Modern Linguistics 现代语言学, 2019, 7(6), 940-945
Published Online December 2019 in Hans. http://www.hanspub.org/journal/ml
https://doi.org/10.12677/ml.2019.76123

Hans汉斯

The Design of Speech Database of Chinese Dialects Read Different Characters

Haixia Jia

Southwest Institute of Nationalities, Southwest Minzu University, Chengdu Sichuan
Email: 317187051@qq.com

Received: Nov. 12th, 2019; accepted: Dec. 3rd, 2019; published: Dec. 12th, 2019

Abstract

With the development of related subjects and computer technology, multi-disciplinary, multi-dimensional, multi-functional voice database has a broad development prospect. This thesis intends to use Chinese dialect as the research object, mainly for literary and colloquial readings and the old and new variant pronunciation, using sociolinguistics, historical comparative linguistics, computational linguistics, geographical linguistics, linguistic typology and other theories, to design a speech database and exhaustively examine the phenomenon of literary and colloquial readings, and the old and new reading, in order to clarify the characteristics, sources, types, changes and differences between the two phenomena through the research methods of big data.

Keywords

Chinese Dialects, Literary and Colloquial Readings, The Old and New Variant Pronunciation, Speech Database

汉语方言异读词语音数据库的设计

贾海霞

西南民族大学西南民族研究院，四川 成都
Email: 317187051@qq.com

收稿日期：2019年11月12日；录用日期：2019年12月3日；发布日期：2019年12月12日

摘　要

随着相关学科及计算机技术的发展，多学科、多维度、多元化、多功能的语音数据库具有广阔的发展前景。本文拟以汉语方言异读词为研究对象，主要是文白异读和新老异读，运用社会语言学、历史比较语

文章引用: 贾海霞. 汉语方言异读词语音数据库的设计[J]. 现代语言学, 2019, 7(6): 940-945.
DOI: 10.12677/ml.2019.76123

言学、计算语言学、地理语言学、语言类型学等理论和方法，设计一个语音数据库，穷尽式地考查汉语方言文白异读和新老异读现象，以期通过大数据的研究方法厘清这两种现象的特点、来源、类型、变迁及二者的区别等问题。

关键词

汉语方言，文白异读，新老异读，语音数据库

1. 引言

语言学的研究对象以“活着”的语言为主，只有在大量语言材料基础上得出的结论才能站得住脚。传统研究中，主要通过口耳相传搜集、整理和加工语言材料，这是一项复杂而又耗时费力的工作。直到计算机出现并用来处理语言数据后，给语言材料的处理工作带来了很大便利。目前，国内外基于语料库的语言学研究都取得了重大进展。随着相关学科及计算机技术的发展，多学科、多维度、多元化、多功能的语音数据库具有广阔的发展前景。

本文拟设计一个异读词专项语音数据库，广泛收集现有的方言异读材料，并有针对性地进行实地调查，补充和完善，综合处理语音数据库中的性别、年龄、地域等信息，进一步提高精准度，以梳理汉语方言异读词的地理分布和历史演变特点，厘清文白异读和新老异读的特点、来源、类型、变迁及二者的区别与联系，也可用于语音识别、方言特征词识别等研究。

2. 概念界定

2.1. 异读

汉字形、音、义的关系错综复杂，异读是一种最常见的现象。本文所说的异读，主要是指文白异读和新老异读。文白异读分为文读音和白读音，新老异读分为新派音和老派音。

关于文白异读的界定有以下几个维度：1) 从语音的语体色彩角度，认为白读是口语形式，文读是书面语形式。文读音用于新词术语、书面语及通用语，而白读音则用于古语词、特色词、土语词及地名。另一种情况是文白的风格差异体现为语用限制，与词汇色彩无关。2) 从历时角度，探讨文白异读的来源问题，认为文读和白读意义上有关联，具有相同的来历，在《切韵》里的音韵地位相同，与古音来源有一定的对应关系[1]。3) 从语言接触角度，认为文度和白读的共存是方言接触在同一系统中的历时体现，也是语音层次分析的依据[2]。4) 从音变方式角度，认为音变发生在一个音节中声母、韵母或声调三个语音单位的某一项，而不是整个字音音节。也就是说，文白差异只是声、韵、调的差异，而不是整个字音的差异[3]。

