Behind-the-Scene Challenges in Sino-Tibetan Low-Resource Language Technology Development

Anonymous ACL submission

Abstract

Despite the persistent effort in developing low-resource language technologies (LRLTs) in NLP, little attention has been paid to the experience of the LRLT practitioners to understand their challenges and opportunities. We employ a qualitative approach and conduct 15 semi-structured interviews with LRLT practitioners working on three Sino-Tibetan low-resource languages. Through our analysis, we identify not only the technical challenges such as the lack of textual data and basic tools, but also the linguistic, social and financial challenges, causing and exacerbating the technical challenges. Fortunately, we also find that most participants are optimistic towards the LRLT development and mitigate the challenges by fostering communities, seeking support from various communities and leveraging the power of large language models. We discuss the complex nature of the challenges in relation to prior literature, and provide recommendations for practitioners, governments and the broader NLP community.

1 Introduction

The development of low-resource language technologies (LRLTs) has become one of the main research topics in the field of natural language processing (NLP). From cross-lingual transfer learning (Conneau et al., 2018; Ruder et al., 2019) to data augmentation (Fadaee et al., 2017; Liu, 2022), researchers have proposed various technical approaches to extend the development of language technologies (LTs, including any digital artifacts for languages from input methods, fonts and language corpora to language models) under low-resource settings.

Nevertheless, the challenges surrounding the LRLT development are not limited to the algorithmic and technical aspects. While some challenges and suggestions can be deduced from analyzing LRLT artifacts created by various communities (Yu et al., 2022), challenges related to the a priori factors for a successful development of LRLTs often go undocumented in the published NLP literature. Revealing these challenges can hopefully bring the attention of those who are or want to support LRLT development to the right place and alleviate the burden for LRLT practitioners by providing what they actually need.

We therefore turn our attention to the qualitative research, specifically through interviewing these practitioners, which provides an opportunity for us to learn about the "behind-the-scene" challenges. We recruited LT practitioners working on Sino-Tibetan low-resource languages (LRLs), and conducted semi-structured interviews, asking participants questions related to a set of broad topics (Knott et al., 2022), with 15 participants working on Cantonese, Southern Min and Tibetan, for having the most resources (Joshi et al., 2020) and publications (Liu and Best, 2025) among all Sino-Tibetan LRLs. We asked these practitioners about their experience in developing LRLTs for the language(s) that they had worked on. Specifically, we are interested in the following research questions:

RQ1. What are the different challenges faced by the Sino-Tibetan LRLT practitioners?

RQ2. What do practitioners identify as ways for them to mitigate these challenges?

Through our analysis, we find that aside from technical challenges, LRLT practitioners also need to overcome linguistic challenges of the language itself when developing the technologies, as well as the lack of financial and resource support due to the nature of LRLT development. These challenges are not standalone but often entangled with each other such that they collectively aggravate the already restrictive low-resource settings. It is also surprising to see how challenges differ based on the language that each practitioner works on,

despite all under the similar low-resource setting. However, LRLT practitioners have attempted to mitigate these challenges somewhat successfully through the support from the communities of LRLT development and related areas, and more recently, they leverage the advancement of large language models (LLMs) to directly help them address some critical issues in LRLT development. We thus discuss on what the broader NLP community can learn from these challenges and opportunities, and our recommendations on how LRLT practitioners can better overcome these challenges. The contribution of this paper can be summarized as:

1. We identify the challenges faced by Sino-Tibetan LRLT practitioners. The challenges may also be applicable to NLP practitioners working on other low-resource settings.

2. We identify the ways through which these practitioners are able to mitigate the challenges, and we suggest recommendations to how these challenges can be overcome for the advancement of LRLT development.

Through conducting interviews and analyzing the topics emerged from them, we reveal deeper insights about the LRLT communities. We hope this work inspires more qualitative and participatory research in NLP, advancing the LTs through a human-centered approach.

2 Related Work

2.1 Development of LRLTs

The LRLT development has become an important subfield of NLP, with its dedicated track in major conferences (e.g. "Low-Resource Methods for NLP" in ACL and EMNLP), workshops (e.g. LoResLM[1]), conferences (e.g. LaTeLL[2]) and journals (e.g. TALLIP[3]). Specifically for the three languages in our study, i.e. Cantonese, Tibetan and Southern Min, most research on LTs centers around topics such as corpus creation (Meelen et al., 2021), machine translation (Hong et al., 2024; Zhou et al., 2024), speech recognition (Yang et al., 2023; Li et al., 2025). More recently, there has been more research on the support of these languages by LLMs, including creating benchmarks for language models (Cheng et al., 2025), or even training LLMs on these languages directly (Lai et al., 2024; Pan et al., 2025; Sha et al., 2025).

2.2 Participatory Research in NLP

Participatory research (Cornwall and Jewkes, 1995; Vaughn and Jacquez, 2020), i.e. collaborative research between researchers and community members, aims to expand our understanding on a research topic through the aspect of a community. This type of research method has been increasingly employed by NLP researchers in recent years to understand the interaction of LTs and various communities. For example, Al Kautsar et al. (2025) used a nationwide survey to understand the different needs of LTs by demographic groups and their reasoning. It also included one interview to provide insights from the government's point of view. Harvey et al. (2025) conducted semi-structured interviews to investigate the challenges faced by practitioners when using publicly available NLP tools to measure representational harms. Siminyu et al. (2023) also conducted interviews with different stakeholders of African LT development, such as content creators and language technologists, to investigate their motivations and challenges as well as provide recommendations for these stakeholders. Nekoto et al. (2020) advocated for a participatory approach to sustainably develop machine translation systems. These studies enrich the field of NLP by expanding our understanding of technological development not just on the technical level, but also how such development influences and is influenced by different communities.

