Modern Linguistics 现代语言学, 2026, 14(4), 238-242
Published Online April 2026 in Hans. https://www.hanspub.org/journal/ml
https://doi.org/10.12677/ml.2026.144288

Hans汉斯

语料库视域下侗语资源数字化保护体系与实践路径

杨云霞

云南师范大学文学院，云南 昆明

收稿日期：2026年3月4日；录用日期：2026年3月27日；发布日期：2026年4月10日

摘　要

侗语属于汉藏语系壮侗语族侗水语支，是生活在侗族聚居区中的侗族人民所使用的语言。是侗族历史文化、思维方式的重要载体，是中华民族语言文化宝库的一部分。随着全球化、现代化的不断发展，侗语正面临使用人口减少、代际传承断裂、资源流失加剧等生存危机，因此数字化保护成为破解语言传承困境的关键路径。语料库具有语言资源数字化存储、管理与应用的功能，可对侗语进行抢救、精准维护与活态传承方面提供支撑。本文基于语料库与数字化保护理念，通过剖析侗语资源数字化现存问题，探索语料规范采集、精准加工、构建存储体系、多元化应用、健全传承机制的实践路径，为侗语及同类少数民族语言的数字化保护提供理论参考与实践借鉴。

关键词

语料库，侗语，数字化保护

Digital Preservation System and Implementation Pathways for Dong Language Resources from a Corpus Perspective

Yunxia Yang

School of Chinese Language and Literature, Yunnan Normal University, Kunming Yunnan

Received: March 4, 2026; accepted: March 27, 2026; published: April 10, 2026

Abstract

The Dong language belongs to the Dong-Shui branch of the Zhuang-Dong language family within the

文章引用: 杨云霞. 语料库视域下侗语资源数字化保护体系与实践路径[J]. 现代语言学, 2026, 14(4): 238-242.
DOI: 10.12677/ml.2026.144288

Sino-Tibetan language family. It is the language used by the Dong people living in Dong-inhabited areas. It is an important carrier of Dong ethnic history, culture, and way of thinking, and is a part of the language and cultural treasure house of the Chinese nation. With the continuous development of globalization and modernization, the Dong language is facing survival crises such as a decrease in the number of speakers, a break in intergenerational transmission, and an acceleration of resource loss. Therefore, digital protection has become a key path to solve the predicament of language transmission. A corpus has the functions of digital storage, management, and application of language resources, and can provide support for the rescue, precise maintenance, and dynamic inheritance of the Dong language. Based on the concept of corpus and digital protection, this paper analyzes the existing problems in the digitalization of Dong language resources, explores the practical paths of standardized collection, precise processing, construction of storage systems, diversified applications, and improvement of inheritance mechanisms, and provides theoretical references and practical references for the digital protection of the Dong language and similar minority languages.

Keywords

Language Database, Dong Language, Digital Protection

1. 侗语资源构建数字化保护体系的必要性

(一) 有利于整合资源

将数字化技术运用于侗语资源的保护与传承，可便于人们整合资源，提高获取各类文化资源的效率、增强对侗族语言文化的学习兴趣，为侗族大歌、侗戏等产业发展提供便利条件。

(二) 摆脱传承困境

现今侗语传承面临着严峻挑战。侗族没有民族文字，侗族大歌、谚语、民间传说等语言资源，均通过口耳相传的方式才得以延续。随着侗族本土文化传承者逐渐离世，珍贵的口头语言资源正不可逆转地永久流失，致使年轻一代能够接触、学习并传承的语言文化日益匮乏。而数字化保护体系可实现语料的永久留存，打破时空限制，破解“人亡语失”的困境。

(三) 赋能活态传承

数字化保护为侗语资源活态传承与可持续发展提供新路径和技术支持，全方位赋能侗语活态传承。通过开发多元应用场景并借助技术手段重塑侗语传承载体，使侗语资源从静态留存到动态使用变化、从“被动保存”到“主动治理”转变、提升侗语研究水平，扩大侗族语言文化影响力。