李如龙认为新老异读是表现了同一个时代的新旧两种读音，这类异读大多是近代方言受普通话的影响所产生的现象[4]。换句话说，新老异读属社会语言学的概念。

曾缇认为“每一种方言的语音层面都有文白异读和新老异读的情况，从历时角度看，文白异读现象是不同历史层次字音的叠置现象，不同时期共同语和方言的语音特点都可能沉淀其中”[5]。文白异读是

本地语音和外来语音两个系统的叠置现象，而新老异读还没有形成稳定的系统，主要是受社会因素的影响造成同一个字的异读现象。

2.2. 文白异读与新老异读的区别

文白异读与新老异读的区别文白异读和新老异读均属语音异读现象，但二者有明显区别，主要体现在以下方面：

1) 所指的时间段不同。王洪君认为文白异读的音类是中古音发展演变的结果，是来源不同、进入当地时间不同的两个(或几个)音韵层次在共时音系中的叠置。各音韵层次间有成系统的同源对应(姊妹方言关系)，但它们不是直接的继承发展关系，而是反映某一历史层面上土语音系与权威方言音系的接触与竞争[6]。而新老异读是一个社会语言学概念，主要指由于代际(一般可分为老中青三代)差异引起的异读，就目前来说，主要指方音受普通话影响而产生的异读。

2) 异读是否成系统。一般文白异读是某一类中古音无例外地发生系统性的异读，偶尔也有一些例外，但是可以利用语音演变规律做出一定解释。但新老异读不成系统，也不稳定，它受发音人年龄、受教育程度、语言环境、职业等社会因素影响。如现在的年轻人文化程度越高，普通话水平相对较高，与外界接触较多，往往存在母语遗忘的趋势，同时也会很自然地将普通话语音融入自己的发音。

3) 发展趋势差异。由于普通话推广工作的不断深入，新派读音会有很大发展，老派读音会渐渐萎缩。即使是由祖父辈抚养的留守儿童，虽习得的是老派读音，但日后如果他们求学、工作，语音也会向新派发音靠拢。文白异读的发展趋势比较复杂，可能是文读音消失或萎缩，白读音继续发展，也可能出现新文读和白读，但是这一过程比较缓慢，要遵循语音本身的发展演变规律。尤其是地名中的文白异读，存古性更高。

以上是基于目前材料所得结论，有待于通过大数据库加以验证和补充完善，尤其是二者的判定目前还没有定论，需要更多的材料来佐证。

3. 汉语方言异读词语音数据库的设计

语音语料库设计的关键问题在于语料，即语音数据的内容或录音文本。

3.1. 语料库及设计准则

语料库作为样本必须具有代表性。本文首先通过现有的资料穷尽式收集汉语十大方言区异读词，然后根据各方言区分布情况及特点适当补充完善，最后形成语料库发音文本。主要的设计准则有：1) 每个异读词都要收集一些语言使用中的自然语料；2) 提取当地方言的特色词作为该方言区的特征词汇；3) 注意当地地名的读音，因为地名中的文白异读现象较为稳定；4) 有意识地结合发音人的年龄、性别、职业、受教育程度等因素，以便做社会语言学考察。

