Modern Linguistics 现代语言学, 2026, 14(11), 41-47
Published Online November 2026 in Hans. https://www.hanspub.org/journal/ml
https://doi.org/10.12677/ml.2026.14111045

Hans汉斯

近三十五年湘方言语料库建设研究综述

胡　鑫

湖北文理学院中文系，湖北　襄阳

收稿日期：2026年9月27日；录用日期：2026年10月6日；发布日期：2026年11月4日

摘　要

从1990年到2024年，湘方言语料库建设成果丰硕，共建成各类语料库23个，研究内容涉及语音、词汇、语法等方面，形成了小型专用、中型综合、大型多媒体并存的多维格局。总体来看，各类语料库建设成效显著，但区域覆盖不均、标注深度不足、技术应用不充分、共享机制不健全等问题仍然突出。系统梳理这一时期湘方言语料库的建设成果与发展规律，可为湘方言研究的科学化、数字化及地域文化传承提供参考借鉴。

关键词

湘方言，语料库，方言数字化，方言保护

A Review of Research on the Construction of Xiang Dialect Corpora over the Past Thirty-Five Years

Xin Hu

Department of Chinese Language and Literature, Hubei University of Arts and Science, Xiangyang Hubei

Received: September 27, 2026; accepted: October 6, 2026; published: November 4, 2026

Abstract

From 1990 to 2024, the construction of Xiang dialect corpora has yielded substantial results, with a total of 23 corpora of various types having been built. The research content covers phonetics, vocabulary, grammar, and other aspects, forming a multidimensional pattern in which small-scale specialized corpora, medium-sized comprehensive corpora, and large-scale multimedia corpora coexist. Overall, the construction of these corpora has achieved remarkable results, but problems such as uneven regional coverage, insufficient annotation depth, inadequate technological application,

文章引用: 胡鑫. 近三十五年湘方言语料库建设研究综述[J]. 现代语言学, 2026, 14(11): 41-47.
DOI: 10.12677/ml.2026.14111045

and an imperfect sharing mechanism remain prominent. A systematic review of the achievements and developmental patterns of Xiang dialect corpus construction during this period can provide reference for the scientific and digitalized study of Xiang dialect and the inheritance of regional culture.

Keywords

Xiang Dialect, Corpus, Dialect Digitalization, Dialect Protection

Copyright © 2026 by author(s) and Hans Publishers Inc.
This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).
http://creativecommons.org/licenses/by/4.0/
Open Access

1. 引言

湘方言作为汉语七大方言之一，主要分布于湖南境内湘江、资江、沅江流域及广西东北部部分地区。根据鲍厚星、陈晖(2005)的研究[1]，湘语内部可划分为长益片、娄邵片、辰溆片、衡州片和永州片五大片区，使用人口超过三千万，被视为湖南地域文化的核心载体。湘方言在古全浊声母的演变规律、语气词系统以及词汇语法结构等方面具有鲜明特色，不仅对汉语类型学研究具有重要意义，也为探讨汉语历史演变和语言接触现象提供了关键样本。

然而，传统湘方言研究长期依赖田野调查，以人工记音和定性分析为主要手段(鲍厚星、陈立中、彭泽润，2000) [2]，在语料规模、量化分析与成果复用等方面存在明显局限。20 世纪 90 年代以来，语料库语言学兴起并渐趋成熟(丁信善，1998；王建新，1998；杨惠中，2002；Biber *et al.*, 1998；McEnery & Hardie, 2012) [3]-[7]，计算机与多媒体技术的进步(常宝宝、俞士汶，2009) [8]带来“数据驱动”的研究转向，推动湘方言语料库建设逐步走向数字化与智能化。鉴于此，本文以 1990~2024 年建成的 23 个湘方言语料库为考察对象，从建设规模、研究核心、技术特征等维度梳理其建设成果与发展规律，指出存在的问题，以期为湘方言研究的科学化、数字化及地域文化传承提供参考借鉴。