(四) 保护语言多样性

数字化通过语音采集、音视频录制、文本转写、语料标注、构建多模态语料库等技术手段，对濒危语言、少数民族语言中的语音系统、词汇语法、口头文学、民间叙事及文化知识体系进行全方位、系统化、标准化记录与永久存储，将原本脆弱易逝的口传语言资源转化为稳定可续的数字资产。这种数字化留存方式打破了传统保护手段易损耗、难保存、易失传的局限。在一定程度上能完整保留不同民族语言的独特结构、表达习惯与文化内涵。通过对各类民族语言、地方语言进行系统性数字存档，数字化保护最大限度维持了世界语言生态的丰富性与差异性，为每一种弱势语言留下了生存空间和发展可能，守护

了语言文化多样性的根脉，推动人类语言生态朝着多元共生、永续传承的方向发展。

2. 侗语资源数字化保护存在的问题

(一) 语料采集难

由于各地区侗语不同、采集缺乏统一的标准。加之采集过程中所使用的工具、标准、方式不同，易导致语料质量存在差异。如：研究者在采集语料的过程中大多倾向于选取知名度高的侗族大歌、戏曲等作为对象，对日常用语则缺乏关注。其次，首先侗语没有文字，大多借助汉字记侗音或用拉丁字母转写。例如，kəi[11] 常出现在贵州省榕江县车江乡侗族人的生活用语中，是因为 kəi[11] 在车江代表鸡蛋；kəi[11] 在贵州省榕江县七十二侗寨的语言系统中的字义为“睾丸”，所以人们很少提 kəi[11] 甚至将其划分为粗俗类语言。因此语义缺乏统一的规范标准，使得语言采集工作难以开展。最后，在整理语料的过程中会出现语料失真、乱码的情况，也会因为口语化语料太多导致语料加工需借助人工的力量才能完成，从而增加了语料采集、整理工作的难度。

(二) 数字技术适配不足

侗语属于低资源型的语言，当下的数字化保护技术多对汉语等主流语言进行研发。侗语语音系统较为复杂，有学者指出：侗语目前有 15 个声调，其中舒声调有 9 个，促声调有 6 个。侗语声调经历了声母清浊对立消失、元音长短对立消失和声母送气对立消失三次分化过程[1]。因侗语声调复杂的语音格局，增加了 ASR 模型的适配难度；目前侗语没有大规模的开放的标注语料库，而主流的 ASR 模型参数庞大[2]需以大量的标注性语料作为训练逻辑，它的算法架构与特征提取维度也都是以汉语等声调结构较为简单的主流语言作为研发对象，还没有对侗语的声调特性进行专项优化，即使 ASR 模型被应用于侗语语音识别的场景中，也会出现舒声调与促声调混淆、同类声调内部识别偏差的问题，无法进行精准自动识别，难以满足侗语语料数字化保护的精度化需求，易造成侗语资源与主流 ASR 模型研发之间的适配矛盾，进而阻碍后续专项核心技术的发展。

(三) 缺乏保障机制

侗族秉持万物有灵、敬畏自然的传统生态观与生命观，认为世间万物皆具灵性与生命，由此孕育出内涵深厚、兼具神圣性与私密性的民族语言体系。侗语资源中不仅承载着日常交流功能，还大量收录了民族禁忌、宗教祭祀、传统仪轨、口传史诗等具有专属文化属性与内部私密性的内容，属于侗族群体共有的精神文化财富。

当前，针对民族语言资源的产权归属界定、文化伦理审查及使用规范等相关保障机制仍显薄弱。在数字化采集、存储与传播过程中，若缺乏严格的伦理审核与前置脱敏处理，部分涉及民族核心信仰与私密仪式的敏感语料极易被不当公开、无序传播或擅自商用，不仅可能触犯民族文化禁忌，还会引发文化侵权、精神权益受损与群体隐私泄露等问题，既损害少数民族文化权益，也为侗语资源的数字化保护与可持续利用带来潜在风险。

3. 侗语资源数字化保护路径

(一) 规范采集

规范化的语料采集是筑牢侗语资源数字化的保护根基。因此需邀请相关领域的专家学者，根据各地区侗语的差异性，明确采集语料的工具参数、确定转写规则、制定语义界定及具体可行的语料采集标准，打破侗语语料采集的乱象，为后续全流程保护奠定基础；扩大语料采集范围，在对侗族大歌、戏曲采集进行采集的同时，还要注重对日常对话、民俗仪式用语等生活化语料的搜集力度，建立语料紧急采集机制，以便开展高龄传承人活态语料的抢救工作；使用专业录音、摄像设备，及时同步记录语料语境、发

