Akha, Dara-ang, Karen, Khamu, Mlabri and Urak Lawoi' language minorities' subjective perception of their languages and the outlook for development of digital tools

Joanna Dolińska
University of Warsaw
j.dolinska@al.uw.edu.pl

Shekhar Nayak
University of Groningen
s.nayak@rug.nl

Sumittra Suraratdecha
Mahidol University
sumittra.sur@mahidol.edu

Abstract

Multilingualism is deeply rooted in the sociopolitical history of Thailand. Some minority language communities entered the Thai territory a few decades ago, while the families of some other minority speakers have been living in Thailand since at least several generations. The authors of this article address the question how Akha, Dara-ang, Karen, Khamu, Mlabri and Urak Lawoi' language speakers perceive the current situation of their language and whether they see the need for the development of digital tools for documentation, revitalization and daily use of their languages. The objective is complemented by a discussion on the feasibility of development of such tools for some of the above-mentioned languages and the motivation of their speakers to participate in this process. Furthermore, this article highlights the challenges associated with developing digital tools for these low-resource languages and outlines the standards researchers must adhere to in conceptualizing the development of such tools, collecting data, and engaging with the language communities throughout the collaborative process.

1 Introduction

The region of Southeast Asia encompasses biologically and linguistically distinct countries. It is characterized by a high level of biodiversity (Tungittiplakorn and Dearden, 2002) and is renowned for its rich linguistic landscape (Enfield, 2021; Lee, 2019; Prasithrathsing, 1988).

However, biodiversity and multilingualism in this region are becoming endangered due to globalization, industrialization and intensive tourism. Available data suggest that the simultaneous disappearance of linguistic and biological diversity is correlated (Gorenflo et al., 2012). This article primarily focuses on six minority language communities in Thailand, located in the provinces of Chiang Mai, Chiang Rai, Nan and Krabi. It presents preliminary findings from interviews with representatives of the Akha, Dara-ang, Karen, Khamu, Mlabri and Urak Lawoi' communities, and discusses efforts to collect audio data for developing voice technology for some of these languages. The aim of this article is to explore concrete steps required to foster more effective collaboration between computer scientists, documentary linguists, and language communities in Thailand. It also proposes technologies for tasks in low-resource settings, particularly for the discussed endangered languages from four selected Thai provinces. Furthermore, it is hoped that the presented findings will initiate a discussion on how to approach the development of digital tools for endangered and low-resource languages.

Data collection has been conducted following the "data statements" practice developed by Bender and Friedman (Bender and Friedman, 2018). The aim of this practice is to create digital tools that avoid the risk of oversimplyfing the situation of any given speech communities and prevent their exclusion, underrepresentation and misrepresentation. The preliminary data presented here was collected in November and December 2023 across four provinces of Thailand: Chiang Mai, Chiang Rai, Nan and Krabi. This data includes approximately 18 hours of recordings and notes involving the participation of 16 adults (3 women and 13 men) aged between 30-94 years at the time of the recordings.

2 Data statements

Curator rationale: Interviews have been conducted with the representatives of language minorities in Thailand in order to learn about their subjective perception of the condition of their languages, to hear out their opinions concerning the needs of their communities and to inquire if these language communities would like to cooperate

Proceedings of the The Seventh Workshop on the Use of Computational Methods in the Study of Endangered Languages, pages 94–99

Language	No. of persons	Province
Akha	2	Chiang Rai
Dara-ang	3	Chiang Mai
Karen	3	Chiang Mai
Khamu	1	Chiang Rai
Mlabri	2	Nan
Urak Lawoi'	5	Krabi

Table 1: Number of native speakers of each minority language that the authors interviewed in 4 different provinces[1] in November and December 2023.

