Documented but Invisible: From Language Documentation to Digital Multilingualism

Hao Lin
Suan Sunandha Rajabhat University
E-mail: raylam0423@hotmail.com, raylam0423@gmail.com

Abstract

Researchers have recorded Southeast Asia's endangered languages for decades, producing dictionaries, grammars, and text collections. Yet these languages are almost absent from the corpora, benchmarks, and language models on which digital multilingualism increasingly depends. This paper asks what blocks the path from documentation to digital multilingualism, taking Thailand and its Mahidol documentation tradition as a critical case. The study is a PRISMA-lite analysis: eight searches run in July 2026 across the ACL Anthology, TCI-THAIJO, and ScholarSpace returned 110 records, of which 105 were included. The coded records confirm the gap: none pairs the documentation tradition with a machine-readable corpus of a Thai minority language; outputs are mostly dictionaries, schoolbooks, and archives made for human readers. Two findings follow. First, the main outputs of language documentation are designed for human readers rather than for computer programs. The study terms this mismatch a typological format barrier. Second, the failure sits in mobilization, the stage at which archived records should be turned back into usable resources. The paper contributes (1) a measured account of this barrier, and (2) a three-step pipeline: extracting structured text, adding annotation in tiers, and releasing corpora in phases with community consent. This conversion, the study argues, is a precondition for digital multilingualism among the region's minority languages.

Keywords: language documentation, digital multilingualism, PRISMA-lite analysis

Introduction

When Himmelmann (2006) defined language documentation as "a lasting, multipurpose record of a language", he added that such records are built for users "whose identity is still unknown." Twenty years later, that unknown user has arrived, and it is not a person. It is the machine: the corpora, speech recognisers, translation systems, and language models through which a language now enters the digital world or disappears from it. Kornai (2013) estimated that only about 170 of the world's roughly 7,000 languages show signs of sustained digital use, and he predicted that few of the rest would survive digitally. A language missing from computable data now faces a second, faster kind of death.

This matters especially in Southeast Asia, one of the world's richest areas of linguistic diversity, with over 1,300 indigenous languages (Lovenia et al., 2024). Thailand alone has more than seventy language communities (Premsrirat, 2007) and a documentation tradition widely regarded as a regional model: the community-based approach developed at the Research Institute for Languages and Cultures of Asia (RILCA), Mahidol University, widely known as the Mahidol model. Under the model, communities design their own orthographies, produce

literacy materials, and teach through mother tongue-based education (Premsrirat & Hirsh, 2018). If any tradition should have prepared its languages for the digital era, it is this widely recognised model.

Yet a puzzle remains. These languages are thoroughly documented, but they are invisible in the computational resources on which digital multilingualism increasingly depends. SEACrowd, a major regional language-technology data initiative, lists nearly 1,000 Southeast Asian languages in its catalogue, yet its benchmark covers only 36 (Lovenia et al., 2024). Cross-indexing a major endangered-language archive with the computational literature, Zariquiey, Oncevay and Vera (2022) found that of 286 archived languages, only 22 have any entry in the ACL Anthology and only 12 appear in any massively multilingual dataset. The two communities have, in effect, been working on the same languages without working together.

This paper asks what blocks the path from documentation to digital multilingualism, and what existing archives can do about it. The question is not whether these languages have been neglected. It is whether they lack data, or whether the data exists in forms that computers cannot use. The question is urgent, as computational research is turning toward low-resource languages now, and a language that computers cannot process is shut out of the tools, from spell-checkers to learning platforms.

The paper proceeds as follows. Section 3 states the research objectives, and Section 4 sets out the scope. Section 5 reviews the three bodies of work behind the puzzle: the documentation tradition, research on digital vitality and language technology, and recent projects that turn records into corpora. Section 6 explains the PRISMA-lite methodology, Section 7 reports the results and the proposed conversion pipeline, Section 8 discusses the implications.

Research Objectives

This study pursues two objectives. The first is to examine the gap between what language documentation produces and what modern computational linguistics requires, within Thailand's multilingual context: where the path from documentation to computation breaks, whether in documentation effort, in archiving, or in a later, less examined stage of the workflow. The second is to propose a way to close that gap: a corpus-based framework that turns existing language archives into machine-readable resources, so that documented languages can take part in digital multilingualism.

