www.sciengine.com/CSD
ISSN 2096-2223
CN 11-6035/N

文献 CSTR：
32001.14.11-6035.csd.2024.0208.zh
文献 DOI：
10.11922/11-6035.csd.2024.0208.zh
数据 DOI：
10.57760/sciencedb.18807

文献分类：计算机科学与技术

收稿日期：2024-12-21
录用日期：2025-09-11
发表日期：2025-09-17

TCST-UT：卫藏方言藏汉语音翻译数据集

黎鑫[1,2]，刘佳洛[1,2]，多杰朋毛[2]，看卓措[2]，戚肖克[3*]，赵小兵[1,2*]

1. 国家语言资源监测与研究少数民族语言中心，北京　100081
2. 中央民族大学信息工程学院，北京　100081
3. 中国政法大学法治信息管理学院，北京　102249

摘要：在大模型时代，多语种语言资源建设具有极为关键的意义。然而，目前公开的藏汉语音翻译数据集资源极为匮乏，这严重制约了藏语在多语种语言资源建设中的发展。为此，本研究充分参考国际语音翻译数据集规范，采用半自动标注方式构建了大规模卫藏方言藏汉语音翻译数据集。首先，基于公开的卫藏方言藏语自动语音识别数据集（M2ASR），利用 Gemini-1.5-pro 大模型将语音对应的藏语转录文本翻译成汉语。随后，专家对翻译结果进行严格审核与校正，最终整理成高质量的卫藏方言藏汉语音翻译数据集。本数据集包含 58,767 条藏语语音－藏语文本－汉语文本三元组，音频数据来自 147 个不同说话人，总时长为 72.08 小时，藏汉文本对数据文件大小为 22 MB。本数据集不仅为藏汉语音翻译研究提供了基础数据，同时也为其他低资源语言的语音翻译数据集构建提供了一定的经验。

关键词：藏汉语音翻译；数据集；半自动标注；低资源语言

数据库（集）基本信息简介

数据库（集）名称	TCST-UT：卫藏方言藏汉语音翻译数据集
数据作者	黎鑫，刘佳洛，多杰朋毛，看卓措，戚肖克，赵小兵
数据通信作者	戚肖克（qixiaoke@cupl.edu.cn），赵小兵（nmzxb_cn@163.com）
数据时间范围	2016—2023 年
地理区域	西藏
数据量	22 MB
数据格式	*.json
数据服务系统网址	https://doi.org/10.57760/sciencedb.18807
基金项目	国家社科基金重大项目（22&ZD035）
数据库（集）组成	数据集共包括 1 个数据文件，output.json 是 58,767 条藏语语音－藏汉文本对数据。

引　言

语音交流自古以来就是人类沟通的基本方式。随着全球化的推进，跨语言的语音交流需求日益增加，实现不同语言之间的无障碍语音交流成为广泛的社会需求[1]。语音翻译技术，尤其是语音到文本的翻译（AST Automatic Speech Translation）[2]，为跨语言沟通提供了重要支持。然而，现有的语音翻译系统主

* 论文通信作者
戚肖克：qixiaoke@cupl.edu.cn
赵小兵：nmzxb_cn@163.com

要面向高资源语言，如英语、汉语等。低资源语言的语音翻译数据稀缺，成为技术发展的主要瓶颈[3]。

目前，对于一些资源丰富的语言，如英语、汉语、法语、德语等，已经形成了较为成熟的语音翻译数据集构建方法。常见的方法主要有两种，一种是利用现有的语音识别（Automatic Speech Recognition）数据集，将源语言文本翻译成目标语言文本，从而生成语音翻译数据集。例如，BÉRARD A [4]以 LibriSpeech[5]公开英语语音识别数据集为基础，对英语文本利用谷歌翻译[6]进行法语对齐，生成了英法语音翻译数据集。这种方法的优势在于可以充分利用已有的大规模语音识别数据集，快速构建起语音翻译数据集，但其翻译质量可能受限于机器翻译系统的性能。另一种方法是利用机器翻译数据集，对源语言文字进行语音合成，生成语音翻译数据集。如 KANO T [7]通过英日机器翻译语料库，结合谷歌语音合成技术[8]生成语音语料库，用于端到端的英日语音翻译研究。此方法能够生成大量数据，但语音合成的自然度可能会影响数据集的质量。

然而，对于低资源语言，如藏语等，构建语音翻译数据集面临着更为严峻的挑战。目前，针对藏汉语音翻译的研究中，公开的语音翻译数据集极为稀缺，主要的已知资源是 TSCT（Tibetan-Chinese Speech Translation）藏汉语音翻译数据集[9]。该数据集基于公开的藏语语音识别数据集 M2ASR[10]的部分数据，借助机器翻译辅助采集，并补充了从微信公众号平台中爬取的藏汉数据，随后进行人工切分与标注，最终由专家审核和校正，共得到 7,270 条藏汉文本数据和约 8.8 小时的藏语语音数据。此外，还有研究利用 OCR（Optical Character Recognition）技术从带有中文字幕的藏语视频中提取语音和文本数据[11]，经过一系列处理后构建藏汉音译数据集。这些方法在一定程度上缓解了藏汉语音翻译数据集的匮乏问题，但仍存在一些不足之处。例如，数据规模有限、数据质量参差不齐、领域覆盖范围有限等。