with the interviewers on developing useful digital tools for their languages. **Language variety:** The represented speakers' variety includes everyday, contemporary, spoken language varieties of Akha, Dara-ang, Karen, Khamu, Mlabri and Urak Lawoi' from the Chiang Mai, Chiang Rai, Nan and Krabi provinces in Thailand. Interviews have been carried out with the support of three translators communicating with the interviewees in Northern, Southern and standard Thai. All interviewees are at least bilingual. Apart from their native languages, they speak Northern Thai in their everyday communication (9 persons), Southern Thai (5 persons) and standard Thai (14 persons). **Speaker demographic:** All 16 interviewed minority language speakers are adults aged between 30 and 94 years. There has been a noticeable tendency for men to be more willing to participate in the interviews than women, with 13 men and 3 women participating. This might result from the fact, that among the interviewed communities men tend to interact more with other social groups and villages. The speakers represent various professions: one community activist, one community activist/social entrepreneur, one healer, one farmer/pastor, five farmers, two farmers/hunter-gatherers, one fisherman/teacher, three fishermen, and one tourism sector employee. All interviewees with exception of one person have attended at least primary school and two conversation partners hold university degrees. Three interviewees migrated to Thailand several decades ago from Myanmar and Laos, while 13 have lived in their province, or a neighbouring province in one case, their entire lives. The interviewees represent diverse religious backgrounds: seven are Buddhist, four follow local beliefs, four are Christian and not known (1 person). Furthermore, all interviewees were asked the following two questions: **[1]** "How would you describe your identity? Are you *Akha /Dara-ang /Karen /Khamu /Mlabri /Urak Lawoi'* **and/or** Thai?" **[2]** "Alternatively, are you Thai and *Akha /Dara-ang /Karen /Khamu /Mlabri /Urak Lawoi*'?" Without exception, all interviewees answered that they feel first *Akha /Dara-ang /Karen /Khamu /Mlabri /Urak Lawoi'*, depending on their community of origin. Almost all interviewees have Thai citizenship, except for one person who has been actively applying for it for several years and continues to strive for full citizenship rights in Thailand. Three interviewees went through distressing refuge experiences in their early adolescence and adulthood periods. **Speech situation:** Twelve interviews took place in the village settings of the interviewees, two in a cafeteria, one at a school, and one in a church community area. The interviewees were informed about the interview topics beforehand and communicated with the interviewers in Northern Thai and standard Thai (10 persons), Southern Thai and standard Thai (4 persons) and English (two persons). The structured interview comprised four sections: general questions, questions about language use, question about the natural environment, and questions related to work and leisure domains. Almost all interviews lasted approximately one hour. All interviewees responded to questions concerning their subjective perception of the sociolinguistic situation within their respective communities. A preliminary set of audio-data has been collected for the Karen language, including a wide array of plant names, and for the Urak Lawoi' language, featuring names of celebrated Deities, basic vocabulary concerning time perception, human body parts, verbs for basic human activities, names of colours and natural phenomena, as well as terms related to fishery, as it is the most essential mode of subsistence for this community. Furthermore, it has been discovered that the Dara-ang community of Christian denomination in the Chiang Mai province has translated Bible into the Dara-ang language with the support

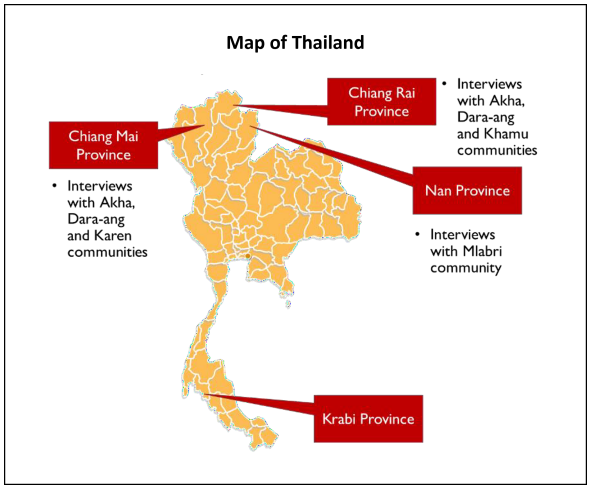

Figure 1: Interviews conducted with the representatives of language minorities in Thailand in November and December 2023.

of foreign religion activists and has supplemented parts of it with numerous audio-recordings which are available through electronic application.[2]

3 Data collection

3.1 Preliminary findings concerning independent efforts of the interviewed language communities to document and revitalize their languages

All interviewed representatives of language communities recognize the need to preserve their language for future generations. They acknowledged the primary role of **oral language** transmission within the family and among members of the same village. Furthermore, the Akha, Dara-ang, Karen and Urak Lawoi' communities each possess their own written standard, which has been developed by these communities with the support of researchers and/or missionaries. Notably, the Akha writing system encompasses Akha varieties spoken not only in Thailand, but also in neighbouring countries. The written forms of Dara-ang and Urak Lawoi' languages are based on the Thai script. However, it should be noted that the transmission of the **written standard** in these communities has not been consistent over the past decades, and not all community members can read it.

In fact, the aforementioned language communities primarily depend on oral transmission, which may also influence the potential development of **digital tools** for these languages in the future. Among the interviewed representatives of language minorities, the Akha group appears to have advanced its own language policy to the most successful stage, yet the motivation to preserve the language remains strong across all the encountered communities. Notably, the motivation to preserve the Dara-ang language among the Christian community in the Chiang Mai province seems to have been inspired by the desire to practice their religion in their first language.

