FormosanBank and why you should use it

Despite repeated calls for a more comprehensive study of human language, the vast majority of studies — including at AMLaP — focus a small number of languages (Collart, 2013). A significant roadblock is the lack of intellectual infrastructure: dictionaries, reference grammars, and most of all, corpora. A lack of corpora not only impedes quantitative analysis and testing computational models, but also the deployment of any experimental paradigm that depends on word frequencies, collocations, surprisal, or other corpus-derived statistics — in other words, nearly anything that would be presented at AMLaP.

Here, we present FormosanBank, a rapidly-growing free-and-open-source meta-corpus of a theoretically-critical family of languages: the indigenous Formosan languages of Taiwan. We present FormosanBank both as a resource for researchers interested in advancing (psycho)linguistic theory but also as a model for similar projects addressing other languages.

Why Investigate Formosan? The Formosan languages — which are unrelated to Chinese — began diverging 5,000 years ago and comprise most of the typological variability in the Austronesian language family, itself one of the world's largest language families by both number of languages and speakers (Li et al., 2024). Formosan languages exhibit a number of typologically unusual patterns that continue to flummox psycholinguistic theory (see supplementary page on Formosan languages). The Formosan languages themselves are diverse, allowing for theoretically-informative comparisons across languages. Unfortunately, all Formosan languages are endangered, so the window of opportunity for investigation is rapidly closing.

Unlike many understudied languages critical to theoretical debates, the Formosan languages are spoken in a country (Taiwan) with ubiquitous rapid transportation, excellent food, low crime, and a large international airport. There is also a robust language science community, which over the last century has compiled reference grammars, dictionaries, and quite a few corpora. Unfortunately, corpora vary in orthography and hardly any are in a machine-readable format. Thus we are (with permission) compile corpora into a common, machine-readable, and publicly-accessible format. Between utilizing existing resources and creating new ones, we aim for >1,000,000 words per language (comparable to the Francis & Kucera corpus, the basis of much classic research) and 10 hours of transcribed speech.

Results & Discussion. Six languages have already reached the threshold for transcribed speech. Total word count is more variable, with three languages at or near criterion (Fig. 1). We outline critical low-hanging fruit for the AMLaP community using the now-available resources. We describe how to extend to other languages.

More about the Formosan languages

The Taiwanese government recognizes 16 extant languages, including "Tao" (alt., "Yami"), which is linguistically considered part of the Malayo-Polynesian subgroup of Austronesian. Formosan languages exhibit a range of theoretically-interesting phenomena. They utilize an unusually restricted number of parts of speech, and indeed the existence of parts of speech is controversial. The phonemic inventory is also strikingly small, with many languages having only 3 vowels (Li et al., 2024). All Formosan languages are verb-initial, and in some the object precedes the subject; a few also allow SVO, probably due to influence from Chinese languages (Li, 2008). Verb-initial and subject-final languages are both rare: 9% and 3%, respectively, of those surveyed by WALS (Dryer, 2013). Formosan languages also make extensive use of reduplication for grammatical purposes. In Thao, for instance, reduplication is used to modify verbal aspect, change adjective intensity, and create instrumental nouns, among other purposes (Chang, 1998). Formosan languages also make use of prefixes, suffixes, infixes, and circumfixes, which in at least some cases have all been attested in the same language (Li et al., 2024).

Most (in)famously, most Formosan languages (along with some closely-related Austronesian languages) exhibit voice (sometimes called "focus" or "topic") system. The exact theoretical characterization remains controversial (Li et al., 2024), but phenomenologically the semantic role of the subject of a verb depends on the verbal affixes. For instance, in Tsou the subject is either the agent, the patient, the instrument, or the beneficiary (examples 1-4). Critically note that the verb is complex in all four voices; there is no "base" form, though the most common appears to be patient voice. Formosan languages vary in how voice is instantiated (Li et al., 2024).

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).

