Rapidly building the missing infrastructure for language science: A case study with Formosan languages

Despite repeated calls for a more comprehensive study of human language, the vast majority of psycholinguistic and language acquisition studies focus a small number of languages (Collart, 2013; Kidd & Garcia, 2022). *One* significant roadblock is the lack of corpora, without which most computational and experimental methods are impossible.

Or so it would seem. In fact, we have discovered a surprising range of languages that lack research corpora but have substantial *latent* corpora: texts and recordings in sufficient quantity to support language science if only they were accessible.

As a case study, we present FormosanBank, a rapidly-growing free-and-open-source meta-corpus of a theoretically-critical family of 16 endangered indigenous languages (Li et al., 2024). As of 2023, there were no machine-readable corpora of any Formosan language. By the middle of 2025, we anticipate corpora of at least 1,000,000 words in each language — comparable to the Francis & Kucera corpus, the basis of much foundational research in English — and at least 10 hours of transcribed speech. We have already reached this threshold in three languages (Figure 1).

We highlight three ingredients to this work: 1) leveraging partnerships with researchers, indigenous groups, and government agencies; 2) marshalling a cross-disciplinary research team; 3) utilizing machine learning to dramatically speed up digitization and transcription. We show how this strategy has allowed us to rapidly identify and obtain existing resources, create new ones, standardize formats, and maintain excellent quality control. We supplement our discussion of FormosanBank with additional examples from our work on Native American and Australian languages.

References

Collart, A. (2013). Ten years of linguistic diversity in language processing conferences. *AMLaP*.

Eberhard, D. M., Simons, G. F., & Fennig, C. D. (2022). *Ethnologue: Languages of the world* (Vol. 22).

Kidd, E., & Garcia, R. (2022). How diverse is child language acquisition research? *First Language*, *42*(6), 703–735.

Li, P. J.-k., Zeitoun, E., & De Busser, R. (2024). *Handbook of formosan languages (3 parts): The indigenous languages of taiwan*. Brill. https://books.google.com/books?id=0Q8M0AEACAAJ

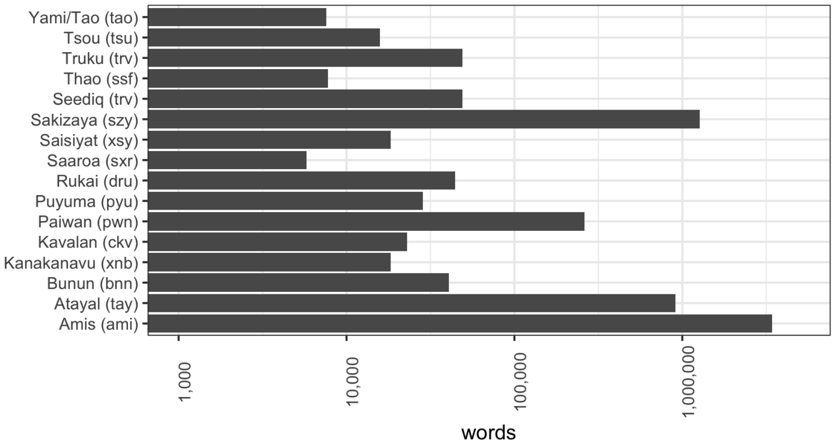

Figure 1: Total words scraped per language to date. Excludes corpora for which we do not yet have republication rights or for which we have rights but have not yet completed scraped.

Language	Dialects	Status	Speakers
Amis (ami)	5	6b (Threatened)	108,000
Atayal (tay)	6	7 (Shifting)	10,000
Bunun (bnn)	5	5 (Developing)	38,000
Kanakanavu (xnb)	1	8b (Nearly Extinct)	4
Kavalan (ckv)	1	8b (Nearly Extinct)	70
Paiwan (pwn)	4	6b (Threatened)	15,000
Puyuma (pyu)	4	8a (Moribund)	1,000
Rukai (dru)	6	6b (Threatened)	2,000
Saaroa (sxr)	1	8b (Nearly Extinct)	25
Saisiyat (xsy)	1	7 (Shifting)	2,000
Sakizaya (szy)	1	7 (Shifting)	590
Seediq (trv)	2	8a (Moribund)	650
Thao (ssf)	1	8b (Nearly Extinct)	4
Truku (trv)	1	8a (Moribund)	650
Tsou (tsu)	1	6b (Threatened)	4,000
Yami/Tao (tao)	1	6b (Threatened)	3,800

Table 1: Language status and speaker population, based on Ethnologue (Eberhard et al., 2022).