Nevertheless, participatory research still remains a niche area in NLP for various communities including not only the LRLT communities but also the broader NLP community. In this paper, we use semi-structured interviews and qualitatively analyze the interview data to understand the various types of challenges and opportunities faced by LRLT practitioners from their experience.

3 Method

We conducted 15 semi-structured interviews, i.e. interviews with a topic guide instead of a closed set of questions, with researchers and software developers who have worked on the development of LTs for Cantonese, Southern Min, and Tibetan after approval from the Institutional Review Board (IRB) at the authors' institution. We detail the procedure for recruitment, interview and analysis below.

[1] https://loreslm.github.io
[2] https://latell.org/2026/
[3] https://dl.acm.org/journal/TALLIP

PID	Language	Base	Years
P1	Cantonese	East Asia	10+
P2	Cantonese	East Asia	10+
P3	Cantonese	East Asia	5-10
P4	Cantonese	North America	5-10
P5	Cantonese	East Asia	3-5
P6	Southern Min	East Asia	10+
P7	Southern Min	East Asia	10+
P8	Southern Min	East Asia	10+
P9	Southern Min	East Asia	10+
P10	Southern Min	East Asia	10+
P11	Southern Min	East Asia	5-10
P12	Tibetan	Europe	10+
P13	Tibetan	Europe	10+
P14	Tibetan	East Asia	5-10
P15	Tibetan	North America	3-5

Table 1: Participants' information showing their assigned participant IDs (**PID**), the Sino-Tibetan LRL that they are working on (**Language**), where they are currently working (**Base**), and the number of years that they have worked in LT development (**Years**).

3.1 Recruitment

We followed the purposive sampling (Etikan et al., 2016) by sending email invitations to recruit participants who have worked in the technologies for at least one of the three targeted languages and have at least three years of experience working in LTs for developing products for LRLs and publishing academic papers about LRLTs. The participants' email addresses were found on their publications (if any) and through Internet searches. In addition, we also employed snowball sampling (Parker et al., 2019) by asking interview participants to connect us to other potential participants. We stopped recruiting participants when we began to notice repetitive names of potential participants and themes (Czarniawska, 2014). In the end, we recruited 15 participants to conduct the interviews. The demographics of the participants are listed in Table 1.

3.2 Interview Procedure

The first author conducted semi-structured interviews with 8 participants in-person and 7 over Zoom. Interviews were conducted in either Mandarin or English, based on the participant's preference. Most interviews lasted from 60-80 minutes and were audio recorded after obtaining the participants' consent. Each interview was divided into three major sections. First we asked about the participant's background in developing LTs and low-resource LTs. Next, we proceeded with questions about the current development of technologies for the target Sino-Tibetan language, including the main research foci and challenges. Lastly, we ended the interview by asking about their opinions on the future LRLT development.

3.3 Data Analysis

To ensure confidentiality, we used a modified version of the OpenAI Whisper model (Radford et al., 2023) on GitHub[4] that could be downloaded and run locally to automatically generate transcriptions for the recorded interviews. We used the `medium` model size with all other hyperparameters set to default. The generated transcripts were examined and coded by the first author using a computer-assisted qualitative data analysis software (CAQDAS) QualCoder[5] for data analysis. With our research questions borne in mind, we conducted an inductive thematic analysis, drawing on grounded theory (Glaser and Strauss, 1967) coding techniques including descriptive coding, i.e. annotating conversation segments by summarizing with a short phrase (Saldaña, 2009), and in vivo coding, i.e. annotating conversation segments using the actual language of the conversation (Glaser and Strauss, 1967), to code everything related to the research questions. Finally, these codes were reorganized and grouped into different categories to answer our research questions.

4 Findings

4.1 Three Layers of Challenges

The core challenge for developing technologies for low-resource languages lies in the term "low-resource" itself, i.e. there is a lack of data (more specifically, textual data, in the current paradigm of NLP) for these languages. Yu et al. (2022) found that most input text (unlabeled text) is collected from the web. However, such data collection method is limited in low-resource context. P6 gave specific figures to illustrate the struggle in Southern Min data collection through web scraping:

> *"The text that we can get from the Internet is only about 10 million characters, and that is because there are some writers donating their work. Otherwise what*

[4] https://github.com/SYSTRAN/faster-whisper
[5] https://github.com/ccbogel/QualCoder

we could have scraped was only about 5 million characters."

10 million Chinese characters take up a storage size around 20-30 MB, which is even smaller than the classic SQuAD dataset (Rajpurkar et al., 2016), the famous English Question-Answering dataset that was annotated and released roughly a decade ago. The large pre-training corpora these days for LLMs, such as C4 (Raffel et al., 2020) which was used to train many LLMs including LLaMA (Touvron et al., 2023), are on the order of terabytes.

While data scarcity is the main challenge for the development of LRLTs, our analysis found challenges with the linguistic difficulties of the languages themselves, the limited resource and financial support and the need for developing basic tools – all of which seem to contribute to the low-resource-ness of these languages.