为此，本研究利用大模型 Gemini-1.5-pro[12]，将公开的多语言少数民族语言自动语音识别（M2ASR）数据集中的藏语文本翻译成中文，并经过专家严格审核与校验，最终整理形成高质量的藏汉语音翻译数据集。数据集包含 58,767 条藏语语音－藏语文本－汉语文本三元组，总时长为 72.08 小时。本数据集的构建不仅为藏汉语音翻译研究提供了一定的数据基础，还推动了低资源藏汉语音翻译的研究发展。此外，这一数据集的高质量特性使其能够在自然语言处理的其他相关方向中得到应用，如藏汉机器翻译、藏语文本生成和语义理解等研究，为藏汉双语的研究和应用开辟了新的可能性。

1 数据采集和处理方法

为了构建藏汉语音翻译数据集，采用清华大学、西北民族大学、新疆大学联合发布的 M2ASR 卫藏方言藏语语音数据作为基础数据，通过半自动化标注方法构建藏汉文本对，最终构建包含藏语语音－藏语文本－汉语文本三元组的语音翻译数据集。数据处理的流程如图 1 所示，整个过程包括评估大模型翻译藏语的能力、清除非藏文文本、去重、机器翻译和数据清洗、专家审核、校验和反向匹配和规范化。

具体处理步骤如下：

（1）藏汉机器翻译模型评估与选择：随着大模型在机器翻译上的能力逐渐提高，本研究采用大模型进行藏汉翻译。首先，评估大模型在藏汉机器翻译任务上的能力。经调研，当下支持藏汉翻译的大模型有 Gemini-1.5-pro、Claude3.5、GLM4[13]、GPT-4o[14]。对比测评的模型中，Gemini-1.5-

pro 是基于 Transformer[15]的稀疏混合专家（SMoE）[16]模型，支持 100 万个 tokens 的上下文窗口，支持文本、图像、音频、视频等多种输入形式，具备卓越的多模态理解和推理能力；Claude3.5 是 Anthropic2024 年 6 月发布的基于 Transformer 模型，支持多模态输入，能够从不完美的图像中准确地转录文本；GLM4 是智谱 AI2024 年 6 月发布的基于 Transformer 模型，支持多语言和多模态；GPT-4o 是 OpenAI2024 年 5 月发布的基于 Transformer 模型，支持多模态输入。从 CCMT2021[17]藏文机器翻译文本中选取十句，对大模型进行 0-shot、1-shot、few-shot 翻译实验[18]，实验结果分别如表 1、表 2 和表 3 所示。经过比对，少样本上的翻译效果优于 0-shot 和 1-shot。另外，Gemini-1.5-pro 和 Claude3.5 在少样本上的翻译效果相当，但 Gemini-1.5-pro 免费，响应速度快且稳定，因此，本研究选择 Gemini-1.5-pro 大模型进行藏汉机器翻译。

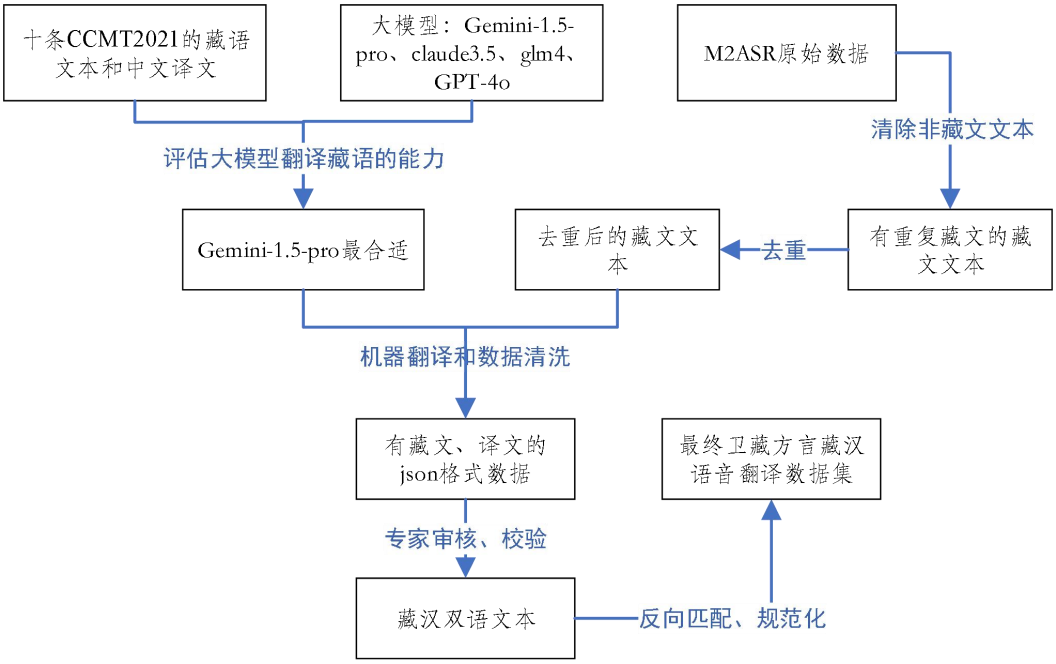

图 1　数据集处理流程

Figure 1　Data processing workflow

表 1　大模型 0-shot 翻译结果比对示例

Table 1　Example comparison of large model 0-shot translation results

例句1	“三农”、社会保障、医疗卫生、教育、文化等方面支出要提高。
glm4	“和平三宝”以及，国际和平组织，文化交流，教育，科研等领域的大规模合作。
GPT-4o	“三农”、社会保障、医疗卫生、教育、文化等方面支出要提高。
claude3.5	在“三农”、社会保障、医疗卫生、教育、文化等方面要加大投入。
Gemini-1.5-pro	在“三农”、社会保障、医疗卫生、教育、文化等方面要加大投入力度。
例句2	去年全国的财政收入比本年度预算超7239亿元。
glm4	在全球气候变化的背景下，从远古时代的文献中得知，已有7239年的气候变化记录。
GPT-4o	去年全国的财政收入比本年度预算超出7239亿元。
claude3.5	去年全国财政收入年度预算超过了72390亿元。
Gemini-1.5-pro	去年全国财政收入突破7239亿元，超出了年初预算。