Scope of the Research

Thailand is the primary research focus, treated as a critical instance of the wider mainland Southeast Asian condition: it combines the region's most established documentation tradition (Premsrirat & Hirsh, 2018) with a typical ecology of minority-language endangerment (Premsrirat, 2007), so a gap found here is structural rather than local. Laos, Myanmar, and regional initiatives such as SEACrowd are consulted for comparison, without claiming full regional coverage. The study works across language documentation, corpus linguistics, and natural language processing, and proceeds as a systematic secondary literature analysis: all empirical claims come from the record retrieved under Section 6, and claims about absence are

bounded by that record. Two terms are fixed throughout: "Thai" is the national language, already well-resourced digitally (Suwanbandit et al., 2023; Thatphithakkul et al., 2024), while "Thailand's minority languages" are its other language communities; and "machine-readable" means explicitly structured in standard, parseable formats (Trilsbeek & Wittenburg, 2006), not merely stored as digital files. Retrieval covers 2000–2026, with earlier classics such as Smalley (1976) kept where their concepts remain in use.

Literature Review

The title of this study joins two ends of one path: from language documentation to digital multilingualism. The review traces that path. It begins with the design of the documentary record, turns to the demands that computation places on such records, and closes with the first attempts to bring the two together.

1. What documentation produced

Language documentation was historically designed around human users. In Thailand, the pioneering Mahidol model catalogued the country's endangered languages (Premsrirat, 2007) and collaborated with communities to produce orthographies, dictionaries, and school curricula (Smalley, 1976; Premsrirat & Hirsh, 2018; Ungsitipoonporn et al., 2021). Whilst invaluable for community revitalisation, these outputs were formatted strictly for print and human readership. Consequently, this human-centred design inadvertently created a 'typological format barrier', validating early warnings that large collections gathered without structural foresight risk ending up as unusable 'data graveyards' (Himmelmann, 2006).

2. What computation requires

Where traditional documentation prioritises human readability, computational linguistics demands explicitly structured, machine-readable formats (Trilsbeek & Wittenburg, 2006). A parser cannot naturally recover linguistic structure from visual print layouts or non-standard fonts (Seifart, 2006). When these computational standards are unmet, the resulting archives suffer from what Adebara (2025) terms 'language data flaring'—data exists but remains trapped in isolated, unusable formats. This failure to mobilise existing records explains why thoroughly documented endangered languages are paradoxically rendered 'low-resourced' in the digital sphere (Hämäläinen, 2021), leading to their near-total absence from computational literature (Zariquiey et al., 2022) and digital vitality metrics (Kornai, 2013; Simons et al., 2022).

3. What has been tried

To overcome this format barrier and make legacy data computable, recent initiatives have begun restructuring archives, albeit unevenly. In Thailand, projects such as the Thai Dialect Corpus (Suwanbandit et al., 2023), LOTUS-TRD (Thatphithakkul et al., 2024), and TTS systems (Geng et al., 2025) successfully apply modern NLP methods, yet they remain restricted to regional dialects of the national language, bypassing minority languages entirely. To bridge this gap, global initiatives like the Pangloss Collection (Adamou et al., 2025) demonstrate that legacy archives can be successfully converted into citable, machine-readable corpora. Replicating this conversion for Southeast Asia (distributing via hubs like SEACrowd; Lovenia et al., 2024) is the necessary next step, provided that such technical interventions

strictly prioritise community consent to avoid extractive research patterns (Bird, 2020; Liu et al., 2022).

Research Methodology

The study is a systematic secondary literature analysis conducted under a PRISMA-lite protocol, a proportionate application of systematic-review reporting standards (identification, screening, inclusion) to the scale of a single study. No primary fieldwork data were collected. All empirical claims derive from the published record retrieved under the protocol below, and claims about absence are bounded by that record.

Searches were executed in July 2026 through a scholarly search engine. Three source families were covered: computational linguistics, indexed via the ACL Anthology; Southeast Asian area scholarship, including TCI-THAIJO and JSEALS; and language documentation, including Language Documentation & Conservation via ScholarSpace. Three Boolean strings mirrored the study's three layers: the contact layer, ("language contact" OR "multilingualism" OR "mother tongue education") AND ("Thailand" OR "MSEA"); the documentation layer, ("language documentation" OR "Mahidol model" OR "orthography") AND ("minority languages" OR "endangered languages"); and the computational layer, ("digital vitality" OR "NLP" OR "corpus" OR "computational") AND ("low-resource" OR "Thai dialects"). Two snowball operations complemented the strings: forward citation-chasing from Zariquiey et al. (2022), which links documentation and computation, and backward citation-chasing from Premsrirat's surveys, filtered for digital and corpus terms and for recency from 2018 onward. The eight batches returned 110 records.