4.1.1 Intrinsic Linguistic Difficulties

Textual data can only be collected when people' thoughts are written down. This is not trivial as pointed out by P5 trying to extend their work to Shanghainese (a dialect of the Wu Chinese language spoken mainly in Shanghai): *"If you have 10 million speaker, you ought to have a standardized orthography, or standardized romanization, and the romanization need to come with a character dictionary, and a compatible input method, and of course, textbooks. I just assume these are going to be taken for granted, but that is not the case."* While all three languages that we study here can be written down in some prescribed orthographies, we found that each of the three languages tells a different reason why this might be proven difficult for low-resource languages.

1. Written Cantonese *"lacks standardization"* (P1). Even when the government has set a standard, *"We might not type according to that standard [orthography published by the authority]"* (P2), i.e. how people type Cantonese often deviates from an existing standard with homophones and even Latin alphabet.

2. Southern Min can be written in *"Chinese characters and two forms of romanizations"* (P10), but most researchers (and speakers) *"cannot read or write"* (P7) the language even when they speak and understand it.

3. While there is a single standard orthography used to write Tibetan, *"We have very, very little text of people writing in their mother tongue"* (P13) and *"even newspapers now in Tibetan sound a lot like classical Buddhist texts, much more than they sound like how people speak on the street"* (P12).

Resource-rich languages such as English have standardized writings adopted by the government and used by their native speakers often to reflect how they use the language in everyday life. While there might be dialectal and individual differences, spellings remain relatively homogeneous for each dialect. However, for LRLs with non-standardized writing or no writings at all, or if the standards are not used by the native speakers, data scarcity will occur as there is little data to be collected.

But what about spoken data – if we only collect spoken data instead of fixating on writing them down in some prescribed orthography, would this be an easier alternative? As it turns out, this form of data collection could be equally challenging for LRLs, if not more. Since a lot of the LRLs are also marginalized languages where they are spoken, many do not have a standardized form of their speeches, and speakers often speak dialects of these languages. P6, P7 and P11 all pointed out how Southern Min spoken in northern versus southern Taiwan carries different accents. Many ASR models would only recognize one dialect of the language, struggling with other dialects (P11) or foreign accents (P9). In some cases, what people often understand as dialects can be mutually unintelligible, such as Amdo Tibetan and Lhasa Tibetan (P12; Tournadre, 2014), so that developing LTs for only one "dialect" of the language will not be usable by the speaker of another "dialect".

Even just on data collection, for written or spoken data, our findings show that there are these intrinsic linguistic challenges under the current LRLT development paradigm that are slowing down its development. The standardized orthography and speech are often the hidden prerequisites for the large scale data collection for LT development, and the lack of them for these LRLs becomes the underlying reason holding back the development for LRLTs, yet changing the status quo is out of the reach of these LT practitioners.

4.1.2 Support Crisis

The lack of data can also be exacerbated by the lack of support from LRLT development communities and local governments. Success on the collection

of data for resource-rich languages, such as English and Mandarin Chinese, depends not only on the availability of data, but also on the collective effort of the NLP community to open-source the collective resources. This includes both large-scale web scraping projects such as the CommonCrawl project[6], as well as platforms such as GitHub and HuggingFace for researchers to upload their collected and annotated data. For LRLs, however, we found an interesting pattern that even when corpora are collected, practitioners are often unwilling to open-source them to the rest of the community:

> *"No, I have never encountered cases of sharing corpora, and it is unlikely [...] I bought a corpus from [a university]. They wouldn't share it with me. I had to buy it from them."* (P11)

> *"So some papers of [nationality] researchers saying 'Oh, we did this and that, and we have this big corpus, and we have these numbers', but there's never any data. So it's like, well, good for you, but can't use any of it [...] It's probably very well funded, but nothing comes out of it [...] One paper, and, you know, there's no corpus [...] It's probably good, but we have no idea what it is."* (P13)

Some participants also justified such practice. P14 commented that *"If you spent all the funding and finally made it [...] you might not be willing to share."* Similarly, P6 attribute this practice to some members of the LRLT community and the "immaturity of the field", saying that *"More opportunities of sharing might come from more people. And [English and Mandarin NLP] is relatively more mature. When something is immature – still in its early stage – such sharing will be difficult."*

The lack of data sharing extends to other developed LTs, forcing practitioners to work independently. There is a lack of collaboration between academia and industry (P3), as well as within academia (P11). P1 commented on the lack of collaboration and its impact, stating that there had not been "*much interaction between the AI people with linguists*", or people who all work on Sinitic languages kept "*reinventing the world*" and *"keep doing silly things again and again"* while P1 firmly believes that increasing the conversations between different communities will benefit the LRLT development. These practices of not sharing resources and collaborating, despite justified for funding and competitiveness, hinder the collective effort for the LRLT development.

Another major part of the support comes from the government, which drafts policies and provides funding for practitioners of technological development. In the case of LRLTs, however, we find that such support can often be insufficient, or sometimes entirely missing. P6 commented that the government support can be limited and *"nominal but not real"*, as both P6 and P7 noted that there had not been specialized governmental agencies and policies for the equality of Southern Min despite the government recognizing the endangeredness of the language. For example, while Southern Min is now taught in school, the education, with only 1-2 hours of lessons every week, is usually insufficient for students' mastery of the language.

