Using ASR to put the interactive back into the corpus of Thulung (Kiranti, Eastern Nepal)

In this presentation, we describe a project in which we have developed and used an ASR model for Thulung, an endangered language of Nepal (Kiranti subgroup, Sino-Tibetan/Trans-Himalayan, Eastern Nepal), with the goal of balancing out a corpus with a strong monological orientation.

The existing annotated corpus of Thulung before the start of the project was predominantly made up of single speaker narratives, collected and analyzed over the course of on- and off fieldwork on the language over the past 20 years. More specifically, the 9-hour corpus contained 36 single-speaker narratives and only 4 conversations. The reasons for this imbalance were a combination of the methodological (difficulties in capturing interactions, owing to their spontaneity), cultural (speakers would much rather be recorded telling their traditional stories, and are uncomfortable giving consent to put their conversations into accessible oral archives), and annotational (e.g. Dingemanse & Liesenfeld 2022; Austin 2021).

An interest in documenting interactives (c.f. Heine 2023) such as interjections (Ameka 1992; Dingemanse 2024) and ideophones (Lahaussois 2023; Lahaussois 2024) has brought to the fore the importance of working with a corpus containing considerably more interactive communicative events. Yet the difficulties in transcribing sufficient such materials have made us consider the possibilities of using ASR to attempt to push through the transcription bottleneck (e.g. Seifart et al. 2018). We fine-tuned XLSR-53, a multilingual pre-trained speech model capable of representing audio recordings in any language as vector embedding, which we had previously successfully used with related minority low-resource languages Na and Japhug (Guillaume et al. 2022; Wisniewski, Michaud & Guillaume 2020; Guillaume, Wisniewski & Michaud 2023). One significant difference was that the materials in the former case studies were largely monological.

We present here the results of our ongoing project, and specifically, the workflow we have adopted to use and improve ASR in the field. This has enabled us to begin to rebalance the corpus—currently consisting of 66 texts of various lengths and type, of which 16, constituting roughly 3 hours, are conversational. We will present quantitative data for our post-editing process, according to communicative event type, and describe some of the difficulties in carrying out this type of work and some promising solutions.

References

Ameka, Felix K. 1992. Interjections: The universal yet neglected part of speech. *Journal of Pragmatics* 18. 101–118.

Austin, Peter K. 2021. Language documentation and language revitalization. In Justyna Olko & Julia Sallabank (eds.), *Revitalizing endangered languages: a practical guide*, 199–212. Cambridge: Cambridge University Press.

Dingemanse, Mark. 2024. Interjections at the Heart of Language. *Annual Review of Linguistics* 10. 257–77.

Dingemanse, Mark & Andreas Liesenfeld. 2022. From text to talk: Harnessing conversational corpora for humane and diversity-aware language technology. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, vol. 1, 5614–5633. Association for Computational Linguistics.

Guillaume, Séverine, Guillaume Wisniewski, Benjamin Galliot, Minh-Châu Nguyên, Maxime Fily, Guillaume Jacques & Alexis Michaud. 2022. Plugging a neural phoneme recognizer into a simple language model: a workflow for low-resource settings. In *Proc. Interspeech 2022*, 4905–4909. Incheon, Korea: ISCA. https://doi.org/10.21437/Interspeech.2022-11314.

Guillaume, Séverine, Guillaume Wisniewski & Alexis Michaud. 2023. From 'Snippet-lects' to Doculects and Dialects: Leveraging Neural Representations of Speech for Placing Audio Signals in a Language Landscape. In *Proc. 2nd Annual Meeting of the ELRA/ISCA SIG on Under-resourced Languages*, 29–33. SIGUL. https://doi.org/10.21437/SIGUL.2023-7.

Heine, Bernd. 2023. *The grammar of interactives*. Oxford: Oxford University Press.

Lahaussois, Aimée. 2023. Ideophonic patterns in Kiranti languages and beyond. *Folia Linguistica* 57(1). 1–36. https://doi.org/10.1515/flin-2022-2053.

Lahaussois, Aimée. 2024. Terminological diversity in descriptions of Kiranti ideophonic lexemes. *Linguistic Typology at the Crossroads* 4(1). 14–43.

Seifart, Frank, Nicholas Evans, Harald Hammarström & Stephen C. Levinson. 2018. Language documentation twenty-five years on. *Language* 94(4). 324–345.

Wisniewski, Guillaume, Alexis Michaud & Séverine Guillaume. 2020. Phonemic transcription of low-resource languages: To what extent can preprocessing be automated? In *Proceedings of the 1st Joint SLTU and CCURL Workshop (SLTU-CCURL 2020)*, 306–315. European Language Resources Association (ELRA).