表 2　大模型 1-shot 翻译结果比对示例

Table 2　Example comparison of large model 1-shot translation results

例句1	“三农”、社会保障、医疗卫生、教育、文化等方面支出要提高。
glm4	“和平三宝”以及，国际和平组织，文化交流，教育，科研等领域的大规模合作。
GPT-4o	“农田”与一般的管理方式、药物救助、教育、科研等需要重视的领域的结合，必须在适当的层面上进行广泛的传播和实施。
claude3.5	要在“三农”、社会保障、医疗卫生、教育、文化等领域加大投入。
Gemini-1.5-pro	要在“三农”、社会保障、医疗卫生、教育、文化等领域加大投入。
例句2	去年全国的财政收入比本年度预算超7239亿元。
glm4	在全球气候变化的背景下，从古代文献中得知，已有7239年的气候变化记录。
GPT-4o	在全国经济与社会发展的背景下，必须在实施中的某些地方加强监管，特别是在7239个被控制的地方
claude3.5	去年全国财政收入年度预算超过了7239亿元。
Gemini-1.5-pro	去年全国财政收入突破7239亿元，超出了年初预算目标。

表 3　大模型 few-shot 翻译结果比对示例

Table 3　Example comparison of large model few-shot translation results

例句1	“三农”、社会保障、医疗卫生、教育、文化等方面支出要提高。
glm4	“和平三宝”以及，国际和平组织，文化交流，教育，科研等领域的大规模合作。
GPT-4o	在农业、农村、农民（“三农”）、社会保障、医疗卫生、教育、文化等方面，必须加大投入力度。
claude3.5	要在“三农”、社会保障、医疗卫生、教育、文化等领域加大投入。
Gemini-1.5-pro	要加大对“三农”、社会保障、医疗卫生、教育、文化等方面的投入力度。
例句2	去年全国的财政收入比本年度预算超7239亿元。
glm4	在去年的财政收入中，比本年度预算超出了7239亿元。
GPT-4o	去年，全国财政收入比年度预算超出7239亿元。
claude3.5	去年全国的财政收入比本年度预算超7239亿元。
Gemini-1.5-pro	去年全国的财政收入比本年度预算超7239亿元。

（2）清除非藏文文本：M2ASR 数据集的藏文文本中仍含有少量非藏文数据，如“F_014_022　[illegible]”，因此先对源数据去除非藏文文本。

（3）去重：M2ASR 数据集的音频数据存在不同的人读相同的文本情况，对应的藏文文本也有很多重复的句子，因此，为了减少翻译藏文文本的工作量，本研究对藏文文本进行去重处理。由剩下的 58,767 条去重后剩余 21,949 条不重复的藏文文本。

（4）机器翻译和数据清洗：采用 Google Ai Studio 平台获取 Gemini-1.5-pro 进行少样本引导翻译。Google Ai Studio 是谷歌推出的一款基于浏览器的集成开发环境（IDE），主要用于生成式 AI 模型的原型设计与实验。少样本引导翻译是指采用（1）中校正后的藏文文本对大模型进行提示，

使得翻译效果更好。大模型的翻译结果中会有很多与数据集无关的信息，比如翻译无关的回答、空行、翻译文本最后的标点符号等，因此需要对生成的文本进行清洗。清洗后，利用编号将藏文文本和译文文本链接起来组成 json 格式，方便专家审核、校验。

（5）专家审核、校对：将整理好的 json 数据送给藏语专家进行审核和校对。在审核过程中，专家记录翻译中存在的问题，并进行校对，获取高质量的翻译结果。

（6）反向匹配、规范化：只清除非藏文文本的 M2ASR 的每个藏文文本对应唯一的编号，编号表明了部分音频文件路径，要给被去重的藏文文本匹配上对应正确的译文，本研究利用 python 的字典记录｛藏文：[编号集]｝的键值对，根据去重时的原理反向匹配译文。此时，得到 58,767 条藏文－汉文文本对，再通过编号信息得到在源数据集的路径信息，并且生成 58,767 条包含音频路径-藏文－汉文字典的 json 格式文件，最终形成卫藏方言藏汉语音翻译数据集（TCST-UT）。