When discussing about practitioners gathering in various conferences, P1 pointed out that while there are conferences for data collection and even LT development for specific languages, *"conference participation is an expensive thing, and low-resource language people don't have that [money]"*. Almost every phase of LT development, from data collection and model training (P5, P12, P14) to student (P8, P9) / participant recruitment (P8, P12) and conference traveling, depends on the availability of funding, but more than often, securing funding is a major hurdle for LRLT practitioners as what they develop can be deemed as unprofitable (P4, P7, P11), not easily commercialized (P8), or less impactful in academia and industry (P5, P12). They often have to *"compete with [...] not just NLP projects"* (P1), and attempt to improve their chances for funding, whether to expand their work to more profitable and commercializable product with the addition of resource-rich languages (P6, P7), or tweak their proposal by adding *"buzzwords"* (P5) / *"sprinkles on a cupcake"* (P12), demonstrating their projects' values appealing to those providing funding. When funding of merely 10 thousand USD cannot even be guaranteed (P5), LRLT development becomes hardly possible, and the chances of people joining or staying in the field of LRLT development quickly diminish.

4.1.3 Underperforming Basic Tools

Besides data scarcity, participants pointed to the lack of basic tools for LT development as a criti-

[6] http://commoncrawl.org

cal challenge for LRLT development. This insufficiency is not limited to the integration to LLMs, but ranges from the need of evaluation tools such as benchmarks (P3, P15) and auxiliary tools such as OCR (P14) and grapheme-to-phoneme (G2P, P5), to the absence of a good tokenizer (P12, P13), font (P14), or even dictionary (P5). Sometimes the practitioners are forced to work with low-performing tools and noisy corpora (P3, P8), as they are the only ones available.

Ironically, LRLTs are expected to be developed close to perfection to be deployed in real life, as commented by some practitioners. P14 cited the risk of uncontrollable generation, noting that generative LRLTs might lead to racial issues or undesirable public opinions if the generated content cannot be controlled well. P5 and P10 both noted the need for tools such as subtitle generators or G2P to have an error rate within 5% or even smaller to be usable in real life. The successful development of these tools is crucial, especially when part of the challenges come from the lack of profitability (§ 4.1.2) where the commercialization of these NLP tools can help alleviate the problem, but the inability to do so for the lack of resources creates a vicious cycle for the development of LRLTs.

Overall, the intrinsic linguistic challenges, support challenges, and developmental challenges all influence and affect each other, contributing to the overall problem of low-resource-ness and amplifying the hardship in developing LRLTs. This might lead the development of LRLTs into the downward spiral described by the Matthew Effect (Merton, 1968), i.e. "the poor get poorer", causing the challenges even more challenging in the future.

4.2 Challenge Mitigations

Despite the challenges, participants in our interviews deeply care about LRLT development and are highly motivated to preserve the identities and cultures of LRL speakers (P5, P11, P14, etc.) especially when they themselves are native speakers, improve the lives of vulnerable speaker population (P8, P9), and by intellectual curiosity (P12, P15). These motivations encourage practitioners to find ways to mitigate the challenges in LRLT development, including seeking community support and utilizing LLMs in their work.

4.2.1 Community Support

In Section 4.1.2, we listed ways how practitioners in LRLT development are not able to receive support from their communities. While community support in terms of resource sharing and funding has been a hindrance to the development of LRLTs, there have been other ways through which practitioners are able to receive support from different communities, including that from research communities, government and interested individuals.

P1 and P6 both mentioned that there have been international research conferences and workshops, as well as competitions, hosted for researchers developing LRLTs, with P1 pointing out EURALI [7] as a specific workshop for LRLT development in Eurasia. Research labs will also collaborate with IT companies (P1, P6), or with other research labs, sometimes in another country (P3, P15). In the case of Cantonese specifically, there have been practitioner-initiated communities for the LT development, in which researchers and developers would communicate through social media platforms to share and discuss their interests and progress in the Cantonese LT development (P1, P4), which can be helpful for those who just get started (P5).

Despite the comments on the limited funding provided by the government, some participants have also recognized the financial support that they received from the government (P6, P8), or talked about instances when the government supports certain projects that they know of (P1, P14). In addition to the government, individuals or non-governmental organizations could also provide financial support. In the case of Tibetan LTs, P13 mentioned that most of their funding came from religious organizations. The contribution of supporting individuals includes not only the financial aspect, but they can also help spreading the message about a new LT product on social media and paying for the product (P10), or coming together and helping with beta testing (P1).

Undoubtedly there is still a need for more support from various communities, whether they are governmental or non-governmental organizations, research groups, or interested benefactors, for the development of LRLTs. Nevertheless, the current progress for LRLT development depends heavily on the financial, emotional and practical support provided by the existing communities.

4.2.2 The Rise of LLMs

We hypothesized that many LRLTs practitioners hold a negative attitude towards LLMs for their limited performances on LRLs (Court and Elsner,

[7] https://sites.google.com/view/eurali/

2024; Jadhav et al., 2025) and the social and ethical considerations (Smart et al., 2024). Surprisingly, the majority carries a positive attitude, finding LLMs to be helpful for LRLT development.

