Modern Linguistics 现代语言学, 2025, 13(11), 9-13
Published Online November 2025 in Hans. https://www.hanspub.org/journal/ml
https://doi.org/10.12677/ml.2025.13111126

Hans汉斯

肇兴智能旅游汉侗双语多模态语料库建设

杨云霞

云南师范大学文学院，云南 昆明

收稿日期：2025年9月22日；录用日期：2025年10月21日；发布日期：2025年11月3日

摘 要

随着信息技术的飞速发展，智能旅游已逐步成为旅游业发展的重要趋势。在少数民族地区，智能旅游的发展不仅能够提升旅游服务质量，还能促进民族文化的传承与保护。肇兴侗寨是拥有1000多年历史的古老村寨，是全国最大的侗族聚居地，享有“侗乡第一寨”的美誉。其拥有十分丰富的侗族文化资源。建设智能旅游多模态语料库，对于推动肇兴智能旅游发展、传承侗族文化、促进民族交流具有重要意义。

关键词

肇兴侗寨，智能旅游，多模态语料库建设

Construction of a Multimodal Corpus for Intelligent Tourism in Both Chinese and Dong Languages in Zhaoxing

Yunxia Yang

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

Received: September 22, 2025; accepted: October 21, 2025; published: November 3, 2025

Abstract

With the rapid development of information technology, intelligent tourism has gradually become an important trend in the development of the tourism industry. In minority areas, the development of intelligent tourism can not only improve the quality of tourism services, but also promote the inheritance and protection of ethnic cultures. Zhaoxing Dong Village is an ancient village with a history of over 1000 years and is the largest settlement of the Dong people in China, enjoying the reputation of “the First Village of Dong”. It is rich in Dong cultural resources. The construction of a multimodal corpus for intelligent tourism is of great significance for promoting the development of intelligent tourism in

文章引用: 杨云霞. 肇兴智能旅游汉侗双语多模态语料库建设[J]. 现代语言学, 2025, 13(11): 9-13.
DOI: 10.12677/ml.2025.13111126

Zhaoxing, inheriting Dong culture, and facilitating ethnic exchanges.

Keywords

Zhaoxing Dong Village, Intelligent Tourism, Multimodal Corpus Construction

1. 引言

在全球数字化浪潮的席卷下，旅游业由传统模式朝智能化方向发展。将信息技术、大数据、人工智能等前沿科技融为一体的智能旅游，逐渐成为游客旅行的首选。此外，随着我国智能旅游市场规模的不断持续扩大，游客对智能导览、多语言服务、个性化旅游攻略等科技化服务的需求日益增长，智能旅游已然成为推动旅游业高质量发展的重要引擎。肇兴侗寨旅游景区位于贵州省黔东南苗族侗族自治州黎平县。自然风光独特、极具侗族民族风情。每年均有大量的国内外游客前往肇兴侗寨旅游、观光。但就目前而言，肇兴侗寨的汉侗双语或汉、英、侗语的语言服务体系还有待开发。既影响了游客的旅游体验感，又阻碍侗族文化的传播、传承。因此，智能旅游多模态语料库的创建就显得尤为重要。

2. 肇兴智能旅游汉侗双语多模态语料库的建设

本研究所构建的多模态语料库包含图像、文本、音频、视频这四大模态语料。本研究语料库具体建设如下。

2.1. 语料设计

笔者将结合肇兴地区的旅游发展现状及现有发展产业为依托，采集具有代表性、典型性、权威性、实用性，能突显肇兴侗寨作为贵州省南部侗族地区重要旅游景点的语料。

2.2. 语料采集

语料采集包括侗语语料收集、中文语料收集。语料来源路径由笔者通过实地田野调查所得、借助互联网和电子文献语料库得到。在电子文献语料库如贵州数字图书馆查找与肇兴侗寨旅游资源相关的语料过程中，发现其侗语和汉语的旅游语料较多。如杨祖华编著的《肇兴体验》[1]，其围绕肇兴及周边村寨，不仅展现侗族鲜活生动的原生态文化，还将侗寨绚丽秀美的自然风光尽数收录，堪称黎平县行走的旅游指南。同时也是采集肇兴汉侗旅游语料的重要来源。本研究的语料类型有：汉侗双语文本语料、音频、视频、图片等语料。汉侗双语文本语料主要是从黎平县旅游网官方平台下载及查阅描写肇兴侗寨历史文化的相关古籍；以实地田野调查的方式借助声飞田野调查软件，Sennheiser 录音话筒、YAMAHA 录音声卡等专业设备收集肇兴景区居民的日常交际用语、侗族歌谣等；借助贵州省旅游局官网、榕江县旅游局官网、旅游攻略网站等平台下载相关的视频和图片材料。视频大多为旅游宣传片、游客拍摄的旅游体验视频、民俗活动视频、导游录制的专业讲解视频等；图像数据则围绕肇兴当地的旅游景点、侗族建筑、服饰、美食展开拍摄。