2 数据样本描述

本研究构建的 TCST-UT 数据集包含 58,767 个样本，来自 147 个不同说话人（66 名男性和 81 名女性）共 72.08 小时卫藏方言的藏语语音数据，继承了 M2ASR 数据集在语音来源、说话人特征和文本类型上的多样性，涵盖了新闻、教育、文化、日常对话等多个领域。平均每个人的音频文件个数为 400 个，每个人总计平均音频时长达到 29 分钟，数据集则包含了 84 种藏文字符、2,470 种中文汉字。数据集中的每个样本都是一个三元组，包括藏语语音、对应的藏语文本及汉语文本。所有翻译文本均经过专家严格校对，确保标注质量。数据集内的藏汉对齐文本文件以结构化格式存储，本文公开发布，文件名为 output.json。藏语音频文件来自 M2ASR 卫藏方言语音识别数据集，发布在 m2asr.cslt.org 网站，可以通过电子邮件请求获取，所以本数据集不直接提供音频文件，只提供数据集中所包含的样本的音频路径与名称。每个样本的音频路径与对应藏文文本和汉文翻译文本存储在 output.json 文件中，其中音频路径指的是在公开数据集中的路径。本数据集为藏语等低资源语言的语音翻译研究提供了数据支持，适用于藏汉语音翻译、藏汉机器翻译、藏语语音识别等相关研究。

具体来说，TCST-UT 数据集中的藏汉对齐文本文件 output.json 的大小为 22 MB。文件将每条样本的音频路径、藏文文本和汉文翻译文本放入一个字典，数据格式为：

“说话人 ID_音频序号”: {
“audio”: 音频文件路径,
“text”: {
“Tibetan”: 音频文件对应的藏语文本
“Chinese”: 音频文件对应的汉语文本
}
}

一些示例如图 2 所示。

3 数据质量控制和评估

本数据集基于公开的卫藏方言藏语自动语音识别数据集（M2ASR），利用大模型 Gemini-1.5-pro 将语音对应的藏语转录文本翻译成中文，结果会有一定的偏差，所以采取了由藏语专家对翻译

结果审核、校验的方式控制数据质量。校验对比部分结果如表 4 所示。


```
"f98-La35_216": {
  "audio": "tibetan_v2_72h_share/F/f98-La35/f98-La35_216.wav",
  "text": {
    "Tibetan": "ཨངས་ཐུང་ཀུན་འགྱུར་ཐྱོས་ཤིའི་ཐྱོར་པྱན་གཡན་ཐྱི་འགྱུར་པ་བཐུན་ནས་ཐྱེ་བཙོས་རིགས་ཐྱི་རིམ་པ་ལོན་འཚེ་དགུ",
    "Chinese": "为什么自然变种会通过永久的变化而达到物种的水平"
  }
},
"f98-La35_217": {
  "audio": "tibetan_v2_72h_share/F/f98-La35/f98-La35_217.wav",
  "text": {
    "Tibetan": "ཡང་ཐོབ་ཐུབ་པ་གསོ་སྐྱོང་བང་བསྟེན་བཀུར་འཁྲུལ་ཞུ་རྒྱུ་ཡིན་དགོས་པ།",
    "Chinese": "也必须能够培养和尊重活佛转世"
  }
},
"f98-La35_218": {
  "audio": "tibetan_v2_72h_share/F/f98-La35/f98-La35_218.wav",
  "text": {
    "Tibetan": "སྐོལ་ཁྲིས་ཀོ་དམ་ལས་ཡུགས་བང་བཙའ་བའི་སྐོལ་ཁྲག་རྣམས།",
    "Chinese": "对于违反学校管理规定的学生"
  }
},
```


图 2　output.json 文件内容示例

Figure 2　Example of output.json file content

表 4　专家校正的部分比对示例

Table 4　Example of expert-corrected partial alignments

出现的问题	机器翻译	人工校正后
语义错误	山南地区佛教协会会长列协·索朗坚赞说：	山南地区佛教协会会长列协图美说
语义错误	向任正非家人表示崇高的敬意	向申展东家人表示崇高的敬意
语义错误	并出席了巴西、俄罗斯、印度、中国四国领导人第二次会晤	并出席了印度等四国领导人第二次会晤
漏翻	健全林业支持保护体系	健全林业支持和保护体系
漏翻	更要尽快恢复正常的社会秩序	九、要尽快恢复正常的社会秩序
语义错误	中尼建交50年来	中尼建交55年来
语义错误	奥巴马与内塔尼亚胡举行了会谈	奥巴马与内塔尼亚胡进行了会谈
语义错误	危地马拉缴获107公斤毒品	危地马拉缴获177公斤毒品

本研究还统计了与本数据集相关的说话人的音频文件数目和音频总有效时长，部分示例如图 3 所示。与本数据集对应的语音数据包含 58,767 个音频文件，音频来自 147 个不同的说话人（66 名男性和 81 名女性），时长总共为 72.08 小时，平均一个人说话时长为 29.42 分钟，平均每个人包含的音频文件数量约为四百个。

另外，本研究采用级联语音翻译系统对数据集进行基线评估实验，该系统由藏语语音识别模型和藏汉机器翻译模型级联组成，其中，语音识别采用了 Conformer 模型对数据进行训练，数据预处理包括：dither + specaug + speed perturb，参数配置信息为：lr 0.0005， warmup_steps 20000，batch size 8，3 gpu，30 epochs，解码信息为 average_num 20。利用 Genmini-1.5-pro 对语音识别的输出进行翻译，采用的是基于提示词的翻译方法，提示词是“请将以下藏文准确翻译为中文，务必确保编号完整，我所提供的编号与藏文句子一一对应，请勿对编号进行任何修改。在翻译时，每条句子的含义独立理解，同时要牢记我此前给出的正确翻译提示”，没有进行微调。具体过程如下：