As all three of the languages in our study are included in various LLMs to a certain extent, many have found that these LLMs could help with some of the most critical issues faced by the LRLT development, including the lack of data and basic tools. Participants such as P1, P3 and P11 discussed leveraging LLMs to generate LRL data (through data augmentation) for model training. P12 described a pipeline for human-LLM collaboration in which an LLM can help with data annotation, which can be expensive if done manually, and practitioners developing filters using different heuristics to improve the quality of the automatic annotations. This aligns with some of the recent literature describing the usage of LLMs to help with various aspects of LRLT development, such as annotation (Kholodna et al., 2024) and data augmentation (Wang et al., 2025). With the challenge in language education (§ 4.1.2), P6 proposed that LLMs can be used to help language instructors to prepare teaching materials, or be customized to check the students' learning progress. Most of the participants believed that "LLM is a trend" (P15), and generally felt optimistic about the positive impact, i.e. helping with basic language tasks, assisting with language education, and potential future capabilities, that LLMs can bring to the development of LRLTs.

However, with all the benefits and practical assistance listed above, many participants are also having concerns about how the LLMs will shape the future for LRLs and LRLTs, even for those who find LLMs useful. *"If we envision our products to be able to translate everything and we don't have to learn but just speak our own languages, then it becomes a challenge [...] If everyone is too lazy to learn, then wouldn't it accelerate the extinction of low-resource languages?"*, a concern of LLMs discouraging people from learning other languages expressed by P9 and also shared by P7. Similarly, P6 was worried about the LLMs' outstanding performance on resource-rich languages potentially forcing users to give up on using LRLs. Some voiced their distrust in using LLMs for LRLT development due the their lack of interpretability (P8), unexplainable reasoning ability (P8) and hallucination in their generations (P5).

As we see, the LRLT practitioners are actively enlisting the assistance from communities that they know of and the capabilities of LLMs to help them combat the innate challenges of LRLT development. However, the help that they are receiving is far from enough, and we call for increasing attention and assistance from relevant stakeholders.

5 Discussion

In this section, we discuss how our findings in § 4.1 corroborate and substantiate prior works discussing LRLT development, as well as implications and recommendations.

The challenges we identified correspond to the aspects categorized in Nigatu et al. (2024) from where the term "low-resource" borrow. The intrinsic linguistic difficulties (§ 4.1.1) and the struggling basic tools (§ 4.1.3) tie to the lack of artifacts, and the lack of support results in the socio-political, resource and agency aspects of the low-resourcedness of these languages. In addition, our findings highlights the complexity of some of these aspects. For example, while Nigatu et al. (2024) and Coto-Solano (2022) both noted the lack of standardization being a feature of low-resourcedness, our findings in § 4.1.1 show that simply having a standardized orthography is not enough to resolve the challenges related to orthography: when multiple standards are present, a standard is not adopted by the native speakers, or if it is only used in certain occasions, LRLs will remain low in digital resource. Similarly, while the challenge to secure funding is universal across all participants, sources of funding where it is difficult to secure are different for different practitioners (§ 4.1.2), resulting in them seeking out to different opportunities too (§ 4.2.1). The situated nature of these challenges calls for case-specific investigations and recommendations.

Moreover, while Nigatu et al. (2024) concludes the causal links between the four aspects, with the resource availability limited by the socio-political issues and impacting the artifacts built. In addition to their acknowledgment of the non-linearity in such causal links (with the examples of socio-political issues impacting artifacts and resources impacting agency directly), our findings show that the causal links can be multidirectional too. For instance, while the lack of funding and human resource impacts the quality of datasets and technologies, the inevitably average-performing technologies also limit their profitability (§ 4.1.2) and for those deciding funding allocation, this becomes a reason against more funding. Curation of datasets

depends on data sharing, but the lack of datasets restricts the amount of shareable data (§ 4.1.1). In these cases, rather than a one-directional causal link where for the four aspects of socio-political, resources, artifacts and agency, each only affects the aspects after it, we see in our findings that they can also affect the aspects before them (first example, artifacts impacting resources) and even themselves (second example, artifacts impacting artifacts).

Last but not least, our interactions with the participants show that the main motivation for LRLT development is often tied to their identities (§ 4.2). Most of the LRLs in the NLP literature happen to be marginalized languages in societies where they are spoken (Smart et al., 2024). Siminyu et al. (2023) has also pointed out that native speakers are motivated to develop LTs for their languages. Sociolinguistics research has shown the identity not just correlated, but produced by the use of one's language (Bucholtz and Hall, 2005; Eckert, 2012). Developing LRLTs for native speakers is therefore a crucial step in promoting the use of their native languages, forging identities, as well as preserving language and culture.

Moving forward, we provide the following recommendations for LRLT practitioners, the broader NLP community and local authorities. While there have been communities formed around Sino-Tibetan LRLT practitioners as we see in § 4.2.1, we still see the potential to improve Sino-Tibetan LRLT development in further expanding communities and increasing connections among each other. Such community can model after Masakhane (Orife et al., 2020), a grassroot organization for African LT development. Acknowledging the lack of focus African LTs, opaqueness of the research effort, and the socio-economic factors hindering connections, Masakhane has expanded from 144 to over 3,000 participants in five years (Issaka et al., 2025), publishing and releasing high-quality research artifacts. This serves as a successful example of how a well connected community is able to overcome the challenges faced by individual practitioners. LRLT practitioners could also engage more with native speakers of the languages by involving them in various phases of LT development to understand the need and disagreement from LT users and evaluate the developed LTs to better assist the identification of future directions (Matias and Price, 2025). Nevertheless, communications and collaboration between practitioners from different regions can be controlled by the local governments. For the collective effort in developing LRLTs, we encourage the governments to ease the restrictions regarding research collaborations. Local governments should also understand how these languages often connect with the speakers' identities and how children are able to achieve bilingualism (Byers-Heinlein and Lew-Williams, 2013). There could be specific policy targeting the education of these marginalized languages, such as having the marginalized language as a mandatory subject in primary and secondary education, and recognizing the official status of these languages and promote their usage in governmental communication. The broader NLP community should continue and increase the effort to welcome LRLT research and practitioners, by lowering the cost for conference registration, having dedicated paper submission tracks, workshops and even conferences. NLP practitioners can continue to develop LRL-specific or language-independent tools. The prevalent use of LLMs in LRLT development suggests the need for effort to build LLMs to be more inclusive. Funding opportunities dedicated to LRLT development could be increased and made available by governments, NLP communities, companies and individuals. Different stakeholders can also convene to discuss and understand each other's perspectives, shaping a more sustainable path for LRLT development.