Records were deduplicated by normalised title, which removed 5 duplicates, and screened on title and abstract against three criteria: publication from 2000 onward, with earlier classics retained where they anchor a concept still in use (Smalley, 1976); topical relevance to language documentation, endangered languages, digital vitality, or low-resource language technology; and relevance to Thailand or mainland Southeast Asia, or direct supply of the study's comparative frame (Kornai, 2013; Hämäläinen, 2021). No record failed screening, and 105 records were included.

With the assistance of large language models (Kimi K3 and Gemini 3.5), titles and abstracts were coded on four dimensions. (1) Thematic block: T1 documentation and revitalisation; T2 contact and multilingual education; T3 digital vitality; T4 NLP and low-resource computation. Records matching none of the keyword sets were coded T0 (Other) and are reported separately. (2) Output type: NLP resource/corpus; archive/documentation practice; orthography/dictionary/grammar; policy/education; revitalisation/community practice; theory/review. (3) Publication year. (4) Keyword sets, normalised to canonical concepts (for example, corpora to corpus; speech recognition to ASR). Three analyses were then computed and visualised using Python: a theme × output-type matrix, a keyword co-occurrence network with edges at a within-record co-occurrence of 3 or more, and a yearly distribution by block.

Research Results

Figure 1 maps the four thematic blocks against output types, revealing a clear split between documentation and computational research. Whereas both T1 (Documentation and revitalisation) and T4 (NLP and low-resource computation) are major producers of 'NLP resource / corpus materials' (n=21 and n=14 respectively), their outputs show little overlap. T4 output is restricted to computational assets, with few contribution to core linguistic outputs such as orthographies, archival practices, or educational policy. Conversely, traditional documentation (T1) remains heavily focused on human-readable formats, specifically static archives (n=10) and community revitalisation manuals (n=10). This divergence confirms a significant format barrier: despite both fields generating data, traditional archival outputs are not being converted into the machine-readable structures essential for modern digital multilingualism.

Figure 1: Output types by thematic block (n = 105 included records).

Figure 2 displays the keyword co-occurrence network, which shows a clear divide between the two research domains. The network shows strong internal clustering among computational concepts, notably 'corpus' (n=35), 'minority langs.' (n=33) and 'low-resource' (n=32), and similar cohesion among educational terms such as 'documentation' (n=18) and 'multilingualism' (n=15). However, the connections between the two clusters are weak. Fieldwork concepts such as 'dialects' (n=9) and 'community' (n=8) rarely co-occur with computational nodes such as 'NLP' (n=6) or 'ASR' (n=6). The network thus visualises a parallel reality: both fields increasingly prioritise endangered languages, but they work in separate methodological ecosystems, without the conceptual overlap needed to turn documented archives into computable resources.

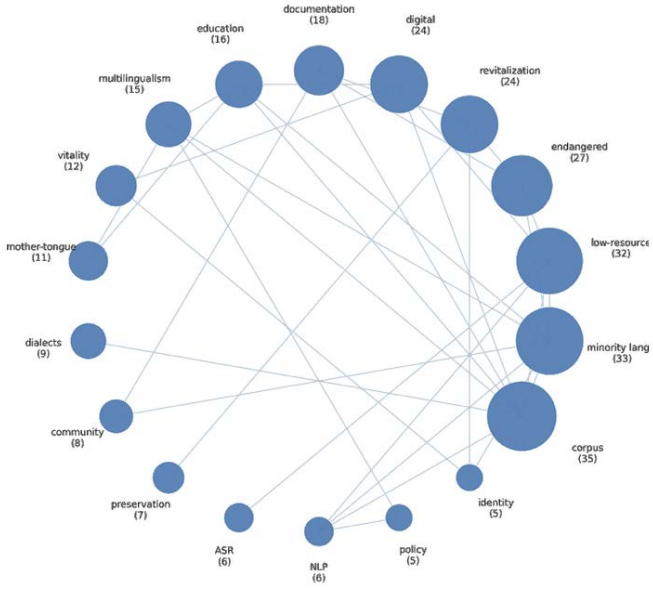

Figure 2: Co-occurrence network of canonical keywords.

