The Syllable Inventory of Basay

A Quantitative Analysis and Typological Assessment

Tsai Yung-kuei (蔡永桂)

June 2026

basay.tw/research/2026-06-basay-syllable/

The Syllable Inventory of Basay: A Quantitative Analysis and Typological Assessment

Abstract

This paper presents a quantitative extraction of the syllable inventory of Basay, an Austronesian language formerly spoken in northern Taiwan by the Basay, a plains indigenous (Pingpu) people. Drawing on a lexical database of 3,324 entries and excluding forms reconstructed for Proto-Austronesian (PAN), we identified 486 distinct syllable types (occurring at least twice) from 2,364 entries. Basay exhibits a typologically notable phonological profile, including retroflex lateral /ɭ/, voiced alveolar lateral fricative /ɮ/, velar nasal /ŋ/, palato-alveolar fricative /ʃ/, and complex onset clusters. With 486 syllable types, Basay substantially exceeds the inventories of Hawaiian (ca. 60) and Japanese (ca. 100), and surpasses Mandarin Chinese (ca. 400, excluding tone), placing it in the mid-to-upper range of the cross-linguistic spectrum. These findings confirm Basay as phonologically complex among Formosan languages and contribute to the documentation of an endangered—now extinct—Austronesian language.

Keywords: Basay, Pingpu, syllable inventory, phonological typology, Austronesian, endangered language

---

1. Introduction

Basay (also Basai) is an Austronesian language spoken by the Basay people, a plains indigenous group who inhabited the Yilan Plain and the northeastern rim of the Taipei Basin in northern Taiwan. Fragmentary vocabulary is attested in Dutch colonial

records of the seventeenth century and in Qing dynasty sources; however, the language is considered to have lost its last native speakers by the early twentieth century (Li 1996, 2000). Basay is currently classified as an extinct language, and documentation and revitalization efforts are ongoing at the Institute of Linguistics, Academia Sinica, Taipei.

The present study has two objectives. First, we extract and describe the syllable inventory of Basay from available lexical data, providing explicit correspondences between the orthographic conventions and the International Phonetic Alphabet (IPA). Second, we situate the resulting inventory within a cross-linguistic typological framework by comparing it with representative languages from around the world.

The quantitative extraction method employed here is not intended to replace a grammar-based phonological analysis. Rather, it aims to provide a distributional description of syllables as observed in attested lexical material—a contribution that is particularly valuable for languages with limited documentation.

---

2. Data and Methods

2.1 Source Data

The analysis is based on the Basay lexical database (`basay_dict.jsonl`; 3,324 entries). Each entry includes the Basay orthographic form, translations into Chinese, Japanese, and English, and a source code. Source codes are as follows: `B` (native Basay vocabulary), `T` (Taiwanese loanwords), `M` (Min Nan/Hokkien loanwords), `S`, `V`, and `PAN` (Proto-Austronesian reconstructions).

Entries coded as `PAN` (960 entries) were excluded from the analysis. PAN reconstructions represent comparative-linguistic abstractions rather than attested modern Basay forms, and their inclusion would distort the inventory of actually recorded vocabulary. The remaining 2,364 entries formed the analytical corpus.

2.2 Orthographic Conventions

Basay is transcribed in a Latin-based orthography. Table 1 presents the correspondences between orthographic symbols and IPA representations.

Table 1. Basay Orthography–IPA Correspondence

Orthography	IPA	Description
`n'`	ŋ	Velar nasal
`s'`	ʃ	Palato-alveolar fricative
`l'`	ɭ	Retroflex lateral approximant
`z'`	ɬ	Voiced alveolar lateral fricative
`o'`	ə	Mid central vowel (schwa)
`'` (coda)	ʔ	Glottal stop (syllable-final position)
`q`	q	Uvular/pharyngeal stop (tentative)
`ts`	ts	Alveolar affricate
`ts'`	tʃ	Palato-alveolar affricate
`j`	j ~ dʒ	Approximant or affricate (context-dependent)
`v`	v	Voiced labiodental fricative

The glottal stop (`'`) functions as a coda consonant belonging to the preceding syllable and does not constitute an independent syllable onset. Accordingly, forms beginning with `'` (e.g., `'a`, `'ul`) were identified as transcription artifacts of a syllable-final glottal stop and excluded from the syllable inventory.