6 Conclusion

In this paper, we explored the challenges faced and the ways employed by the LRLT practitioners to mitigate these challenges through semi-structured interviews with practitioners working on Sino-Tibetan LRLTs. While the obvious challenge is the lack of textual data, we find that the linguistic difficulties for LT development, lack of funding and research support, and the need for usable basic tools. These challenges both hinder the LRLT development, and provide an explanation to the low-resource-ness of the languages. Despite these difficulties, we found that practitioners are generally optimistic towards the progress of LRLT development, with the help that they receive from the LRLT and language communities and the utility of LLMs for basic language tasks. We identify these challenges and mitigations, calling for more attention and support from the NLP communities, and advocate for the increasing development of LRLTs for and by the native speakers.

Limitations

Our recruitment process involves emailing practitioners whom we can find online and snowball sampling by connecting with more participants through practitioners we have already connected with. We acknowledge that this recruitment process likely results in our participants being those who are already connected with other practitioners and who are open to voice out their experience. We also faced challenges in our recruitment process when participants did not respond to emails, or even refuse to participate in interviews. The three languages selected, i.e. Cantonese, Southern Min and Tibetan, are languages with the richest resources among all Sino-Tibetan LRLs. Practitioners working on other Sino-Tibetan LRLs might be facing more severe challenges and less ways to mitigate them. We therefore consider our findings to be a subset of all challenges and mitigations faced by Sino-Tibetan LRLT practitioners. Future research can try to expand the investigation to practitioners that are less connected with the broader NLP community and other practitioners, as well as comparing challenges across different communities.

References

Muhammad Dehan Al Kautsar, Lucky Susanto, Derry Tanti Wijaya, and Fajri Koto. 2025. What do indonesians really need from language technology? a nationwide survey. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 7231–7256, Suzhou, China. Association for Computational Linguistics.

Mary Bucholtz and Kira Hall. 2005. Identity and interaction: a sociocultural linguistic approach. *Discourse Studies*, 7(4-5):585–614.

Krista Byers-Heinlein and Casey Lew-Williams. 2013. Bilingualism in the early years: What the science says. *Learn Landsc*, 7(1):95–112.

Tsz Chung Cheng, Chung Shing Cheng, Chaak-ming Lau, Eugene Lam, Wong Chun Yat, Hoi On Yu, and Cheuk Hei Chong. 2025. HKCanto-eval: A benchmark for evaluating Cantonese language understanding and cultural comprehension in LLMs. In *Proceedings of the 29th Conference on Computational Natural Language Learning*, pages 1–11, Vienna, Austria. Association for Computational Linguistics.

Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. XNLI: Evaluating cross-lingual sentence representations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2475–2485, Brussels, Belgium. Association for Computational Linguistics.

Andrea Cornwall and Rachel Jewkes. 1995. What is participatory research? *Social Science & Medicine*, 41(12):1667–1676.

Rolando Coto-Solano. 2022. Evaluating word embeddings in extremely under-resourced languages: A case study in Bribri. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 4455–4467, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Sara Court and Micha Elsner. 2024. Shortcomings of LLMs for low-resource translation: Retrieval and understanding are both the problem. In *Proceedings of the Ninth Conference on Machine Translation*, pages 1332–1354, Miami, Florida, USA. Association for Computational Linguistics.

Barbara Czarniawska. 2014. *Social Science Research: From Field to Desk*, first edition edition. SAGE Publications Ltd, 55 City Road, London.

Penelope Eckert. 2012. Three waves of variation study: The emergence of meaning in the study of sociolinguistic variation. *Annual Review Anthropology*, 41:87–100.

Ilker Etikan, Sulaiman Abubakar Musa, Rukayya Sunusi Alkassim, and 1 others. 2016. Comparison of convenience sampling and purposive sampling. *American journal of theoretical and applied statistics*, 5(1):1–4.

Marzieh Fadaee, Arianna Bisazza, and Christof Monz. 2017. Data augmentation for low-resource neural machine translation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 567–573, Vancouver, Canada. Association for Computational Linguistics.

Barney G. Glaser and Anselm L. Strauss. 1967. *The discovery of grounded theory: Strategies for qualitative research*. Aldine de Gruyter.

Emma Harvey, Emily Sheng, Su Lin Blodgett, Alexandra Chouldechova, Jean Garcia-Gathright, Alexandra Olteanu, and Hanna Wallach. 2025. Understanding and meeting practitioner needs when measuring representational harms caused by LLM-based systems. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 18423–18440, Vienna, Austria. Association for Computational Linguistics.