Figure 3 traces the chronological distribution of the literature over the past two decades. Prior to 2015, publication volume remained low and was dominated almost entirely by language documentation (T1) and multilingual education (T2). From 2021 onwards the volume rises sharply, and the composition changes with it: studies in NLP and low-resource computation (T4) first appear in 2020 and expand rapidly, accounting for half of all records in 2024, while digital vitality work (T3) appears steadily from 2015. This growth shows that the computational community has begun to turn toward endangered and low-resource languages, the very subject matter of the documentation tradition. Yet traditional documentation (T1) does not decline during this turn; it reaches its own peak in the same period. The hotspot is shared in time only: both bodies of work have grown in the same years, yet Figure 1 shows that their outputs remain separate, and Figure 2 demonstrates that their vocabularies rarely co-occur. As both bodies of work continue to grow separately, the need to convert documentary records into machine-readable resources grows with them.

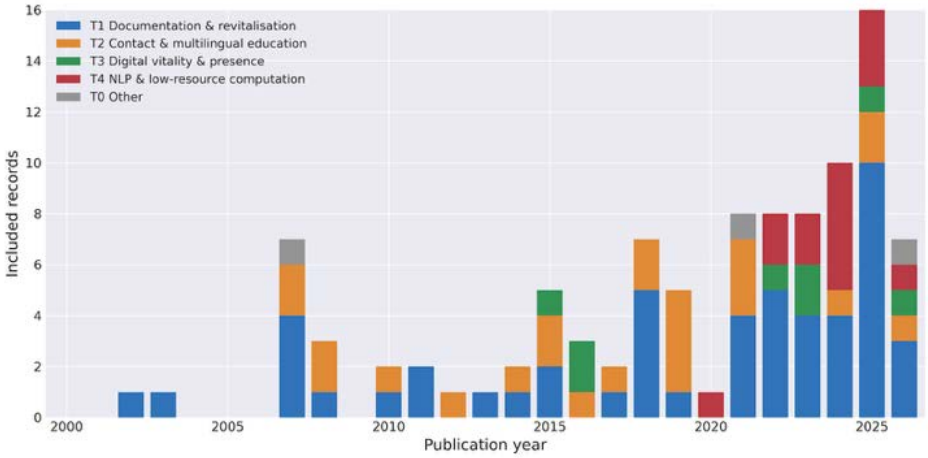

Figure 3: Yearly distribution of included records by thematic block.

The quantitative findings illustrate a clear divergence in research outputs. This confirms that the disconnect between documenting languages and processing them computationally is a structural issue. To address the typological format barrier and to summarise beneficial practices in corpus conversion, the study proposes a three-step framework comprising extraction, tiered annotation, and integration. Crucially, to avoid extractive research practices, this proposed pipeline treats community consent as embedded governance rather than a secondary addition (Bird, 2020; Liu et al., 2022).

Step 1: Data Extraction. Rather than defaulting to optical character recognition (OCR), a computational process that attempts to extract text from scanned images, extraction strategies might first prioritise 'born-digital' documentary deposits. These are files created directly within software environments where source texts, grammatical glosses, and translations are already explicitly aligned (Track A). Print-only legacy materials often require OCR and script normalisation (Track B), but this approach is generally reserved as an error-prone fallback. By harvesting pre-structured digital files first, researchers can systematically navigate the typological barrier. This approach is supported by the archive-scale republication demonstrated by the Pangloss Collection (Adamou et al., 2025).

Step 2: Tiered Annotation. Because full syntactic treebanks, which are complex databases mapping the complete grammatical structure of sentences, often end up being impractical for 'zero-resource' languages currently lacking digital data, annotation strategies could be designed to be cumulative and cost-effective. A foundational 'Bronze' tier includes parallel text (sentences matched alongside their translations) and digitised lexica. This basic level can often be extracted almost automatically from existing documentary glosses. Subsequent 'Silver' tiers for word formation (morphology) and 'Gold' tiers for sentence structure (syntax) could remain optional and would depend on native-speaker validation. This tiered logic aligns with observations on resource constraints and suggests that the feasibility demonstrated by Thai dialect corpora (Suwanbandit et al., 2023; Thatphithakkul et al., 2024) might be extended to minority languages at a manageable cost.

Step 3: Integration and Dissemination. Finally, conversion efforts benefit from being piloted alongside active community partnerships, such as those established within the Mahidol mother-tongue education networks (Premsrirat & Hirsh, 2018), prior to any regional scaling. The resulting corpora could then be deposited into existing platforms like SEACrowd (Lovenia et al., 2024) under graduated licences, which allow the community to select varying levels of access and permission for their data. Similar to the AmericasNLP initiative (Mager et al., 2021), this integration stage introduces legacy data into modern shared computational tasks. Ultimately, achieving even a foundational Bronze tier produces a machine-readable parallel corpus, which serves as a minimal increment to help transition a language from being computationally invisible to becoming a measurable entity within digital vitality frameworks (Kornai, 2013; Simons et al., 2022).