2.3 Syllable Extraction Procedure

Syllables were extracted using the following procedure:

1. Each Basay entry was cleaned by removing annotations, parenthetical material, and alternative forms; only the primary form was retained.
2. A syllabification algorithm based on the template (C*)V(V?)(C?) was applied, treating vowel letters (a, e, i, o, u, and special vowel characters) as syllable nuclei. A following consonant was assigned to the onset of the next syllable if that syllable began with a vowel.
3. Forms beginning with spaces, punctuation, numerals, hyphens, or the apostrophe were removed as noise.
4. Syllables with a frequency of one were excluded to reduce the influence of transcription variation and loan-word artifacts.
5. The remaining syllables (frequency ≥ 2) constitute the final inventory.

---

3. The Syllable Inventory

3.1 Overall Statistics

The procedure yielded 486 distinct syllable types with a frequency of two or more. The frequency distribution is summarized in Table 2.

Table 2. Syllable Types by Frequency Band

Frequency Band	Types	Percentage
High (≥ 50 occurrences, ★)	27	5.6%
Medium (10–49 occurrences, ☆)	113	23.3%
Low (2–9 occurrences)	346	71.2%
Total	**486**	**100%**

High-frequency syllables are almost exclusively of the simple CV type (la, ma, sa, ta, se, ka, pa, etc.), consistent with the CV preference widely attested across the Austronesian family.

3.2 Distribution by Onset

Table 3 presents the number of syllable types per onset category.

Table 3. Syllable Types by Onset (frequency ≥ 2)

Onset	IPA	Types	Representative syllables
∅ (null onset)	—	15	a, i, u, au, o
b	b	21	ba, be, bu, bo, bun
h	h	20	ha, hi, he, ho, hu
j	j ~ dʒ	6	ja, jan, jen, ju
k	k	31	ka, ku, ki, ke, kə
l	l	40	la, li, lu, lan, lai
l’	ɭ	5	ɭa, ɭi, ɭai, ɭal, ɭe
m	m	39	ma, man, mu, mi, mal
n	n	37	na, nan, nu, ni, nə
n’	ŋ	6	ŋa, ŋo, ŋan, ŋu
p	p	30	pa, pu, pan, pə, pi

Onset	IPA	Types	Representative syllables
q	q	21	qa, qu, qo, qai, qul
r	r	17	ru, ri, ra, re, rit
s	s	49	sa, se, su, si, san
s’	ʃ	3	ʃi, ʃa, ʃe
t	t	37	ta, te, tu, ti, tan
ts	ts	15	tsu, tsa, tse, tsat
ts’	tʃ	2	tʃi, tʃa
v	v	24	va, vu, van, vi
w	w	8	wa, wan, wai, wak
y	j	2	ja, jan (y-spelling)
z	z	22	za, zu, zo, zan
z’	ʮ	5	ʮa, ʮu, ʮian, ʮaz
Clusters (ml’, mn, kn, tm, etc.)	various	28	ml’a, kna, tma

The lateral /l/ yields the largest onset class (40 types), followed by /s/ (49 types when high-frequency sibilants are included). The predominance of coronals reflects patterns common to Austronesian languages, in which coronals participate heavily in morphological processes such as prefixation and infixation.

3.3 Syllable Structure Types

The following syllable structure templates were attested:

- **V**: a, i, u (vowel nucleus only)
- **CV**: la, ma, sa, ba (most frequent; canonical type)
- **CVC**: lan, man, tan, bun
- **CVV**: lau, mai, tiu (diphthong nucleus)
- **CVVC**: laan, maan (long vowel + coda)
- **CCV**: kna, tma, ml’a, sja (onset cluster)
- **CCVC**: knat, mnan, tmat

The CV template is by far the most frequent, and syllables of CCV or greater complexity are relatively rare. This distribution is consistent with the typological markedness hierarchy of syllable structures (Blevins 1995; Maddieson 2006), whereby greater structural complexity implies lower frequency and cross-linguistic distribution.