Kung Yin Hong, Lifeng Han, Riza Batista-Navarro, and Goran Nenadic. 2024. CantonMT: Cantonese-English neural machine translation looking into evaluations. In *Proceedings of the 16th Conference of the Association for Machine Translation in the Americas (Volume 2: Presentations)*, pages 133–144, Chicago, USA. Association for Machine Translation in the Americas.

Sheriff Issaka, Keyi Wang, Yinka Ajibola, Oluwatumininu Samuel-Ipaye, Zhaoyi Zhang, Nicte Aguillon Jimenez, Evans Kofi Agyei, Abraham Lin, Rohan Ramachandran, Sadick Abdul Mumin, Faith Nchifor, Mohammed Shuraim, Lieqi Liu, Erick Rosas Gonzalez, Sylvester Kpei, Jemimah Osei, Carlene Ajeneza, Persis Boateng, Prisca Adwoa Dufie Yeboah, and Saadia Gabriel. 2025. The african languages lab: A collaborative approach to advancing low-resource african nlp. *Preprint*, arXiv:2510.05644.

Suramya Jadhav, Abhay Shanbhag, Amogh Thakurdesai, Ridhima Sinare, and Raviraj Joshi. 2025. On limitations of LLM as annotator for low resource languages. In *Proceedings of the 8th International Conference on Natural Language and Speech Processing (ICNLSP-2025)*, pages 277–282, Southern Denmark University, Odense, Denmark. Association for Computational Linguistics.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The state and fate of linguistic diversity and inclusion in the NLP world. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6282–6293, Online. Association for Computational Linguistics.

Nataliia Kholodna, Sahib Julka, Mohammad Khodadadi, Muhammed Nurullah Gumus, and Michael Granitzer. 2024. Llms in the loop: Leveraging large language model annotations for active learning in low-resource languages. In *Machine Learning and Knowledge Discovery in Databases. Applied Data Science Track: European Conference, ECML PKDD 2024, Vilnius, Lithuania, September 9–13, 2024, Proceedings, Part X*, page 397–412, Berlin, Heidelberg. Springer-Verlag.

Eleanor Knott, Aliya Hamid Rao, Kate Summers, and Chana Teeger. 2022. Interviews in the social sciences. *Nat. Rev. Methods Primers*, 2(1).

Yen-Chun Lai, Yi-Jun Zheng, Wen-Han Hsu, Yan-Ming Lin, Cheng-Hsiu Cho, Chlh-Chung Kuo, Chao-Shih Huang, and Yuan-Fu Liao. 2024. Construction of large language models for taigi and hakka using transfer learning. In *2024 27th Conference of the Oriental COCOSDA International Committee for the Coordination and Standardisation of Speech Databases and Assessment Techniques (O-COCOSDA)*, pages 1–6.

Longhao Li, Zhao Guo, Hongjie Chen, Yuhang Dai, Ziyu Zhang, Hongfei Xue, Tianlun Zuo, Chengyou Wang, Shuiyuan Wang, Jie Li, Jian Kang, Xin Xu, Hui Bu, Binbin Zhang, Ruibin Yuan, Ziya Zhou, Wei Xue, and Lei Xie. 2025. Wenetspeech-yue: A large-scale cantonese speech corpus with multi-dimensional annotation. *Preprint*, arXiv:2509.03959.

Evelyn Kai-Yan Liu. 2022. Low-resource neural machine translation: A case study of Cantonese. In *Proceedings of the Ninth Workshop on NLP for Similar Languages, Varieties and Dialects*, pages 28–40, Gyeongju, Republic of Korea. Association for Computational Linguistics.

Shuheng Liu and Michael Best. 2025. A survey of NLP progress in Sino-Tibetan low-resource languages. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 7804–7825, Albuquerque, New Mexico. Association for Computational Linguistics.

J. Nathan Matias and Megan Price. 2025. How public involvement can improve the science of ai. *Proceedings of the National Academy of Sciences*, 122(48):e2421111122.

Marieke Meelen, Élie Roux, and Nathan Hill. 2021. Optimisation of the largest annotated tibetan corpus combining rule-based, memory-based, and deep-learning methods. *ACM Trans. Asian Low-Resour. Lang. Inf. Process.*, 20(1).

Robert K. Merton. 1968. The matthew effect in science. *Science*, 159(3810):56–63.

Wilhelmina Nekoto, Vukosi Marivate, Tshinondiwa Matsila, Timi Fasubaa, Taiwo Fagbohungbe, Solomon Oluwole Akinola, Shamsuddeen Muhammad, Salomon Kabongo Kabenamualu, Salomey Osei, Freshia Sackey, Rubungo Andre Niyongabo, Ricky Macharm, Perez Ogayo, Orevaoghene Ahia, Musie Meressa Berhe, Mofetoluwa Adeyemi, Masabata Mokgesi-Selinga, Lawrence Okegbemi, Laura Martinus, and 28 others. 2020. Participatory research for low-resourced machine translation: A case study in African languages. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2144–2160, Online. Association for Computational Linguistics.

Hellina Hailu Nigatu, Atnafu Lambebo Tonja, Benjamin Rosman, Thamar Solorio, and Monojit Choudhury. 2024. The Zeno's paradox of 'low-resource' languages. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 17753–17774, Miami, Florida, USA. Association for Computational Linguistics.

