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

Joshua K. Hartshorne[1], Emily Prud'hommeaux[2],
Li-May Sung[3], Éric Le Ferrand[2]

[1] Communication Sciences and Disorders, MGH IHP
[2] Computer Science Department, Boston College
[3] Linguistics Department, National University of Taiwan

Introduction

Goal:

Facilitate psycholinguistic and acquisition studies of the world's languages while it's still possible, starting with the 16 endangered Formosan languages of Taiwan.

Background:

- The vast majority of psycholinguistic and language acquisition studies focus a small number of languages (Collart, 2023; Kidd & Garcia, 2022).
- Major roadblock: Lack of corpora (for computational analysis, word frequencies, surprisal, etc.)
- ~50% of languages are already gone and as many as 90% would be by the end of the century

Why Formosan Languages?

- Cover every major branch of Austronesian family, one of largest in world.
- Formosan languages challenge existing theory (voice system, no clear parts of speech, etc.)
- Preexisting standardized written form, reference grammars, dictionaries, large "latent" corpus.

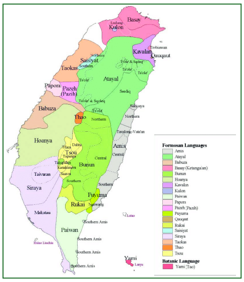


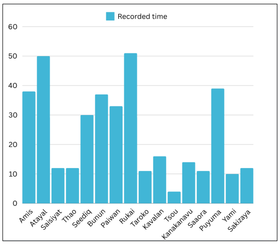
Fig 1: Historical distribution of Formosan languages

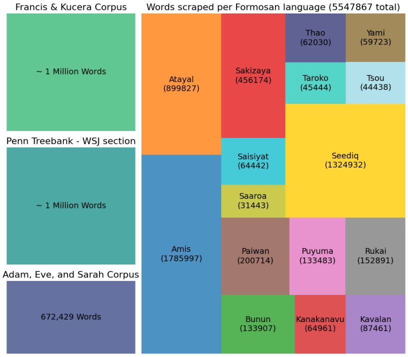
Language	Dialects	Status	Speakers
Amis (ami)	5	6b Threatened	108,000
Atayal (tay)	6	7 (Shifting)	10,000
Bunun (bnn)	5	5 (Developing)	38,000
Kanakanavu (xnb)	1	8b (Nearly Extinct)	4
Kavalan (ckv)	1	8b (Nearly Extinct)	70
Paiwan (pwn)	4	6b (Threatened)	15,000
Puyuma (pyu)	4	8a (Moribund)	1,000
Rukai (dru)	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 (ssf)	1	6b (Threatened)	3,800

Table 1: Language status and speaker population, based on Ethnologue (Eberhard et al., 2022). NOTE: Yami/Tao is not "linguistically" Formosan.

Current Status

Text & Audio processed & permission obtained. NOTE: Quality Control is ongoing…





References

- Collart, A. (2024). A decade of language processing research: Which place for linguistic diversity?. Glossa Psycholinguistics, 3(1)
- Kidd, E., & Garcia, R. (2022). How diverse is child language acquisition research? First Lang, 42, 703–735.
- Eberhard, D. M., Simons, G. F., & Fennig, C. D. (2022). Ethnologue: Languages of the world (Vol. 22).
- Michailovsky, B., Mazaudon, M., Michaud, A., Guillaume, S., François, A., & Adamou, E. (2014). Documenting and researching endangered languages: the Pangloss Collection.
- Prud'hommeaux, E., Jimerson, R., Hatcher, R., & Michelson, K. (2021). Automatic speech recognition for supporting endangered language documentation.

Acknowledgments

This work was supported by NSF under Grant No. 2319296. We are grateful for their generous support. We also want to thank our consultants who made this research possible, Yuyang Liu, Xuan Ruan, Ûi-iû Kán, Chuan-Jie Lin, Damián Blasi, Severine Guillaume, Isabelle Bril, and especially the Indigenous Languages Research and Development Foundation of Taiwan.

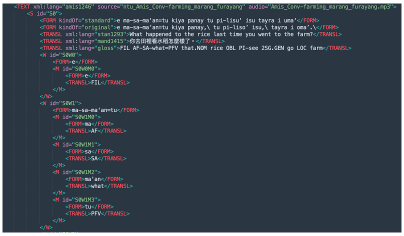
Method

Data Collection:

- Leveraging partnerships with researchers, indigenous groups, and government agencies
- Processing & reformatting latent corpus:
  - Published corpora
  - Indigenous YouTube
  - Dictionaries (with example sentences)
  - Instructional materials
  - Wikipedias
  - Radio & TV transcripts
- Obtaining permission for republication

Data Processing and QC:

- **Initial processing:** Hand-verified OCR if necessary; Remove extraneous text; Alignment of translations, audio, if any; Convert to modified Pangloss format.



```
<TEXT xml:lang="amis1246" source="ntu_Amis_Conv-farming_marang_furayang" audio="Amis_Conv-farming_marang_furayang.mp3">
    <S id="S0">
        <FORM kindOf="standard">e ma-sa-ma'an=tu kiya panay tu pi-lisu' isu tayra i uma'</FORM>
        <FORM kindOf="original">e ma-sa-ma'an=tu kiya panay,\ tu pi-liso' isu,\ tayra i oma'.\</FORM>
        <TRANSL xml:lang="stan1293">What happened to the rice last time you went to the farm?</TRANSL>
        <TRANSL xml:lang="mand1415">你去田裡看水稻怎麼樣了•</TRANSL>
        <TRANSL xml:lang="gloss">FIL AF-SA-what=PFV that.NOM rice OBL PI-see 2SG.GEN go LOC farm</TRANSL>
        <W id="S0W0">
            <FORM>e</FORM>
            <M id="S0W0M0">
                <FORM>e</FORM>
                <TRANSL>FIL</TRANSL>
            </M>
        </W>
        <W id="S0W1">
            <FORM>ma-sa-ma'an=tu</FORM>
            <M id="S0W1M0">
                <FORM>ma</FORM>
                <TRANSL>AF</TRANSL>
            </M>
            <M id="S0W1M1">
                <FORM>sa</FORM>
                <TRANSL>SA</TRANSL>
            </M>
            <M id="S0W1M2">
                <FORM>ma'an</FORM>
                <TRANSL>what</TRANSL>
            </M>
            <M id="S0W1M3">
                <FORM>tu</FORM>
                <TRANSL>PFV</TRANSL>
            </M>
        </W>
```

- **Automated Validation:** XML, orthography, frequent words, translations, etc.
- **Manual Quality Control:** Manually review flagged segments and random samples.

Next Steps

Data Priorities:

- Finish automating Quality Control
- Incorporate more glossed corpora
- Obtain rights for
  - Indigenous YouTube
  - Radio & TV
- Obtain rights for more glossed corpora

Publish v.1

Bootstrapping the Corpora:

- Automatic Speech Recognition for transcription (Prud'hommeaux et al., 2021)
- Apply to ongoing Paiwan data collection

Use the Corpora!

- Machine Translation (requested by indigenous partners)
- Classifiers for finding unusual syntactic patterns (requested by colleague)
- Comparison of voice system across languages (using parallel corpora)

Next Next Steps

- Design psycholinguistic studies
- Collect child-directed speech (limited # of languages)