体中文词汇库。提供了3913个单字、34233个双字词、7143个三字及3355个四字词的丰富词汇资料。

对于CLD中的48644个单词，研究者提供了广泛的分类预测因子，包括频率、复杂性、邻近密度（neighborhood density）、正字法音系一致性（orthography-phonology consistency）和信息理论（information-theoretic）度量。

网址：https://static-content.springer.com/esm/art%3A10.3758%2Fs13428-018-1038-3/MediaObjects/13428\_2018\_1038\_MOESM1\_ESM.txt

网址：http://www.chineselexicaldatabase.com

参考文献：Sun, C. C., Hendrix, P., Ma, J., & Baayen, R. H. (2018). Chinese lexical database (CLD). Behavior research methods, 50(6), 2606-2629.

该数据库包含1020个单字、1022个双字、949个三字和587个四字简体中文单词的词汇判断数据集（MELD-SCH）。

研究者发现单词长度与反应时间之间的U形关系，这在中文中还没有报道过。MELD-SCH能够提供高质量的规范数据和不同语言变量的信息，促进汉语词汇识别的研究。它还鼓励研究人员将他们的实证发现（主要基于一个字符和两个字符的单词）扩展到不同长度的单词。

网址：https://static-content.springer.com/esm/art%3A10.3758%2Fs13428-017-0944-0/MediaObjects/13428\_2017\_944\_MOESM1\_ESM.xlsx

参考文献：Tsang, Y. K., Huang, J., Lui, M., Xue, M., Chan, Y. W. F., Wang, S., & Chen, H. C. (2018). MELD-SCH: A megastudy of lexical decision in simplified Chinese. Behavior research methods, 50(5), 1763-1777.

汉字书写的心理语言学数据库

数据库包含1600个频率分布广泛的汉字，采用听写任务，总共203名被试来书写这些汉字，采集了被试的书写潜伏期、书写时长、书写正确率，并收集了1600汉字的14个词汇变量。

研究结果发现，字频、习得年龄、语境是影响正字法通达、运动执行和书写正确率的共同因素；语音变量（是否为形声字、规则性、同音字密度）影响正字法通达，但不影响运动执行；语义变量（表象性和具体性）只影响书写正确率。研究结果对汉字书写产生机制有着重要启发。作为第一个大规模的汉字书写的心理语言学数据库，该数据库可以作为二次数据分析的资源以及书写实验材料制作的工具。

网址：https://osf.io/7s9kq/

参考文献：Wang, R., Huang, S., Zhou, Y., & Cai, Z. G. (2019). Chinese character handwriting: A large-scale behavioral study and a database. Behavior research methods, 1-15.

词语小世界

词语联想实验是一个大规模的科学研究。该网站研究者致力于构建世界上主要语言的大脑词库，并普及这些信息。

与一般词库或者字典不同，词语联想词库帮助我们深入了解人脑中最重要的词及其含义的构成。心理学家、语言学家、神经学家等其他领域的研究人员也可以利用词语联想词库检验有关语言表达和语言处理的新理论。这些知识可以运用于诸多领域，包括文化差异，母语或二语的新词习得与忘记。

详情请见原网址。

网址：https://smallworldofwords.org/zh/project/home

8

语料库在线

该语料库是教育部语言文字应用研究所肖航教授建立的，网站提供在线现代汉语语料库检索、古代汉语语料库检索和语料库字词索引。

除此之外，提供了语料分析处理工具，如语料分词和词性标注、语料汉语拼音自动标注和语料字词频率统计。

网址：http://corpus.zhonghuayuwen.org/

9

CCL现代汉语语料库

CCL汉语语料库总字符数为783,463,175,其中现代汉语语料库总字符数为581,794,456。

CCL语料库及其检索系统为纯学术非盈利性质，语料库中的中文文本未经分词处理，检索系统以汉字为基本单位。主要功能特色在于：支持复杂检索表达式（比如不相邻关键词查询，指定距离查询等等）；支持对标点符号的查询（比如查询“?”可以检索语料库中所有疑问句）；支持在“结果集”中继续检索：用户可定制查询结果的显示方式（如左右长度，排序等）；用户可以从网页上下载查询结果（text文件）。

网址：http://ccl.pku.edu.cn:8080/ccl\_corpus/index.jsp?dir=xiandai

新词语研究资源库

由国家语言资源监测与研究有声媒体中心提供在线查询1995年至2016汉语新词语。

网址：http://ling.cuc.edu.cn/newword/showcls2.aspx?classid=94

1. 在描述性统计中我们有一些描述数据需中性的指标，比如平均数、中位数、众数。 [↑](#footnote-ref-1)