2.3. 语料整理

首先，本研究对收集到的纸质文本语料进行扫描。利用 OCR 软件将 PDF 版本的文本语料转化为可

实际操作的 Word 文档。之后对语料中出现的错别字、多余空格、文字的格式、移除不合乎规范的注释、乱码等现象进行校对和降噪；其次，借助 FFmpeg 软件和云水印软件对图片和音视频语料进行格式转换和去水印。因侗族没有民族文字，侗语及侗族文化均靠侗族人民世代口口相传，所以田野调查得来的音频语料还需利用云龙国际音标软件对其进行转写。最后，为增强语料库建设的规范性，不仅所有的语料需要统一格式如视频转码为 MP4 格式、音频为 WAV 格式、图片为 JPG 格式，还需将它们储存为清晰、规范的电子文档。

2.4. 语料加工

本研究主要从语料标注和语料对齐这两个方面对收录整理的生语料进行加工处理。周燕[2]指出：标注是旨在为语料库中的文本内容增添描述性标签的过程。本研究将先使用 ELAN 软件工具对语料进行机器标注，之后再进行人工标注。中文语料则使用进行依据文本的语料信息对文本的词语性质进行词性标注，之后由人工进行校对。标注是语料加工过程中必不可少的步骤，常宝宝等[3]指出：加工标注的本质就是将语料中隐含的语言知识转化为明确可见的形式。

谢家成[4]指出：语料对齐为源语语料、译语语料分别放置于不同文本，并将两种类型的语料按段落或句子之间的关系进行对齐处理。由于汉语、侗语在语序、句子结构、语义表达等方面存在差异。陈锦娟[5]谈到：有时中文句子的原文翻译可能需多个英文句子才能准确表达，而多个中文句子的意思有时能浓缩为单个英文句子。因此，要实现逐句、翻译能完全对应十分困难。首先，本研究出现的汉语、侗语文本语料需统一字符和标点符号。其次，在借助专业的软件工具进行对齐的同时，还需投入一定量的人工对齐，在语料的加工过程中，人工可根据实际情况选择使用 Excel 或 WPS 等办公软件对其进行批量整理。

2.5. 语料的检索运用

熊婧[6]指出，语料库建成后，从语料库中获取信息并进行分析的过程称之为检索。检索类似于语料中的“搜索引擎”，它能让研究者高效地定位、提取并分析语言模式，从海量的数据中找到有价值的语言证据。本研究的语料库不仅包含汉、英、侗三种语言，还涵盖三者之间的语义翻译、句式对齐等关系。因此，可将三者存在的高频词汇、句式结构、不同交际场景的常用词汇等进行提取、观察、或者统计分析，为之后的文本研究奠定基础。而本研究所使用的检索工具为 CUC_ParaConc。它可用来检索双语、多语等语料，拥有跨编码检索能力，能够兼容 ANSI、Unicode、UTF-8 等主流纯文本格式。CUC_ParaConc 还具有排序、多语检索、自动检测并识别文本的对齐形式等功能。此外，还有正则式检索。正则表达式是语料库建设与检索中的核心技术，能够高效、灵活地解决复杂的模式匹配问题，能极大程度地提高语言工作者的检索效率。

2.6. 语料库的后期管理和维护

语料库建设是一个复杂而漫长的系统过程。即使建成之后，后期仍需投入大量的精力对其进行维护和完善，语料库才能继续运行。需定期清理无效、重复、无意义的数据，建立数据安全系统，防止外界对语料库进行恶意攻击，导致数据遭到破坏，时常更新语料数据、扩大语料库的数据规模，同时对语料库中进行更新过的语料版本进行标记，方便追踪查找。

3. 肇兴景区多模态语料库建设的意义

3.1. 促进当地旅游业发展

本语料库将不同类型的语料进行整合。例如：文本语料包含景区的景点简介、侗族文化、民族风情、

旅行攻略等；图像和视频资料有侗族吊脚楼、服饰、侗戏等；音频语料则包括侗族大歌、侗语版日常高频交际用语，侗语讲解等。这些语料不仅为肇兴景区智能旅行系统奠定了信息数据基础，还帮助全国各地的游客更加深入地了解侗族文化的韵味。此外，还可根据语料库的特点，研发 AI 智能旅游系统，借助语音识别、旅行数字人、语音合成技术，为前来肇兴游玩的旅客提供汉侗双语的语音讲解服务。游客可根据自身需求说出想要了解的景点或民俗文化，AI 导游对此进行相应的讲解。游客可使用 AI 智能旅游系统中的语音翻译功能，与当地侗族居民进行沟通交流，增强游客旅行的便利性和满意度，促进当地旅游业的发展。