为保证综述的严谨性与可追溯性，本文对文献的检索与纳入过程作如下说明：检索时间窗口为 1990 年 1 月至 2024 年 12 月，主要检索中国知网(CNKI)、万方数据、维普中文期刊服务平台等中文数据库，并辅以国家语委及中国语言资源保护工程相关成果发布平台、湖南师范大学等高校学位论文库，国外文献以 Web of Science 等外文数据库补充检索；检索词以“湘方言”“湘语”“长沙话”“双峰话”“湘南土话”等方言名称，分别与“语料库”“多媒体语料”“有声数据库”“方言标注”等主题词组合检索，并追溯相关专著、学位论文与项目报告所附参考文献。纳入标准为：以湘方言(含湘语五大片区及湘南土话等归属存在争议的湘语相关方言)语料库建设为核心议题，或包含成规模的语料采集、标注与建库内容；排除仅以零星语料为例证、未形成系统性语料库的定性研究，以及重复发表、无法获取全文的文献。此前，谭晓平已对近十年汉语语料库建设研究进行过系统综述[9]，本文在吸收其框架的基础上，将考察时段延伸至 2024 年，并聚焦建设规模、研究核心与技术特征三个维度展开梳理。

2. 湘方言语料库建设的总体情况

(一) 发展历程

近三十五年，湘方言语料库建设与语料库语言学的发展同步演进，大致经历三个阶段：1990~2000 年为早期探索期，语料库多依附方言调查报告和工具书编纂而建，以文本型为主；2001~2010 年为稳步发展

期，多媒体技术开始应用于语料建设，中型综合语料库相继出现；2011~2024 年为快速发展期，尤其自 2015 年中国语言资源保护工程启动以来(曹志耘，2015) [10]，大型多媒体语料库集中建成，智能化探索逐步展开。

(二) 总体规模与分布

截至 2024 年，经对相关期刊论文、学位论文、项目报告等文献的系统梳理，国内外已建成各类湘方言语料库 23 个，其中国内建设成果 21 个，占绝对主导；国外成果仅 2 项，且均为基于既有语料库的对比研究，尚未形成独立的建设实践。从研究主体看，国内核心力量集中于湖南师范大学、湖南大学等高校，鲍厚星、陈晖、罗昕如、李斌等学者构建了“语料库建设–专题研究”的闭环模式，为湘方言研究提供了重要的数据支撑与方法示范。此外，2015 年“湖南方言调查‘响应’计划”正式启动，由汪涵发起，以媒体与社会力量推动湖南方言的数字化记录与保存[11] [12]。

(三) 国外研究状况

国外湘方言语料库相关研究相对薄弱，尚未形成独立的语料库建设成果，相关研究主要集中在两个方向：一是汉学家基于现有中文语料库开展的类型学对比研究，如日本学者中嶋幹起对湘方言语音语料的整理与分析，并与吴语、粤语等方言进行对比；二是对湘方言与周边方言接触现象的研究，如基于湘南土话词汇语料探讨湘方言与客家话、西南官话的接触机制。整体而言，国外研究缺乏对湘方言语料库建设技术与应用的深度参与，成果数量与研究视野均不及国内。

3. 按建设规模划分的湘方言语料库建设

湘方言语料库的建设规模与研究目标高度相关，形成“小型专用–中型综合–大型多媒体”的清晰梯度，各类语料库在技术要求、研究定位上各有侧重。

(一) 小型专用语料库(10 万字以内，共 10 个)

此类语料库聚焦单一方言点的特定研究主题，规模为 1~10 万字，以自然口语语料为主，标注围绕研究主题展开，建设周期短、针对性强。如李斌(2013) [13]的双峰方言语气词语料库，收录 1.8 万字口语语料，标注 18 个单用语气词、31 组双连语气词、6 组三连语气词，支持语气词连用检索与实验语音分析；何铃玉(2015)的安仁方言语气词语料库[14]，收录 8 万字口语语料，运用声飞、ELAN 等软件重点标注句末语气词的分布与功能；罗昕如(1998)的新化方言代词语料库收录 5 万字语料[15]，聚焦人称代词、指示代词、疑问代词的用法；彭泽润(1999) [16]的衡山方言语音语料库收录 6 万字语料，标注古全浊声母演变、声调分化与文白异读。小型专用语料库以专题深度见长，为湘方言的精细化研究奠定了语料基础。