按照约 8:1:1 的比例，将数据集根据说话人随机分为训练集、验证集和测试集，并在划分过程中确保同一说话人的数据分配到同一子集中。然后，采用训练集和验证集训练语音识别模型，模型使用 Conformer 架构，并在 WeNet 平台上进行实验[19]。训练完成后，将测试集数据输入模型，模型输出的语音识别结果通过 Gemini 大模型翻译为中文。将翻译结果与数据集内对应的中文文本比对，利用 nltk.translate.bleu_score 模块的 sentence_bleu 函数分别计算 BLEU-1、BLEU-2、BLEU-3、BLEU-4。级联语音翻译系统的平均 BLEU 值如表 5 所示。同时，部分句子的单句 BLEU 值及不同 BLEU 值的分布情况分别展示在图 4 和图 5 中。从结果来看，模型在基础词汇翻译方面表现较为精准，BLEU-1 的分数达到了 43.83%，在处理词语组合方面也有比较好的表现，这也反映了本数据集质量比较高。然而，基线模型在面对长句及复杂语义结构时翻译能力有所下降。未来可以利用本数据集深入研究藏汉语音翻译算法，以提升模型性能。

L_F_0_01 - 音频文件个数: 404, 总时长: 22.95 分钟
L_F_0_06 - 音频文件个数: 410, 总时长: 24.60 分钟
L_F_0_09 - 音频文件个数: 409, 总时长: 34.65 分钟
L_M_0_06 - 音频文件个数: 402, 总时长: 22.34 分钟
L_M_0_08 - 音频文件个数: 404, 总时长: 24.34 分钟
L_M_0_12 - 音频文件个数: 409, 总时长: 23.66 分钟
L_M_0_15 - 音频文件个数: 409, 总时长: 26.58 分钟
L_M_0_19 - 音频文件个数: 407, 总时长: 22.09 分钟
L_F_0_13 - 音频文件个数: 410, 总时长: 28.50 分钟
L_M_0_01 - 音频文件个数: 389, 总时长: 24.07 分钟
L_M_0_16 - 音频文件个数: 408, 总时长: 23.78 分钟
L_F_0_05 - 音频文件个数: 410, 总时长: 31.58 分钟
L_F_0_12 - 音频文件个数: 409, 总时长: 26.86 分钟