Conclusion and Discussion

This study set out to investigate why Southeast Asia's thoroughly documented minority languages remain largely absent from modern digital platforms. By conducting a systematic literature analysis, the research directly addressed its first objective, which was to examine the methodological gap between traditional language documentation and modern computational linguistics. The quantitative results demonstrate a clear structural disconnect. Traditional archiving and modern natural language processing operate in parallel but entirely isolated spheres. This confirms that the underlying issue is not a lack of linguistic data. Rather, it is a typological format barrier that causes a failure in data mobilisation, leaving rich historical archives trapped in human-readable formats that computers cannot process.

To address the second objective, the study proposed a three-step corpus-conversion framework comprising data extraction, tiered annotation, and community integration. The primary innovation of this research lies in shifting the academic focus away from collecting new field data. Instead, it offers a structured, post-hoc pathway to restructure existing legacy archives into machine-readable formats. By breaking the conversion process into affordable, cumulative tiers (starting at a foundational 'Bronze' level) and embedding community consent directly into the workflow as a core governance mechanism, the framework offers a realistic and ethical solution for zero-resource languages.

Whilst this study provides a framework, it is important to acknowledge certain boundaries. The systematic review focused primarily on Thailand, meaning the specific format barriers identified might differ slightly in regions with alternative archiving histories. Furthermore, the proposed corpus-conversion pipeline is currently a theoretical model derived from secondary literature. Although it synthesises established methods from independent projects, the complete workflow remains to be tested in practice. Therefore, the immediate implication for future research is the necessity of an empirical pilot study. Researchers should prioritise testing this conversion pipeline alongside active community partnerships to assess its practical feasibility and cost-effectiveness. Finally, these findings suggest a broader shift in methodology for linguists: future language documentation projects should incorporate computational formatting and machine readability from their very inception. Many regioanl minority languages have been documented for decades, yet they remain computationally

invisible. Converting their records into machine-readable resources is therefore no longer just a technical exercise; it is the precondition for these languages to survive and take part in global digital multilingualism.

Acknowledgement:

During the preparation of this work, the author used Gemini 3.6 Flash (Google DeepMind) and Kimi K3 (Moonshot AI) to assist with literature searches, language polishing, and to improve the readability of the figures. After using these tools, the author reviewed and edited the content as needed and takes full responsibility for the content of the publication.

References

Adamou, E., Guillaume, S., & Michaud, A. (2025). The Pangloss Collection: Opening up research data on endangered and underdocumented languages. *Language*, *101*(1), e38–e59. https://doi.org/10.1353/lan.2025.a954237

Adebara, I. (2025). *AI and language data flaring in Africa: Addressing the low-resource challenge* (Policy Brief No. 216). Centre for International Governance Innovation. https://www.cigionline.org/publications/ai-and-language-data-flaring-in-africa-addressing-the-low-resource-challenge/

Bird, S. (2020). Decolonising speech and language technology. In D. Scott, N. Bel, & C. Zong (Eds.), *Proceedings of the 28th International Conference on Computational Linguistics* (pp. 3504–3519). International Committee on Computational Linguistics. https://doi.org/10.18653/v1/2020.coling-main.313

Geng, Y., Xu, J., Liang, Z., Yang, J., Shi, X., & Shen, X. (2025). Empowering global voices: A data-efficient, phoneme-tone adaptive approach to high-fidelity speech synthesis. *arXiv preprint arXiv:2504.07858*. doi.org

Hämäläinen, M. (2021). Endangered languages are not low-resourced! In M. Hämäläinen, N. Partanen, & K. Alnajjar (Eds.), *Multilingual facilitation* (pp. 1–11). University of Helsinki. https://doi.org/10.31885/9789515150257

Himmelmann, N. P. (2006). Language documentation: What is it and what is it good for? In J. Gippert, N. P. Himmelmann, & U. Mosel (Eds.), *Essentials of language documentation* (pp. 1–30). Mouton de Gruyter. https://linggwistiks.wordpress.com/wp-content/uploads/2009/04/essentials-of-language-documentation.pdf

Kornai, A. (2013). Digital language death. *PLoS ONE*, *8*(10), Article e77056. https://doi.org/10.1371/journal.pone.0077056