(二) 中型综合语料库(10 万~100 万字，共 8 个)

此类语料库覆盖单一方言点或某一片区的语音、词汇、语法多维语料，语料类型包括口语对话、书面语料、民间文艺等，标注涉及音标、词性、句法结构，具备基础检索与统计功能，多为方言单点研究专著的配套成果。如鲍厚星等(1999)的长沙方言综合语料库[17]，收录了 30 万字语料，涵盖语音系统、核心词汇与常用语法句式；夏剑钦(1998)的浏阳方言研究语料库[18]，收录 24 万字语料，含语音转写、词汇标注与语法标注；陈晖(2006)的娄邵片方言对比语料库[19]，收录 25 万字语料，覆盖娄底、邵阳、双峰等 6 个方言点；沈若云(1999)的湘南土话词汇语料库收录 35 万字语料[20]，系统梳理了湘南土话特色词汇与古语词保留情况。整体而言，中型综合语料库实现了从“专题研究”到“系统描写”的跨越，支撑了湘方言单点研究专著的撰写与片区内部特征对比。

(三) 大型多媒体语料库(≥100 万字/含音视频，共 5 个)

此类语料库规模宏大、覆盖范围广，整合音频、视频、文本、音标等多模态数据，多采用标准化编码，具备高级检索与多模态分析功能，多为国家级或省级项目成果，兼顾研究与保护功能。如中国语言

资源有声数据库(湖南部分)为国家语委重大项目(李宇明，2010) [21]，覆盖 48 个方言点，含 120 万字文本语料与 300 小时音视频语料；李斌(2012) [22]的湘方言多媒体语料库收录 8 个代表点 150 万字语料、200 小时音视频，实现“音视频－文本－音标－译文”逐句对齐；湘南土话基础语料音库(2022)为省级项目，收录 10 个湘南土话方言点 80 万字语料、150 小时音频，其中邓永红(2014)调查了湘南 16 个市县[23]；侯精一主编的《现代汉语方言音库》含长沙话、湘潭话音档[24]，收录 2000 个单字、1000 条词语、500 句短句的语音数据。

4. 按研究核心划分的湘方言语料库建设

根据标注重点与研究定位，湘方言语料库可分为语音型、词汇型、语法型三类，三类并非完全割裂，部分综合语料库跨类覆盖，合计 23 个(含 3 个跨维度综合库重复计入)。

(一) 语音型语料库(7 个)

语音型语料库以语音数据为核心，标注内容涵盖声韵调值、文白异读、变调规则及轻声儿化现象，语料以单字、双音节词、短句为主，均配套音频数据，多用 Praat、Audacity 等工具，服务于语音系统描写、语音演变追踪与方言分区验证。如《现代汉语方言音库》的长沙话、湘潭话音档，收录基础单字、常用词语与语法短句的录音，是湘方言语音研究的基础参照库；彭泽润(2003)的衡山方言语音语料库，覆盖前山话、后山话两个子方言点，标注古全浊声母“舒声保留浊音、入声清化”等特征，如“白”前山话读[ba]、后山话读[pʰa]，支撑地理语言学视角下的语音变异研究(彭泽润、鲍厚星，2005) [25]；瞿建慧在博士论文《湘语辰溆片语音研究》(2008) [26]基础上建成湘语辰溆片语音语料库(2021)，收录辰溪、溆浦、沅陵等 6 个方言点的语音数据，专项标注阳去调值分化与蟹假果摄元音演变，在以往零散调查的基础上，为辰溆片语音的系统性语料建设提供了重要补充。此类语料库以量化统计验证了长益片古全浊声母“平声不送气、入声部分送气”、娄邵片“入声调类保留完整”等语音规律，为湘语五片分区体系提供了实证依据。

(二) 词汇型语料库(8 个)