基于以上准则收集的原始语料文本由固定词、句子、短文和自由话题四部分组成。

3.2. 发音合作人的选择

发音合作人的选择非常重要，直接决定了样本的代表性。为确保采集方言语料的质量和代表性，所选择的方言发音人必须是土生土长、口齿清晰、发音地道、用于传统、语速适中。鉴于本文的研究目的，我们既要选择那些文化程度不高、生活范围狭小，善于交际聊天、很少受普通话影响的年龄在五、六十岁的发音人，还要有意识地控制发音人的年龄、性别、职业、受教育程度等因素，寻找各年龄段发音人作为参照，这样才能全方位、科学地判定文白异读和新老异读，进而揭示其演变规律，预测其发展趋势。

录音时，发音合作人的发音方式包括自然语音、描述语音、回答语音和朗读语音，语调为中性[7]。

3.3. 数据库建设技术及软件

1) 语音录制及处理

尽量选配专业录音房、录音麦克风等设备，或安静的室内房间，无背景噪声。录音软件用 Cool Edit，采用单声道录制，采样率 11,025 kHz，以 PCM 格式 16 比特量化编码的 wav 格式保存。如果录音条件受限，也可结合电话录音，软件用 IRNO3000，采用 8 kHz 采样率，语音存储为 16 比特量化、PCM 格式编码的 wav 格式。

每次录音结束后，都由通晓该方言的人员人工检验语音文件，及时修正和补录。然后对噪声进行预处理，如过长的静音段、电流噪声、咳嗽声、喘息声和笑声等。

2) 文本标注

为后期采用语音合成技术模拟仿自然语句，需对句子、短文和自由话题进行标注，采用 Praat 软件完成。参照汉语音段标注系统 SAMPA-C 进行分级标注。语音库的标注主要包括文字标注和音节标注两部分。文字标注部分用汉字 + 拼音形式，以便为语音识别、切分和合成系统使用。对于副语言学现象，可用相应的副语言学符号表示。音节标注部分需用国际音标标出相应的当地方言语音，以便与中古音及其他方言点语音进行对比。另外声调标注可设定相应符号，如 0 表示轻声，1 表示阴平，2 表示阳平，3 表示上声，4 表示去声。具体情况可根据语料情况来设定。

3.4. 数据库建设技术及软件

1) 数据库功能

检索功能。数据库应具有强大的检索功能，可以通过语音和文本一框式检索，也可使用多条件组合检索，能实现普通话及各方言之间的双向浏览，如输入普通话字词，可以检索到各方言各年龄段、职业等相应的异读字词或语音；能通过语音输入，从调类、调值、语音特点等方面匹配查出该语音的市、县、乡地名信息；能根据中古音类检索其在各方言区的演变结果，并绘制出该语音类型地理分布图；也可根据用户的实际需求和浏览习惯来确定检索条件和方式。

学习功能。该数据库的建成将成为人们研究汉语方言异读字(词)，尤其是文白异读和新老异读不可或缺的工具。可以直接点击数据库中的字词，并选择方言点，便可听到该方言的发音和相关语音资料等信息。如技术条件允许，我们将适当运用语音切分和合成技术，用户不仅可以听到语料库中原生态的发音，而且还可以听到采用语音合成技术模拟而成的仿自然语句。

分析功能。系统可以运用内置的语音演变条件和路径规律，对照各方言文白异读字今读音与中古音，推导出其演变条件和路径，并与其他汉语方言区该音类进行横向对比加以考察，以分析其语音层次。另外还可以对比分析文白异读与新老异读的区别，分析哪些是文白异读，哪些是新老异读。

下载功能。用户可以将检索、分析和比较的结果输出到 Excel 表格，然后下载和打印；可以输出用户浏览和使用情况的统计数据和分析等；可以按规定格式下载语音文件。

维护功能。这一功能只能由系统管理员操作。一是数据编辑功能，可复制、粘贴、剪切、替换、插入数据等；二是系统维护功能，可管理和操作数据、用户信息、日志及系统升级等；三是拓展功能，可根据用户反馈信息和研究需求对数据库进行再设计，或添加新模块，以加强或拓展数据库功能。