图 3　说话人的文件数目和音频时长示例

Figure 3　Example of the number of files and audio duration of speakers

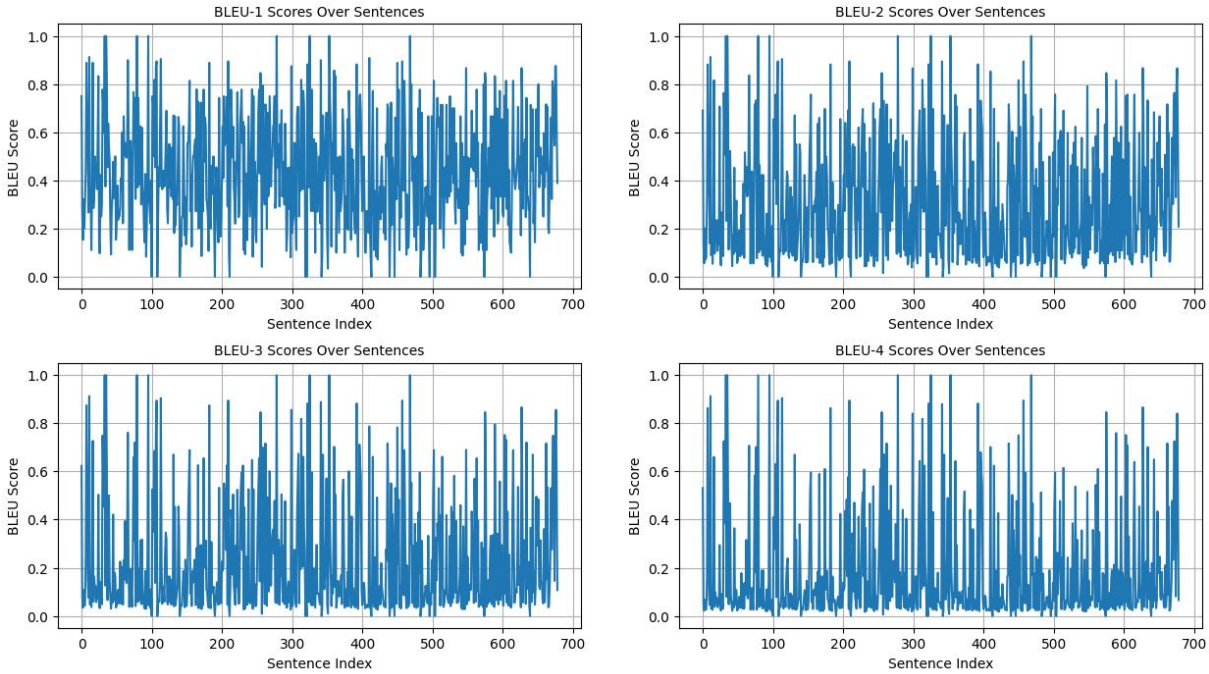

图 4　BLEU 的实验结果

Figure 4　BLEU experimental results

表 5　平均 BLEU 结果

Table 5　Average BLEU results

BLEU-1	43.83
BLEU-2	30.55
BLEU-3	23.21
BLEU-4	18.15

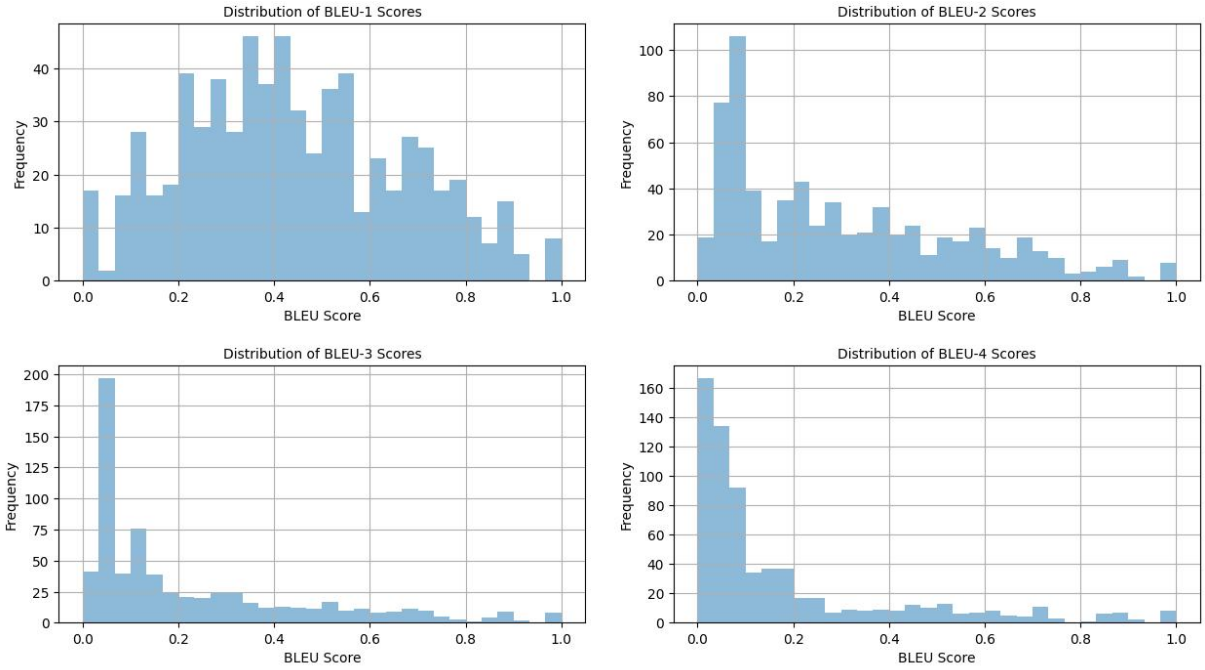

图 5　BLEU 分数的分布图

Figure 5　Distribution of BLEU scores

为了对人工校正的效率进行量化评估，从原数据的藏文文本中随机抽取了 300 条样本。针对这批样本，比较了两种翻译策略的时间成本：（1）直接进行人工翻译的全人工标注方式；（2）先经过 Gemini-1.5-pro 生成初步译文，再进行人工校对的半自动标注方式。通过记录两种方式所需的时间，计算时间节省百分比。结果显示，人工翻译 300 条文本所需时间约为 240 分钟，而经过 Gemini-1.5-pro 翻译之后再进行人工校对的总耗时约为 60 分钟，即半自动标注方法使得时间效率提升了约 75%。这体现了本研究通过半自动标注方法能够更高效地构建本研究的数据集。

最后，还对 Gemini-1.5-pro 生成的翻译进行了误差分析。具体来说，选择 600 条经 Gemini-1.5-pro 生成的藏文翻译结果，请藏语母语者进行了人工校对，同时记录了错误类型及数量。通过统计分析，错误类型大致可以分为四类，最终统计结果为：语义错误 27 条，漏翻 7 条，顺序错误 3 条，多翻 2 条。Gemini-1.5-pro 生成翻译的误译类型统计柱状图如图 6 所示。

4　数据价值

目前，公开的藏汉语音翻译数据集资源较为匮乏，这在很大程度上阻碍了大模型时代低资源藏语智能信息处理领域产学研的发展与应用[20]。本数据集可广泛应用于机器翻译的研究，为低资源藏语的自然语言处理领域研究提供了数据基础，有助于推动藏汉语言交流以及少数民族语言智能信息