词汇型语料库以词汇数据为核心，标注内容包括词汇读音、词义解释、搭配规则与来源，语料涵盖核心词汇、民间俗语及专业领域词汇，主要用于词汇对比、语言接触与方言词典编纂。如《湖南汉语方音字汇》配套语料库收录 26 个方言点的 8000 个单字[27]，标注“字形－方言读音－普通话读音”对应关系，是早期词汇对比研究的核心数据源，如“吃”在长沙话读[tɕʰi]、娄底话读[tɕʰia]、永州话读[tɕʰie]，清晰呈现湘方言词汇读音的区域差异；罗昕如(2006)的湘方言词汇对比语料库[28]，覆盖 10 个代表点 2000 条核心词汇，揭示了湘方言词汇的“南北差异”——北部受官话影响较深，南部保留更多中古汉语词汇；沈若云(1999)的湘南土话词汇语料库聚焦宜章、桂阳等湘南土话点，标注“爷”“娘”“跕”等特色词汇，印证了湘南土话与客家话的语言接触历史。此类语料库有效支撑了《长沙方言词典》《娄底方言词典》等工具书的编纂。

(三) 语法型语料库(8 个)

语法型语料库以语法数据为核心，标注内容包括词性、句法结构、虚词用法与特殊句式，语料以自然口语为主，主要用于语法特征描写、虚词功能分析与句法演变研究。如李斌的双峰方言语法语料库，基于 ELAN 工具建设，专项标注 18 个语气词的单用与连用规律，修正了“湘方言语气词连用自由”的传统认知；李永明在《衡阳方言》基础上建成衡阳方言语法语料库(2008) [29]，收录 30 万字语料，聚焦“把”字句与重叠式形容词，系统梳理了衡阳方言语法体系；陈晖在《湘方言语音研究》基础上先后建成娄邵片方言对比语料库与湘语娄邵片语法语料库，后者覆盖 6 个方言点，标注声母送气影响声调分化、无入声调类等现象，为汉语语法类型学提供了新样本。

5. 按技术特征划分的湘方言语料库建设

根据语料载体与技术应用水平，湘方言语料库可分为文本型、多媒体型和智能化探索型三类，反映了从“机读化”到“数字化”再到“智能化”的技术演进脉络。

(一) 文本型语料库(6 个)

此类语料库以文字文本为唯一载体，无音频视频数据，语料存储为 Word、Excel 文档或纸质档案，标注仅含基础文字转写(部分含国际音标)，无检索系统，需人工查阅，是早期“机读化起步”阶段的成果，以 1990~2000 年为主。如《湖南汉语方音字汇》配套语料库(1993)、杨时逢《湖南方言调查报告》所附调查语料集(1974) [30]、鲍厚星《东安土话研究》词汇表(1998) [31]等。文本型语料库无法还原口语语音特征，语料复用性较低，仅能支撑基础词汇或语音对比。

(二) 多媒体型语料库(15 个)

此类语料库整合文本、音频、视频多模态数据，多采用 ELAN 工具实现音视频与文本逐句对齐(李斌，2012)，标注涵盖文本层、国际音标层、普通话译文层，部分含语用场景标注，具备关键词检索、时间点定位等基础功能，以 2001~2024 年为主。如李斌(2013)的双峰方言多媒体语料库，语料关联音频视频，支持语气词连用检索与语音波形分析；中国语言资源有声数据库(湖南部分)含 300 小时音视频，覆盖 48 个方言点，收录民间故事、歌谣等口头文化语料，采用 TEI 标准化编码；李斌等(2012)的湘方言多媒体语料库通过 200 小时音视频实现“音视频－文本－音标－译文”四层对齐。多媒体型语料库突破了传统文本局限，实现了方言的“立体化保存”，但其建设规范与检索功能仍需完善(范俊军，2013) [32]。

6. 结语

回顾近三十五年湘方言语料库建设历程，整体呈梯次演进、多维并进之势：规模上，小型专用、中型综合、大型多媒体梯度完整；载体上，由文本型为主转为多媒体型为主，并开启智能化探索；研究上，语音、词汇、语法三类语料库并进互补，多与专题研究深度结合。