When asked about any potential need for digital tools to preserve their community languages, interviewees from the Akha, Dara-ang, Karen, Mlabri and Urak Lawoi' communities expressed interest. Akha interviewees were primarily interested in developing digital tools for educational purposes, while the Karen speakers would like to use the digital tools to preserve the indigenous knowledge of the Karen community pertaining to the names and images of healing plants and their application. Mlabri and Urak Lawoi' interviewees both shared the views that it would be beneficial to use their native languages when using mobile phones and one of the Urak Lawoi' interviewees positively approached the idea of developing digital educational materials in Urak Lawoi' language.

3.2 Future plans for the development of voice technology

The interviewee from the Akha community has shown a considerable interest in creating a voice technology tool for the Akha language, especially in the speech-to-text tool. Given the existing abundance of the Akha language written materials for educational purposes, the authors of this article have drafted a plan with the Akha community representative to start Akha language recordings in 2024 in order to provide the data for the development of the audio tool. A positive outcome of this cooperation between sociolinguists, a computational linguist and language activist is to be expected, since it combines Akha community experience, professional academic knowledge and skills, as well as the self-driven motivation of the Akha community to create Akha language resources.

The second future opportunity to develop digital tools for one of the above languages is to create an offline and online educational brochure focused on the Karen language, especially on the plants with healing characteristics, which are applied in the

[2]The interviewers were advised by the Dara-ang community to ask the Christian Pastor about the need and feasibility of the development of digital tools, as he would have the knowledge about the authors rights of the translation and recording of the Bible.

traditional Karen medicine and constitute a component of the Karen indigenous knowledge. After consultations with the Karen healer, the authors of this article alongside with the Karen healer carried out video recordings of several dozen of plants, alongside their name in Northern Thai and Karen languages. These names have been recorded in the audio format as well. This initiative will be implemented at the beginning of the year 2024 with the active participation of the Karen community. Like in the case of the Akha community, the most important factor pointing towards the potentially positive outlook of this project is the strong engagement and motivation of the Karen healer to preserve the knowledge about healing plants and herbs that she inherited from her both grandfathers. Thirdly, the collected Urak Lawoi' data encompassing names of local Deities, basic vocabulary concerning time perception, human body parts and fishery, verbs for basic human activities, names of colours and natural phenomena could be the basis for the development of a basic audio dictionary for the Urak Lawoi' language in the future, if the Urak Lawoi' interviewees will be interested in a further cooperation.[3]

3.3 Outlook for multilingual ASR for these endangered languages

There have been recent efforts in the speech technology research towards developing tools which are not pre-dominantly text-driven to support many endangered languages which lack in textual resources (Ewan Dunbar and Dupoux, 2022). A range of strategies has been developed specifically for speech-to-text or automatic speech recognition (ASR) for low resource languages (Nayak and Kodukula, 2019). Multi-lingual and cross-lingual approaches (Schultz and Kirchhoff, 2006) have essentially led to effective use of limited resources of these languages for improved ASR performance. Specifically, the self-supervised learning based models such as wav2vec 2.0 which are pre-trained on large number of diverse languages make the development of speech-to-text simpler as the models aren't trained from scratch for such languages (Babu et al., 2021; Baevski et al., 2020). Also, the untranscribed audio data can be utilized for pre-training these models. Our future goal is to expand our database for these languages and build a common multilingual ASR model supporting these endangered languages utilizing self-supervised and transfer learning methodologies.

3.4 Challenges concerning the development of digital tools for the Akha, Dara-ang, Karen, Khamu, Mlabri and Urak Lawoi' languages

Having analyzed the interviews with the above-mentioned communities, the authors came to the following conclusions. First of all, asking the representatives of minority speech communities, if they would be interested in co-developing digital tools for their language needs to be preceded by a description of how their audio-data could be processed and what kind of benefit it would mean for the community.

A second question refers to the motivation to create such tools. Based on the interview results, it seems that the Akha community representative would like to promote and standardize the Akha language within the Akha community, whereas the Mlabri speakers offered to help acquire Mlabri language skills by the foreigners willing to learn it. Another challenge is the written standard of a given language. If a speech-to-text tool is developed, in which script does the text need to appear? Even if only one type of script has been consequently applied in the history of this language research, orthography variants can also induce complications while developing digital tools. Not only the motivation to develop particular tools is important, but also the target audience. An essential question to ask before developing digital tools for a given low-resource language concerns the data that it needs to encompass, whether they should be accompanied by pictures (if the tool is for children), whether it should include religious and worldview contents (but then which religion and which worldview, as the above-mentioned communities often follow various religions and beliefs within their groups).