Liu, Z., Richardson, C., Hatcher, R., & Prud'hommeaux, E. (2022). Not always about you: Prioritizing community needs when developing endangered language technology. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 3933–3944). Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.acl-long.272

Lovenia, H., Mahendra, R., Akbar, S. M., Miranda, L. J. V., Santoso, J., Aco, E., Fadhilah, A., Mansurov, J., Imperial, J. M., Kampman, O. P., Moniz, J. R. A., Habibi, M. R. S., Hudi, F., Montalan, R., Ignatius, R., Lopo, J. A., Nixon, William., Karlsson, B. F., Jaya, J., . . . Cahyawijaya, S. (2024). SEACrowd: A multilingual multimodal data hub and benchmark suite for Southeast Asian languages. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing* (pp. 5155–5203). Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.emnlp-main.296

Mager, M., Oncevay, A., Ebrahimi, A., Ortega, J., Rios, A., Fan, A., Gutierrez-Vasques, X., Chiruzzo, L., Giménez-Lugo, G., Ramos, R., Meza Ruiz, I. V., Coto-Solano, R., Palmer, A., Mager-Hois, E., Chaudhary, V., Neubig, G., Vu, N. T., & Kann, K. (2021). Findings of the AmericasNLP 2021 shared task on open machine translation for indigenous languages of the Americas. In *Proceedings of the First Workshop on Natural Language Processing for Indigenous Languages of the Americas* (pp. 202–217). Association for Computational Linguistics.

Moshagen, S. N., Pirinen, T. A., & Trosterud, T. (2013). Building an open-source development infrastructure for language technology projects. In *Proceedings of the 19th Nordic Conference of Computational Linguistics (NoDaLiDa 2013)* (NEALT Proceedings Series 16, pp. 343–352). Linköping University Electronic Press.

Premsrirat, S. (2007). Endangered languages of Thailand. *International Journal of the Sociology of Language, 2007*(186), 75–93. https://doi.org/10.1515/IJSL.2007.043

Premsrirat, S., & Hirsh, D. (Eds.). (2018). *Language revitalization: Insights from Thailand.* Peter Lang.

Seifart, F. (2006). Orthography development. In J. Gippert, N. P. Himmelmann, & U. Mosel (Eds.), *Essentials of language documentation* (pp. 275–299). Mouton de Gruyter.

Simons, G. F., Thomas, A. L. L., & White, C. K. K. (2022). Assessing digital language support on a global scale. In *Proceedings of the 29th International Conference on Computational Linguistics* (pp. 4299–4305). International Committee on Computational Linguistics. https://aclanthology.org/2022.coling-1.379

Smalley, W. A. (Ed.). (1976). *Phonemes and orthography: Language planning in ten minority languages of Thailand* (Pacific Linguistics, Series C, No. 43). Department of Linguistics, Research School of Pacific Studies, Australian National University.

Suwanbandit, A., Naowarat, B., Sangpetch, O., & Chuangsuwanich, E. (2023). Thai dialect corpus and transfer-based curriculum learning investigation for dialect automatic speech recognition. In *Proceedings of Interspeech 2023* (pp. 4069–4073). International Speech Communication Association. https://doi.org/10.21437/Interspeech.2023-1828

Thatphithakkul, S., Thangthai, K., & Chunwijitra, V. (2024). The development of LOTUS-TRD: A Thai regional dialect speech corpus. In *2024 27th Conference of the Oriental COCOSDA International Committee for the Co-ordination and Standardisation of Speech Databases and Assessment Techniques (O-COCOSDA)* (pp. 1–6). IEEE. https://doi.org/10.1109/O-COCOSDA64382.2024.10800335

Trilsbeek, P., & Wittenburg, P. (2006). Archiving challenges. In J. Gippert, N. P. Himmelmann, & U. Mosel (Eds.), *Essentials of language documentation* (pp. 311–335). Mouton de Gruyter. https://doi.org/10.1515/9783110197730.311

Ungsitipoonporn, S., Watyam, B., Ferreira, V., & Seyfeddinipur, M. (2021). Community archiving of ethnic groups in Thailand. *Language Documentation & Conservation*, *15*, 267–284. http://hdl.handle.net/10125/24975

Zariquiey, R., Oncevay, A., & Vera, J. (2022). CLD² language documentation meets natural language processing for revitalising endangered languages. In *Proceedings of the Fifth Workshop on the Use of Computational Methods in the Study of Endangered Languages* (pp. 20–30). Association for Computational Linguistics. https://aclanthology.org/2022.computel-1.4/