处理领域的不断发展[21]。

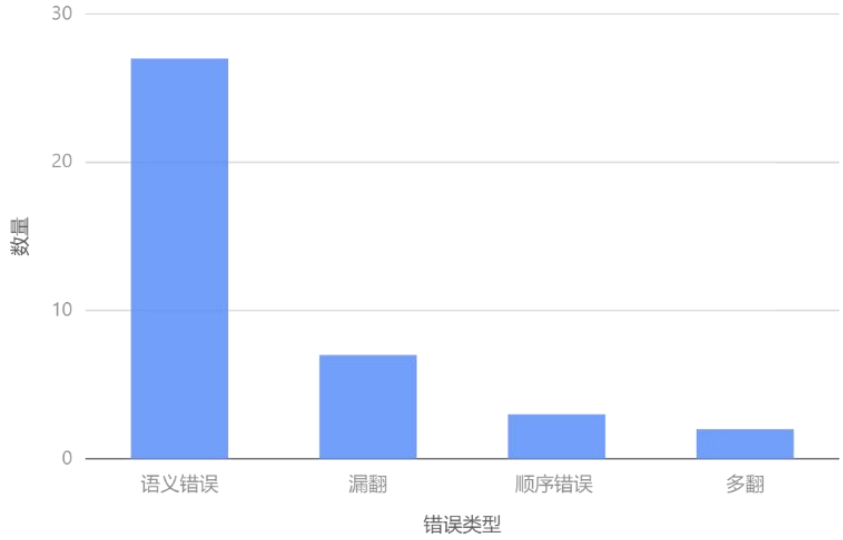

图 6　大模型翻译错误类型统计图

Figure 6　Statistical chart of translation error types by Large Language Models

藏汉数据为本藏汉语音翻译数据集的数据整理加工过程具有独特性。采用大模型进行翻译结合专家审核校验的加工方法，既充分发挥了人工智能的高效性，又保证了翻译的准确性和专业性[22]。这种先进的加工方法为藏汉语音翻译数据集的质量提供了有力保障。

此外，本数据集不仅可用于藏汉语音翻译研究，还可应用于自然语言处理的其他相关方向，如藏汉机器翻译、藏语文本生成、语义理解等。

数据作者分工职责

黎鑫（2002—），女，湖北省咸宁市人，硕士研究生，研究方向为语音识别和语音翻译。主要承担工作：数据集的生成与整合、论文撰写。

刘佳洛（2001—），男，江西省赣州市人，硕士研究生，研究方向为语音识别和语音翻译。主要承担工作：数据质量控制。

多杰朋毛（2003—），女，青海省海南州贵德县人，本科生，专业为中国少数民族语言文学。主要承担工作：数据的审核与校对。

看卓措（2005—），女，青海省黄南州同仁市人，本科生，专业为计算机科学与技术。主要承担工作：数据清洗、审核与校对。

戚肖克（1985—），女，山东省菏泽市人，博士，副教授，研究方向为语音信号处理、自然语言处理。主要承担工作：数据整理和论文写作指导，数据质量控制。

赵小兵（1967—），女，内蒙古自治区呼和浩特市人，博士，教授，研究方向为自然语言处理。主要承担工作：数据质量控制与综合管理。

参考文献

[1] HJARVARD S. The globalization of language[J]. Nordicom Review, 2004, 25(1-2): 75-97.

[2] KUREMATSU A, MORIMOTO T. Automatic Speech Translation[M]. CRC Press, 2023.

[3] MAGUERESSE A, CARLES V, HEETDERKS E. Low-resource languages: a review of past work and future challenges[EB/OL]. 2020: arXiv: 2006.07264. https://arxiv.org/abs/2006.07264

[4] BÉRARD A, BESACIER L, KOCABIYIKOGLU A C, et al. End-to-end automatic speech translation of audiobooks[C]//2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Calgary, AB, Canada. IEEE, 2018: 6224–6228. DOI:10.1109/ICASSP.2018.8461690.

[5] PANAYOTOV V, CHEN G G, POVEY D, et al. Librispeech: an ASR corpus based on public domain audio books[C]//2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). South Brisbane, QLD, Australia. IEEE, 2015: 5206 – 5210. DOI: 10.1109/ICASSP.2015.7178964.

[6] WU Y H, SCHUSTER M, CHEN Z F, et al. Google’s neural machine translation system: bridging the gap between human and machine translation[EB/OL]. 2016: arXiv: 1609.08144. https://arxiv.org/abs/1609.08144.

[7] KANO T, SAKTI S, NAKAMURA S. End-to-end speech translation with transcoding by multi-task learning for distant language pairs[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2020, 28: 1342–1355. DOI:10.1109/TASLP.2020.2986886.

[8] WANG Y X, SKERRY-RYAN R, STANTON D, et al. Tacotron: towards end-to-end speech synthesis[EB/OL]. 2017: arXiv: 1703.10135. https://arxiv.org/abs/1703.10135

[9] 赵小兵, 刘佳洛, 周毛克, 等. 藏汉语音翻译数据集[J/OL]. 中国科学数据, 2024, 9(4): 21–29. DOI:10.11922/11-6035.csd.2024.0023.zh. [ZHAO X B, LIU J L, ZHOU M K, et al. A dataset of Tibetan-Chinese speech translation[J/OL]. China Scientific Data, 2024, 9(4): 21 – 29. DOI:10.11922/11-6035.csd.2024.0023.zh.]

[10] WANG D, ZHENG T F, TANG Z Y, et al. M2ASR: Ambitions and first year progress[C]//2017 20th Conference of the Oriental Chapter of the International Coordinating Committee on Speech Databases and Speech I/O Systems and Assessment (O-COCOSDA). Seoul, Korea. IEEE, 2018: 1 – 6. DOI:10.1109/ICSDA.2017.8384469.

[11] 徐晓娜, 谭晶, 赵悦. 基于OCR技术辅助构建藏汉音译数据集的方法及系统: CN116468054A[P]. 2023-07-21. [XU X N, TAN J, ZHAO Y. Method and system for auxiliary construction of Tibetan-Chinese transliteration data set based on OCR technology: CN116468054A[P]. 2023-07-21.]