Iroro Orife, Julia Kreutzer, Blessing Sibanda, Daniel Whitenack, Kathleen Siminyu, Laura Martinus, Jamiil Toure Ali, Jade Abbott, Vukosi Marivate, Salomon Kabongo, Musie Meressa, Espoir Murhabazi, Orevaoghene Ahia, Elan van Biljon, Arshath Ramkilowan, Adewale Akinfaderin, Alp Öktem, Wole Akin, Ghollah Kioko, and 6 others. 2020. Masakhane – machine translation for africa. *Preprint*, arXiv:2003.11529.

Leiyu Pan, Bojian Xiong, Lei Yang, Renren Jin, Shaowei Zhang, Yue Chen, Ling Shi, Jiang Zhou, Junru Wu, Zhen Wang, Jianxiang Peng, Juesi Xiao, Tianyu Dong, Zhuowen Han, Zhuo Chen, Yuqi Ren, and Deyi Xiong. 2025. Advancing large language models for tibetan with curated data and continual pre-training. *Preprint*, arXiv:2507.09205.

Charlie Parker, Sam Scott, and Alistair Geddes. 2019. Snowball sampling. *SAGE research methods foundations*.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21(1).

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392, Austin, Texas. Association for Computational Linguistics.

Sebastian Ruder, Ivan Vulić, and Anders Søgaard. 2019. A survey of cross-lingual word embedding models. *J. Artif. Int. Res.*, 65(1):569–630.

Johnny Saldaña. 2009. *The coding manual for qualitative researchers.* Sage Publications Ltd.

Jiu Sha, Mengxiao Zhu, Chong Feng, and Jizhuoma Ci. 2025. Tibetan-llama 2: Large language model for tibetan. *ACM Trans. Asian Low-Resour. Lang. Inf. Process.*, 24(12).

Kathleen Siminyu, Jade Abbott, Kọ́lá Túbọ̀sún, Aremu Anuoluwapo, Blessing K. Sibanda, Kofi Yeboah, David Adelani, Masabata Mokgesi-Selinga, Frederick R. Apina, Angela Thandizwe Mthembu, Arshath Ramkilowan, and Babatunde Oladimeji. 2023. Consultative engagement of stakeholders toward a roadmap for african language technologies. *Patterns*, 4(8):100820.

Andrew Smart, Ben Hutchinson, Lameck Mbangula Amugongo, Suzanne Dikker, Alex Zito, Amber Ebinama, Zara Wudiri, Ding Wang, Erin van Liemt, João Sedoc, Seyi Olojo, Stanley Uwakwe, Edem Wornyo, Sonja Schmer-Galunder, and Jamila Smith-Loud. 2024. Socially responsible data for large multilingual language models. *Preprint*, arXiv:2409.05247.

Nicolas Tournadre. 2014. The tibetic languages and their classification. In Thomas Owen-Smith and Nathan Hill, editors, *Trans-Himalayan Linguistics*, pages 105–130. De Gruyter Mouton, Berlin, Boston.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aur'elien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. Llama: Open and efficient foundation language models. *ArXiv*, abs/2302.13971.

Lisa M. Vaughn and Farrah Jacquez. 2020. Participatory Research Methods – Choice Points in the Research Process. *Journal of Participatory Research Methods*, 1(1).

Yurong Wang, Min Lin, Qitu Hu, Longjie Bao, Shuangcheng Bai, and Yanling Li. 2025. Large and small models for collaborative cross-lingual data augmentation in entity relationship extraction for low-resource languages. *Journal of King Saud University Computer and Information Sciences*, 37(4):56.

Jiahao Yang, Jianguo Wei, Kuntharrgyal Khysru, Junhai Xu, Wenhuan Lu, Wenjun Ke, and Xiaokang Yang. 2023. Effective fine-tuning method for tibetan low-resource dialect speech recognition. In *2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, pages 960–965.

Xinyan Yu, Trina Chatterjee, Akari Asai, Junjie Hu, and Eunsol Choi. 2022. Beyond counting datasets: A survey of multilingual dataset construction and necessary resources. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 3725–3743, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Mingjun Zhou, Quzong Gesang, Nuo Qun, Tashi Nyima, and Dongrub Rinchen. 2024. Tibetan-chinese machine translation enhanced on cross-lingual pretrained model. In *2024 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, pages 1618–1623.

A Appendix

A.1 Participant Consent Form

Here is the full text for the consent form with redaction for anonymity.

> You are being asked to be a volunteer in a research study. The purpose of this study is to investigate the assets and challenges of the research communities working on low-resource Sino-Tibetan language technology by interviewing experts in this field. The interview will take approximately 1 hour to complete.
>
> To protect your privacy, your records will be kept under a code number rather than by name. Your name will only be collected when signing the consent forms. The signed consent forms will be kept in a secured locker and unless you give specific consent otherwise, only study staff will be allowed to look at them. Your name and any other fact that might point to you will not appear when results of this study are presented or published.
>
> The risks involved are no greater than those involved in daily activities. You will not likely directly benefit for joining this study, but your participation will help the NLP research community to understand the direction for future development of low-resource language technology.
>
> We will comply with any applicable laws and regulations regarding confidentiality. To make sure that this research is being carried out in the proper way, the [redacted] IRB may review study records. The [redacted] may also look at study records.
>
> If you have any questions about the study, you may contact [redacted] at telephone [redacted], or via email at [redacted]. If you have any questions about your rights as a research subject, you may contact [redacted] at [redacted]. Thank you for participating in this study.