(1) mo **t-m-eaphʉ** to oko ta skayʉ si ino.
AV.RLS put-AV OBL child OBL cradle NOM mother.
Mother put the child into a cradle.

(2) i-si **teaph-a** ta skayʉ to ino to oko.
UV.RLS-3S.GEN put-PV OBL cradle OBL mother NOM child.
Mother put the child into a cradle.

(3) i-si **teaph-i** to oko ta ino ta skayʉ.
UV.RLS-3S.GEN put-IV OBL child OBL mother NOM cradle.
Mother put the child into the cradle.

(4) i-si **teaph-neni** to tacʉmʉ to ino ’e oko.
AV.RLS-3S.GEN put-BV OBL banana OBL mother NOM child.
Mother put bananas (in a cradle) for the child.

Zeitoun, 2005

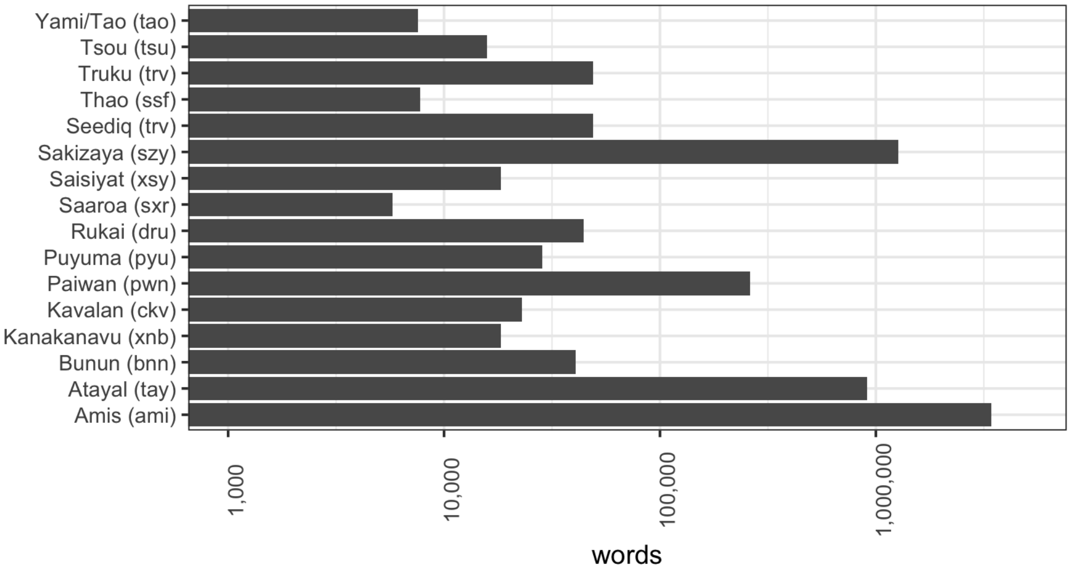

Figure 1: Total words scraped per language to date. Excludes corpora for which we do not yet have republication rights or for which we have rights but have not yet completed scraped.

Chang, M. L. (1998). Thao reduplication. Oceanic Linguistics, 277–297. **Collart, A.** (2013). Ten years of linguistic diversity in language processing conferences. AMLaP. **Dryer, M. S.** (2013). Order of subject, object and verb (v2020.3). In M. S. Dryer & M. Haspelmath (Eds.), The world atlas of language structures online. **Eberhard, D. M., Simons, G. F., & Fennig, C. D.** (2022). Ethnologue: Languages of the world (Vol. 22). **Li, P. J.-k.** (2008). The great diversity of Formosan languages. Language and Linguistics, 9(3), 523–546. **Li, P. J.-k., Zeitoun, E., & De Busser, R.** (2024). Handbook of formosan languages (3 parts): The indigenous languages of Taiwan. Brill. **Zeitoun, E.** (2005). Tsou. In K. A. Adelaar & N. Himmelmann (Eds.), The Austronesian languages of Asia and Madagascar. Routledge.