同时也要看到，当前湘方言语料库建设仍存在若干亟待补齐的短板。其一，区域覆盖失衡。语料库建设集中于长益片、娄邵片等核心片区，湘南土话、湘西乡话等濒危方言以及部分县域仍未建成系统性语料库，方言资源的系统采集与数字化记录覆盖面不足。其二，标注深度有限。多数语料库仅完成语音转写、词性标注等基础加工，语义、语用、语篇等层级标注普遍缺失，制约了语料在句法、语义等深层研究中的利用价值。其三，技术应用滞后。检索与处理多停留在关键词检索、词频统计等浅层，语音自动转写、方言识别等智能化功能尚未大规模落地，难以支撑高效率的语料加工。其四，共享机制缺失。各库建设标准、标注规范与数据格式不统一，多数中小型语料库处于封闭或半封闭状态，资源重复建设、利用率偏低。其五，应用场景狭窄。跨方言、跨语言对比与国际合作研究薄弱，方言 AI 教学、数字文化传播等成果转化与应用拓展仍显不足。

深入剖析可见，上述短板并非单纯的技术问题，而是体制、资金与技术三重因素长期交织的结果。从体制层面看，现有成果多为课题驱动型建设：中国语言资源有声数据库(湖南部分)、湘南土话基础语料音库等国家与省级项目以结项验收为导向，项目结束后缺乏持续的机构化运维；而李斌(2013)、何铃玉(2015)等依托学位论文建成的语料库，多随论文完成而封存于个人或院系，缺乏制度化承接，导致“重建设、轻运维、难共享”。从资金层面看，中小型语料库多依赖课题经费与研究者个人投入，经费一次性、短期化，难以支撑标注深化与长期维护；《现代汉语方言音库》等早期成果亦因建成年代较早、缺乏持续更新机制而渐成静态资源。从技术层面看，湘方言自动转写、语音识别等工具研发滞后，现有标注多依赖人工逐层完成，如 ELAN 下的“音视频－文本－音标－译文”四层对齐耗时费力，直接制约了标注深度与语料规模的扩展，也使跨库共享面临格式转换的技术门槛。

针对上述问题，未来湘方言语料库建设应从标准、组织与技术三个层面提出更具操作性的推进方案。第一，制定面向湘语内部差异的最小化标注规范。湘语内部新湘语与老湘语差异显著，标注规范应兼顾两类方言的语音、词汇与语法特征：音系层统一采用国际音标并标注调值(五度制或调值数字)，其中老湘语(以娄邵片为代表)应专项标注古全浊声母的演变类型(如舒声保留浊音、入声清化等)，新湘语(以长益片为代表)则应侧重官话接触带来的语音与词汇变异标记；词汇层统一标注词形、词义、词性与来源(古语词、官话借词、方言词)，并区分核心词与特色词；语法层统一标注词性与虚词功能，尤需为语气词建立“单用–连用–叠用”的功能位标注，为特殊句式(如“把”字句、重叠式形容词)设定统一的句式标签。规范遵循最小集原则，各库在共享核心字段的基础上可扩展自定义标注层，以兼顾可比性与灵活性。

第二，倡议由湖南师范大学等核心机构牵头，联合湖南大学、湘潭大学、衡阳师范学院等高校及相关项目组，基于现有开源平台建立“湘方言语料库共享联盟”，并制定分步实施路线图：近期(约 1 年)成立联盟并制定统一的数据标准，包括元数据方案、语料编码(参照 TEI 标准)、标注规范与共享协议，完成存量语料库的清点与版权梳理；中期(1~2 年)依托开源技术栈(如 ELAN 标注工具、Praat 语音分析软件与开源检索框架)搭建共享平台，将存量语料分批标准化转换并上线；远期(2~3 年)开放检索与下载接口，建立增量采集、联合标注与多方共建机制，形成“共建–共享–共治”的可持续生态。

如此，湘方言语料库建设方能在“研究–保护–传承”的链条上形成闭环，为湘方言研究的科学化、数字化及地域文化传承提供坚实支撑。

参考文献

[1] 鲍厚星, 陈晖. 湘语的分区(稿) [J]. 方言, 2005(3): 261-270.