3.2. 促进侗族文化保护与传承

随着现代化进程的不断发展，侗族人民没有专属于自己本民族的文化，诸多侗族文化仅以口头故事、歌谣的形式传承。因此，侗族文化面临着众多挑战。多模态语料库的建设在一定程度上可打破时空的限制，将侗族语言、歌谣、传统技艺等以数字化的形式储存下来。此外，语料库可作为教育资源，供年轻一代的侗族青少年学习。增强他们对本民族文化的认同感和自豪感，激发他们学习和传承侗族文化的积极性，促进侗族文化的发展。

3.3. 推动民族地区经济发展

首先，语料库的创建可为游客提供更加优质、便捷的服务，增强游客的旅行体验感，从而吸引全国各地更多的游客前往肇兴旅行、刺激消费，提高当地的旅游收入。其次，有助于肇兴当地的传统旅游业向智能化、数字化、科技化方向发展，促进其可持续发展。除此之外，旅游业的发展可带动餐饮、住宿、交通等产业的兴起，为当地居民提供更多的就业岗位，扩大劳动者的经济来源，促进少数民族地区的经济发展。

3.4. 增强各民族之间的交流与融合

语料库的建设，可打破外来游客与当地居民之间的语言壁垒。在实际的沟通交流过程中，不仅加深游客对侗族文化的了解，还有利于侗族人民学习吸收其他民族的优秀文化。不同民族之间的文化碰撞，增强了彼此之间优秀文化的多元发展，构建了和谐、友好、团结的民族关系。

4. 语料库建设过程中遇到的挑战及对策

多模态语料库打破了传统单模式的语料库。肇兴智能旅游汉侗双语多模态语料库是一项长期、复杂、涉及多个学科理论知识的综合性、系统性工程。语料库在建设过程中，主要遇到以下问题：语料收集困难。语料来源十分广泛、牵涉领域众多、不同类型的格式及语料质量存在差异。需要预先对语料的收集进行详细地策划，包括数据来源、收集类型、标准、技术路线等，确保后期整理工作能正常开展；语料标注复杂。因语料库为不同模态的语言且含有音视频语料和文本语料之间的跨模态标注，不同模态的语料标注存在难以兼容的现象。因此，本研究统一标注元素的定义及分类系统。首先与侗族的专家学者合作筛选出高频词语并进行标注，建立“双语对齐 + 场景适配”的标注元素定义规范。其次，构建旅游资源、自然景观、鼓楼，三级结构 + 适配机制。如：借用官方使用的标准名词对肇兴侗寨景区进行标注，侗族建筑类运用不同的几何图形进行标注等；缺乏技术支撑。多模态语料库的创建需要运用语料储存、分析、检索、语音、图像识别、语音合成等技术。但目前的技术若对多种模态语料进行一体化的管理、标注、融合及分析还存在一定难度，这也是我们今后需要克服的技术难点。

5. 结语

为解决外来游客到少数民族地区旅行双方语言不通的问题，本文提出建设多模态语料库。它是民族

文化发展和当代技术深入融合的一次创新尝试。建库之初是基于对侗族文化数字化的保护，逐步发展为兼容多模态语言的数字文化库。在很大程度上填补了侗族文化在智能旅游领域数据的空白，还探索出了一条少数民族地区“文化保护–旅游升级–人民收入多元–经济发展”的协同路径机制，成为保护民族文化根脉、推动区域旅游高质量发展、促进民族交流与交融的重要根基。虽然汉侗双语多模态语料库当前出现很多不足之处，但随着技术的进步、各界专家与学者的不断完善，相信它会成为维护侗族文化的“数字基因库”、推动肇兴侗寨旅游升级的“核心引擎”、促进民族交流融合的“桥梁纽带”，为我国少数民族文化的传承与发展贡献重要力量。

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] 杨祖华. 肇兴体验[M]. 海口: 海南出版社, 2008.

[2] 周燕. 海南景区多语翻译平行语料库的建设[J]. 品位·经典, 2023(2): 51-53.

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

[4] 谢家成. 小型英汉平行语料库的建立与运用[J]. 解放军外国语学院学报, 2004(3): 45-48.

[5] 陈锦娟. 衢州生态旅游汉英双语平行语料库构建研究[J]. 科技视界, 2016(24): 61-62.

[6] 熊婧. 江西省旅游翻译语料库的创建和应用[J]. 南昌师范学院学报, 2021, 42(5): 77-80.