3.4 Typologically Marked Phonemes

Four phonemes in particular distinguish Basay from most other Austronesian and Formosan languages.

Retroflex lateral /ɭ/ (orthographic `l'`): Retroflex consonants are uncommon in the Formosan branch of Austronesian. In Basay, /ɭ/ functions as an independent phoneme, generating 14 distinct syllable types (ɭa, ɭi, ɭai, ɭal, ɭe, etc.).

Voiced alveolar lateral fricative /ɮ/ (orthographic `z'`): This phoneme is typologically rare worldwide and is assigned a dedicated IPA symbol. In Basay, five syllable types with /ɮ/ onset are attested: ɮa, ɮu, ɮian, ɮaz, ɮə.

Voiced labiodental fricative /v/: While /v/ occurs in some Philippine languages, it is uncommon in the Formosan branch, making its presence in Basay a distinctive feature (Li 1996). Twenty-four syllable types with /v/ onset are recorded.

Complex onset clusters (ml’, mn, kn, tm, etc.): Forms such as ml’a, kna, tma, and mnan stand in sharp contrast to the canonical CV syllables of Polynesian languages and point to a more conservative phonotactic profile within the Malayo-Polynesian subgroup.

---

4. Typological Assessment

4.1 Cross-Linguistic Comparison of Syllable Inventory Size

The count of syllable types varies considerably across methods and definitions, particularly with respect to the treatment of tone and vowel length. Table 4 adopts a consistent baseline of non-tonal, non-length-differentiated syllable types for comparability.

Table 4. Syllable Inventory Size: Cross-Linguistic Comparison

Language	Types (approx.)	Maximum syllable shape	Notes
Hawaiian	ca. 60	CV, V	8 consonants, 5 vowels
Japanese	ca. 100	CV (+ moraic nasal/ geminate)	Mora-timed
Mandarin Chinese	ca. 400	CVC (n/ŋ codas only)	ca. 1,300 with tone

Language	Types (approx.)	Maximum syllable shape	Notes
Basay (this study)	**486**	**CCVC**	Frequency ≥ 2; clusters included
German	ca. 2,700	CCCCVCCCCC	Rich derivational morphology
English	ca. 10,000–15,000	CCCVCCCC	Large lexicon; extensive borrowing
Telugu	ca. 12,000	Complex	Dravidian; agglutinative

Sources: Neergaard & Huang (2019) for Mandarin; Fenk-Oczlon & Fenk (2021) for cross-linguistic data; for English, see also the speech recognition literature (cf. the estimate of >15,000 in subword modelling studies).

Basay's 486 syllable types clearly exceed the inventories of tone languages with strict phonotactic constraints (Hawaiian, Japanese, Mandarin) while remaining far below the inventories of languages that permit elaborate consonant clusters in both onset and coda (English, German). This positions Basay in the mid-range of the cross-linguistic spectrum.

4.2 Phoneme Inventory and Syllable Complexity

Fenk-Oczlon & Fenk (2021) demonstrate, on the basis of 61 languages from 18 families, a significant positive correlation between phoneme inventory size and mean phonemes per syllable. Maddieson (2006) similarly reports a positive association between consonant inventory size and syllable complexity in the WALS database.

Basay instantiates this generalization: its large consonant inventory—which includes the typologically marked phonemes /l̪/, /ʒ/, /ʔ/ (coda), /q/, /tʃ/, and /v/—combines with permissive onset phonotactics to generate a syllable inventory approaching 500 types. The correlation between inventory richness and syllable complexity thus holds for Basay as a specific case.

4.3 Position within the Formosan Branch

Formosan languages constitute the deepest-branching subgroups of the Austronesian family and are widely regarded as the most conservative (Blust 1999). Within Taiwan's indigenous languages, Atayal and Paiwan are known for their relatively complex syllable structures compared to Polynesian languages. Basay shares this tendency but further distinguishes itself through the co-occurrence of /l̪/ and /ʒ/—a combination not

reported for other Formosan languages in available documentation—and through onset clusters of the type CCC (e.g., ml’a /lə/).