[2] 鲍厚星, 陈立中, 彭泽润. 二十世纪湖南方言研究概述[J]. 方言, 2000(1): 47-54.

[3] 丁信善. 语料库语言学的发展及研究现状[J]. 当代语言学, 1998(1): 5-13.

[4] 王建新. 语料库语言学发展史上的几个重要阶段[J]. 外语教学与研究, 1998(4): 53-59.

[5] 杨惠中. 语料库语言学导论[M]. 上海: 上海外语教育出版社, 2002.

[6] Biber, D., Conrad, S. and Reppen, R. (1998) Corpus Linguistics: Investigating Language Structure and Use. Cambridge University Press. https://doi.org/10.1017/cbo9780511804489

[7] McEnery, T. and Hardie, A. (2012) Corpus Linguistics: Method, Theory and Practice. Cambridge University Press. https://doi.org/10.1017/cbo9780511981395

[8] 常宝宝, 俞士汶. 语料库技术及其应用[J]. 外语研究, 2009(5): 43-51.

[9] 谭晓平. 近十年汉语语料库建设研究综述[C]//北京大学对外汉语教育学院. 第七届北京地区对外汉语教学研究生论坛文集. 北京: 北京大学出版社, 2014: 26-31.

[10] 曹志耘. 中国语言资源保护工程的定位、目标与任务[J]. 语言文字应用, 2015(4): 10-17.

[11] “湖南方言调查‘响应’计划”项目启动[J]. 方言, 2015(3): 231.

[12] 汪涵发起湖南方言调查[J]. 语文教学与研究, 2015(9): 3.

[13] 李斌. 用 ELAN 自建汉语方言多媒体语料库及其应用研究: 以双峰方言语气词的研究为例[D]: [博士学位论文]. 长沙: 湖南师范大学, 2013.

[14] 何铃玉. 湖南安仁方言语气词研究及其语料库建设[D]: [硕士学位论文]. 长沙: 湖南师范大学, 2015.

[15] 罗昕如. 新化方言研究[M]. 长沙: 湖南教育出版社, 1998.

[16] 彭泽润. 衡山方言研究[M]. 长沙: 湖南教育出版社, 1999.

[17] 鲍厚星, 崔振华, 沈若云, 伍云姬. 长沙方言研究[M]. 长沙: 湖南教育出版社, 1999.

[18] 夏剑钦. 浏阳方言研究[M]. 长沙: 湖南教育出版社, 1998.

[19] 陈晖. 湘方言语音研究[M]. 长沙: 湖南师范大学出版社, 2006.

[20] 沈若云. 宜章土话研究[M]. 长沙: 湖南教育出版社, 1999.

[21] 李宇明. 论中国语言资源有声数据库的建设[J]. 中国语文, 2010(4): 356-363+384.
[22] 李斌. 用 Elan 建设单点方言多媒体语料库[J]. 方言, 2012(2): 178-190.
[23] 邓永红. 湘南土话基础语料音库建设初论[J]. 暨南学报(哲学社会科学版), 2014, 36(4): 6-10+163.
[24] 侯精一. 现代汉语方言音库[M]. 上海: 上海教育出版社, 1994-1999.
[25] 彭泽润. 衡山南岳方言的地理语言学研究[J]. 语言文字应用, 2005(1): 144.
[26] 瞿建慧. 湘语辰溆片语音研究[D]: [博士学位论文]. 西安: 陕西师范大学, 2008.
[27] 湖南省公安厅《湖南汉语方音字汇》编纂组. 湖南汉语方音字汇[M]. 长沙: 岳麓书社, 1993.
[28] 罗昕如. 湘方言词汇研究[M]. 长沙: 湖南师范大学出版社, 2006.
[29] 李永明. 衡阳方言[M]. 长沙: 湖南人民出版社, 1986.
[30] 杨时逢. 湖南方言调查报告[M]. 中央研究院历史语言研究所, 1974.
[31] 鲍厚星. 东安土话研究[M]. 长沙: 湖南教育出版社, 1998.
[32] 范俊军. 汉语方言自然口语语料库建设的几个基本问题[J]. 学术研究, 2013(2): 153-158.