[12] TEAM G, GEORGIEV P, LEI V I, et al. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context[EB/OL]. 2024: arXiv: 2403.05530. https://arxiv.org/abs/2403.05530

[13] GLM T,: , ZENG A H, et al. ChatGLM: a family of large language models from GLM-130B to GLM-4 all tools[EB/OL]. 2024: arXiv: 2406.12793. https://arxiv.org/abs/2406.12793

[14] ISLAM R, MOUSHI O M. GPT-4o: the cutting-edge advancement inMultimodal LLM[C]//Intelligent Computing. Cham: Springer, 2025: 47–60. DOI:10.1007/978-3-031-92611-2_4.

[15] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[J]. Advances in neural information processing systems, 2017, 30.

[16] FEDUS W, ZOPH B, SHAZEER N. Switch transformers: scaling to trillion parameter models with simple and efficient sparsity[J]. Journal of Machine Learning Research, 2022, 23(1): 5232 – 5270.

DOI:10.5555/3586589.3586709.

[17] 第十七届全国机器翻译大会(CCMT2021)在西宁召开[J]. 中文信息学报, 2021, 35(10): 136. DOI:10.3969/j.issn.1003-0077.2021.10.016. [The 17th national machine translation conference (CCMT2021) was held in Xining[J]. Journal of Chinese Information Processing, 2021, 35(10): 136. DOI:10.3969/j.issn.1003-0077.2021.10.016.]

[18] BROWN T B, MANN B, RYDER N, et al. Language models are few-shot learners[C]//Proceedings of the 34th International Conference on Neural Information Processing Systems. Vancouver, BC, Canada. ACM, 2020: 1877–1901. DOI:10.5555/3495724.3495883.

[19] ZHANG B B, WU D, PENG Z D, et al. WeNet 2.0: more productive end-to-end speech recognition toolkit[C]//Interspeech 2022. ISCA, 2022: 1661-1665.. DOI:10.21437/interspeech.2022-483.

[20] ARIVAZHAGAN N, BAPNA A, FIRAT O, et al. Massively multilingual neural machine translation in the wild: findings and challenges[EB/OL]. 2019: arXiv: 1907.05019. https://arxiv.org/abs/1907.05019.

[21] NEUBIG G, HU J J. Rapid adaptation of neural machine translation to new languages[EB/OL]. 2018: arXiv: 1808.04189. https://arxiv.org/abs/1808.04189.

[22] KOCMI T, BOJAR O. Trivial transfer learning for low-resource neural machine translation[EB/OL]. 2018: arXiv: 1809.00357. https://arxiv.org/abs/1809.00357.

论文引用格式

黎鑫, 刘佳洛, 多杰朋毛, 等. TCST-UT:卫藏方言藏汉语音翻译数据集[J/OL]. 中国科学数据, 2025, 10(3). (2025-09-17). DOI: 10.11922/11-6035.csd.2024.0208.zh.

数据引用格式

黎鑫, 刘佳洛, 多杰朋毛, 等. TCST-UT:卫藏方言藏汉语音翻译数据集[DS/OL]. V3. Science Data Bank, 2025. (2025-08-07). DOI:10.57760/sciencedb.18807.

TCST-UT: a dataset of Tibetan-Chinese speech Translation in the Ü-Tsang dialect

LI Xin[1,2], LIU Jialuo[1,2], DUO Jiepengmao[2], KAN Zhuocuo[2], QI Xiaoke[3*], ZHAO Xiaobing[1,2*]

1. National Language Resource Monitoring & Research Center of Minority Languages, Beijing 100081, P. R. China

2. School of Information Engineering, Minzu University of China, Beijing 100081, P. R. China

3. School of Information Management for Law, China University of Political Science and Law, Beijing 102249, P. R. China

Abstract: In the era of large language models, the construction of multilingual language resources is of great significance. However, the publicly available Tibetan-Chinese speech translation datasets are

currently very limited, which significantly hinders the development of Tibetan within the context of multilingual language resource construction. To address this, this paper presents a large-scale dataset of Tibetan-Chinese speech translation constructed using a semi-automatic annotation approach, adhering to international speech translation dataset standards. Initially, leveraging the publicly available Tibetan automatic speech recognition dataset for the Ü-Tsang dialect (M2ASR), we employed the Gemini-1.5-pro large language model to translate the corresponding Tibetan transcripts into Chinese. These translations underwent rigorous review and correction by experts, resulting in a high-quality dataset of Ü-Tsang dialect Tibetan-Chinese speech translation. This dataset comprises 58,767 speech-text pairs, totaling 72.08 hours of audio from 147 different speakers, with a total size of 22 MB. This dataset can provide foundational data for Tibetan-Chinese speech translation research. Furthermore, it can offer insights for the construction of speech translation datasets for other low-resource languages.

Keywords: Tibetan-Chinese Speech translation; dataset; semi-automatic annotation; low-resource Languages

Dataset profile

Title	TCST-UT: a dataset of Tibetan-Chinese speech Translation in the Ü-Tsang dialect
Data authors	LI Xin, Liu Jialuo,Duo Jiepengmao,Kan Zhuocuo,Qi Xiaoke,Zhao Xiaobing
Data corresponding author	QI Xiaoke (qixiaoke@cupl.edu.cn), ZHAO XiaoBing (nmzxb_cn@163.com)
Time range	2016-2023
Geographical scope	Xizang
Data Volume	22MB
Data formate	*.json
Data service system	https://doi.org/10.57760/sciencedb.18807
Source of funding	National Social Science Foundation of China (22&ZD035)
Dataset composition	The dataset consists of one data files—output.json, which contains 58,767 Tibetan speech-Tibetan-Chinese text pairs.