辅助功能。提供各方言点的语音系统介绍，数据库建库原则和使用说明，以及一些简单的汉语方言异读词调查表和相关语音、释义对照表。

4. 需要注意的几个方面

4.1. 异读词的界定

文白异读和新老异读至今没有明确的界定标准，我们之所以建立这样一个语音数据库就是为了通过大数据的方法，为其界定和分类提供依据，所以在数据库建设过程中，如果把握不准不能随意判定，需与各地方言和普通话进行横向比较，再与中古音进行纵向比较，以期得出合理、可信的解释。另外，随着数据量的不断扩大和认识的不断深入，还需回头审查之前的结论，以保证数据库的正确性和可信度。

4.2. 调查点和发音合作人的选择

调查点的选择可以借鉴现有成果来选定。对于已经有丰富文白异读现象的方言点，应根据其特点来选代表点。对于没有文白异读研究成果的方言点，需根据已有方言点的分布情况推测其是否有可能存在此现象，再进一步深入挖掘，使该数据库尽可能覆盖各大方言区。

由于该数据库采用社会语言学方法选取发音合作人，最后还要在此基础上判定新老异读与文白异读的区别和联系，以及二者的发展趋势，所以在选择发音合作人时，应充分考虑其社会因素，如性别、年龄、职业、受教育程度等，以保证样本和数据的可对比性。

4.3. 数据库标准化

目前，我国数据库系统的研制和开发没有统一的标准和建库规范，这也是很多数据库虽然建成了，但是并没有公开或广泛使用的主要原因之一，所以，我们在建库时应统筹考虑现有语音数据库的优缺点，尽可能扩大其适用领域。

4.4. 数据采集

数据是数据库各种功能得以实现的最基本保证，而方言设计种类多，范围广，难免存在与真实语音相悖的情况，尤其异读词更是一个方言特色词的代表。尽管在设计之初，我们规定发音合作人的发音方式包括自然语音、描述语音、回答语音和朗读语音，语调为中性。但由于普通话推广工作的广泛深入，发音合作人难免受普通话或调研员语音影响，在采集数据时，我们应想尽一切办法采集最纯、最地道的方言，以保证其客观性和特殊性。

5. 结语

文白异读是汉语中较常见的语言现象，我国学者很早就开始关注这一现象了，可至今没有彻底搞清其来源、层次、类型、特点及演变趋势。随着普通话的推广，社会语言学又将新老异读列为研究对象，关于文白异读与新老异读的区分又成为新的研究热点，有待挖掘的空间很大，这必将进一步加深我们对语音演变路径的认识。由此看来，我们建立汉语方言异读词语音数据库将为文白异读和新老异读研究提供大数据支持，不仅有利于汉语方言异读词对比研究，而且有利于从地理学角度探求汉语方言的演变路径及规律，具有重要的应用价值和意义。

基金项目

教育部人文社会科学研究青年基金项目“《通用彝文规范方案》认知度及现行字库使用情况调查研究”资助(16YJC740028)。西南民族大学中央高校基本科研业务费专项资金“基于彝语动态流通语料库的汉语借词使用及语序特点研究”资助(2019NQN31)。

参考文献

[1] 张振兴. 漳平(永福)方言的文白异读[J]. 方言, 1989(3): 171-179.
[2] 王福堂. 文白异读和层次区分[J]. 语言研究. 2009(1): 1-5.
[3] 陈忠敏. 重论文白异读与语音层次[J]. 语言研究, 2003, 23(3): 43-59.
[4] 李如龙. 汉语方言学[M]. 北京: 高等教育出版社, 2001: 81.
[5] 曾缇. 奇台方言中的文白异读及新老异读[J]. 昌吉学院学报, 2016(2): 54-59.
[6] 王洪君. 历史语言学方法论与汉语音韵史个案研究[M]. 北京: 商务印书馆, 2014: 314.
[7] 洪拓夷. 汉语方言语音数据库建设构想[J]. 图书情报工作, 2009, 53(5): 83-86.