Another question relates to the so called "heritage data" for a given language, which means data compiled by the community members themselves, researchers and missionaries in the past. The question which arises here is how to incorporate such data and how to understand the situation related to the authors rights in terms of these data (Blokland et al., 2019).

[3]Two Urak Lawoi' interviewees explicitly stated that they would like to have such tools for their native language in the future.

4 Ethical statement

The design of the project and its implementation follow the "The TRUST Code - A Global Code of Conduct for Equitable Research Partnerships" https://www.globalcodeofconduct.org. The participants of the research were both informed about the consent procedures and the goal of the research in Thai language (as all of them are fluent Thai speakers). Question concerning consents have been asked without recording. If the participants allowed the recording of the interview (which all of them did), they responded to the consent questions once again while being audio-recorded. After the interviews the participants received audio-recordings and photos from the meetings from the interviewers.

4.1 Limitations

Since the authors interviewed only 16 representatives of 6 various language minority groups in Thailand, the collected qualitative data understandably cannot be indicative of the whole speaker populations. Nevertheless, the presented results allow for determining the future direction of the research devoted to the self-perception of language minorities in Thailand. The qualitative research method in the form of a structured interview was perceived by the authors as the best choice for introductory field work research on the sociolinguistic status of minority languages in Thailand and the interest of their speakers in developing digital tools.

5 Acknowledgements

Joanna Dolińska would like to acknowledge the seminal role of the University of Warsaw New Ideas 3A Grant "Interdependence of multilingualism and biodiversity in the Chiang Mai and Satun provinces in Thailand" (2023-2024), as well as the Scholarship for Short-term Visiting Scholars at the Mahidol University for the conceptualization and implementation of this research project together with Sumittra Suraratdecha. Furthermore, Joanna Dolińska would like to acknowledge the importance of Short-Term Scientific Action (STSM) within the LITHME (Languange in the Machine-Human Area) COST Action at the Voice Technology Program, Campus Fryslân, University of Groningen, which inspired the cooperation with Shekhar Nayak on the development of voice technology tools for endangered languages, as well as the conceptualization of this article and future development of voice technologies for the described endangered languages.

References

Arun Babu, Changan Wang, Andros Tjandra, Kushal Lakhotia, Qiantong Xu, Namal Goyal, Kritika Singh, Patrick von Platen, Yatharth Saraf, Juan Pino, and Alexei Baevski. 2021. Xls-r: Self-supervised cross-lingual speech representation learning at scale. arxiv preprint.

Alexei Baevski, Henry Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in neural information processing systems*, 33:12449–12460.

Emily M. Bender and Batya Friedman. 2018. Data statements for natural language processing: Toward mitigating system bias and enabling better science. *Transactions of the Association for Computational Linguistics*, 6:587–604.

Rogier Blokland, Nikko Partanen, Michael Rießler, and Joshua Wilbur. 2019. Using computational approaches to integrate endangered language legacy data into documentation corpora: Past experiences and challenges ahead. *Proceedings of the Workshop on Computational Methods for Endangered Languages*, 2(5).

Nick Enfield. 2021. *The Languages of Mainland Southeast Asia (Cambridge Language Surveys)*. Cambridge University Press, Cambridge.

Nicolas Hamilakis Ewan Dunbar and Emmanuel Dupoux. 2022. Self-supervised language learning from raw audio: Lessons from the zero resource speech challenge. *IEEE Journal of Selected Topics in Signal Processing*, 16(6):1211–1226.

Larry Gorenflo, Suzanne Romaine, Russel A. Mittermeier, and Kristen Walker-Painemilla. 2012. Co-occurrence of linguistic and biological diversity in biodiversity hotspots and high biodiversity wilderness areas. *Proc Natl Aca Sci U S A*, 22;109(21):8032–8037.

Hugo Yu-Hsiu Lee. 2019. Rethinking globalization, english and multilingualism in thailand: A report on a five-year ethnography. *3L – Language Linguistics Literature – The Southeast Asian Journal of English Language Studies*, 2(1):69–84.

Shekhar Nayak and Sri Rama Murty Kodukula. 2019. *Unsupervised speech Signal to Symbol Transformation for Zero Resource Speech Processing*. Doctoral dissertation, Indian Institute of Technology Hyderabad.

Amara Prasithrathsing. 1988. Sociolinguistic research on thailand languages. *Language Sciences*, 10(2):236–272.

Tanja Schultz and Katrin Kirchhoff. 2006. *Multilingual speech processing*. Elsevier, Amsterdam.

Waranoot Tungittiplakorn and Philip Dearden. 2002. Biodiversity conservation and cash crop development in northern thailand. *Pacific Linguistics*, (C-43):2007–2025.