音人信息，确保语料来源地真实性及采集地规范性。

(二) 精准加工

首先针对语料加工出现的效率低、质量差等问题，进行精细化、智能化加工，加强语料采集与存储的联通性，搭建专业化加工团队，采取人工校对与使用数字化工具的方式，解决语料失真、乱码问题；其次，优化加工流程，规范标注是语料标注的蓝图[3]。因此对声调、语义、句式内容等按照实际情况，采用偏误或“偏误标注 + 基础标注”的模式进行标注[4]；最后，建立加工质量审核机制，对加工处理后的语料进行检验、统一格式，确保语料的质量。

(三) 构建存储体系

鉴于数字资源存在保存难、风险高等情况，构建安全的数据存储体系十分重要。创建多模态资源、多功能的侗语语料存储平台。完善平台运行的安全机制，构建数据异地备份、多重加密体系，并定期对存储的数字资源进行检测与修复，保障数据安全，实行资源动态储存机制。此外，严格制定规范的存储权限，确保民族文化不被滥用，维护民族情感。

(四) 推动资源多元化应用

打破数字资源“重存档、轻活化”困境，紧扣文旅融合现实需求，推动侗语数字资源落地应用，实现保护与利用的良性循环。一是打造侗语学习数字化载体，开发“侗听 APP”等学习小程序，以便捷化、趣味化的方式助力外来旅客学习侗语，增强文旅体验感，进一步吸引旅客前往侗族地区游玩，实现语言传播与文旅发展的双向赋能。二是拓展非遗数字化传承场景，鼓励侗族非遗传承人依托各类多媒体平台开展线上教学，聚焦侗族琵琶歌、酒令歌等特色非遗文化，通过直播授课、短视频教学等形式，拓宽文化传承渠道，推动侗族文化活态传承与创新发展。三是研发文旅适配语音导览产品，整合侗语数字语料资源，结合侗族景区、非遗场馆等文旅场景，打造多语种、智能化语音导览服务，让游客在游览过程中沉浸式感受侗语文化魅力，推动数字资源与文旅场景深度融合，让数字资源真正绽放芳华。

(五) 健全传承机制

构建完备的数字化传承保护机制，有助于守住民族语言根脉。一是培养语言素养与数字技术兼备的专业人才，加强对侗族文化传承人、文化工作者的数字化技能培训，建立人才激励机制，引更多专业人才投身侗语数字化传承事业，筑牢人才根基。二是不断优化传承机制，健全相关的配套保障机制，实现侗语资源的长盛不衰与活态传承。立足侗语数字化发展的现实需求，建立动态的保护机制，确保语言传承工作平稳运行；三是加强产权与伦理保障，明确资源的产权归属，规避文化侵权与隐私泄露风险，为资源数字化保护体系创造安全的环境。

4. 结语

语言资源的数字化保护是顺应时代发展的必然趋势，数字技术为濒危语及低资源型语言的发展提供了新契机。本文以语料库作为研究视角，论述了对侗语进行数字化保护的意义，阐述了当下数字化保护存在的问题并根据不足提出了规范采集语料、精准加工、构建语料存储体系等保护路径。少数民族语言数字化保护并非简单的资源数字化存档，而是集众多学科技术于一体的系统性工程。未来需进一步优化语料库建设标准、强化产学研协同发力，推动语料资源从“被动保存”向“主动活化”转型，才能助力民族语言文化永续传承发展，维护语言生态的多样性筑牢根基。

基金项目

2025 年云南师范大学研究生科研创新基金一般项目“肇兴智能旅游汉侗双语多模态语料库建设”(项目编号：YJSJJ25-B46)。

参考文献

[1] 石林. 侗语声调的共时表现和历时演变[J]. 民族语文, 1991(5): 26-34.

[2] 时小虎, 袁宇平, 吕贵林, 等. 自动语音识别模型压缩算法综述[J]. 吉林大学学报(理学版), 2024, 62(1): 122-131.

[3] 张宝林. 汉语中介语语料库建设的现状与对策[J]. 语言文字应用, 2010(3): 129-138.

[4] 张宝林. 关于通用型汉语中介语语料库标注模式的再认识[J]. 世界汉语教学, 2013, 27(1): 128-140.