4.4 Speaker Population and Phonological Complexity

Lupyan & Dale (2010) and Fenk-Oczlon & Fenk (2021) argue that languages spoken by smaller communities tend to retain or develop more complex phonological systems, partly because they are less subject to the simplifying pressures of adult second-language acquisition. Basay, as an extinct language of a small indigenous community with no recorded late-stage simplification, exemplifies this pattern: its complex phonology appears to have been maintained until the point of last attestation.

---

5. Methodological Caveats

Three limitations of the present study deserve explicit acknowledgement.

First, the syllabification algorithm operates on a purely phonotactic basis and does not take morphological boundaries into account. In a morphologically complex language such as Basay, affixation may create syllable-internal boundaries that differ from those assigned by the algorithm, potentially inflating the count of CCV and CCVC structures.

Second, the results are contingent on the transcriptional consistency of the source database. Where different fieldworkers or transcription conventions have been applied across entries, the same phoneme may be represented by different orthographic strings, leading to both over-counting and under-counting of syllable types. The exclusion of hapax legomena (frequency-1 forms) partially mitigates this problem.

Third, the figure of 486 syllable types represents the syllables attested in the recorded lexicon, not the full set theoretically generated by the phonological grammar. The attested inventory is a subset of the possible inventory, bounded by lexical coverage.

---

6. Conclusion

This study extracted 486 distinct syllable types (frequency ≥ 2) from the Basay lexical database and assessed their typological significance. The principal findings are as follows:

1. Basay’ s syllable inventory of 486 types exceeds those of Hawaiian (ca. 60), Japanese (ca. 100), and tone-free Mandarin Chinese (ca. 400), placing it in the mid-to-upper range of the cross-linguistic distribution.
2. The typologically marked phonemes /l̪/, /ʒ/, /v/, and complex onset clusters (ml’ , kn, tm, etc.) are the primary drivers of this inventory size.
3. These characteristics are consistent with the cross-linguistic generalization that larger consonant inventories co-occur with more complex syllable structures, and they establish Basay as one of the phonologically more complex languages within the Formosan branch of Austronesian.
4. Quantitative syllable extraction from lexical databases constitutes a viable complementary method for phonological documentation of under-described and extinct languages.

Future work should address the moraic and prosodic structure of Basay, conduct a systematic comparison with closely related Pingpu languages (Ketagalan, Kavalan), and, where audio recordings exist, supplement the orthographic analysis with acoustic phonetic data.

References

Blust, R. (1999). Subgrouping, circularity and extinction: Some issues in Austronesian comparative linguistics. In E. Zeitoun & P. J.-K. Li (Eds.), Selected papers from the Eighth International Conference on Austronesian Linguistics (pp. 31–94). Academia Sinica.

Blevins, J. (1995). The syllable in phonological theory. In J. A. Goldsmith (Ed.), The handbook of phonological theory (pp. 206–244). Blackwell.

Fenk-Oczlon, G., & Fenk, A. (2021). Linguistic complexity: Relationships between phoneme inventory size, syllable complexity, word and clause length, and population size. Frontiers in Communication, 6, 626032. https://doi.org/10.3389/fcomm.2021.626032

Li, Paul Jen-kuei (1996). The Formosan Tribes and Languages in I-Lan. Yilan: Yilan County Government.）

Li, P. J.-K. (2000). 臺灣南島語言的語音符號系統. Taipei: Crane Publishing.

Lupyan, G., & Dale, R. (2010). Language structure is partly determined by social structure. PLOS ONE, 5(1), e8559. https://doi.org/10.1371/journal.pone.0008559

Maddieson, I. (2006). Correlating phonological complexity: Data and validation. Linguistic Typology, 10(1), 89–118.

Neergaard, K. D., & Huang, C.-R. (2019). Constructing the Mandarin phonological network: Novel syllable inventory used to identify schematic segmentation. Complexity, 2019, 6979830. https://doi.org/10.1155/2019/6979830

Institute of Linguistics, Academia Sinica (Ed.). Basay Lexical Database (`basay_dict.jsonl`). Taipei: Academia Sinica.

Data extraction and analysis were conducted in Python. The lexical database is publicly available through Academia Sinica.