巴賽語音節目錄及其音韻類型學定位

—— 基於詞典資料的計量分析 ——

蔡永桂（Yung-kuei Tsai）

2026年6月

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

巴賽語音節目錄及其音韻類型學定位

——基於詞典資料的計量分析——

摘要

本文利用巴賽語詞典資料庫（3,324筆條目），對曾使用於台灣北部的南島語系語言巴賽語（Basay）進行音節目錄的計量提取，並就其音韻結構加以描述，同時與世界語言的音節規模進行類型學比較。以排除南島語族原始語（PAN）重建形式後的2,364筆條目為分析對象，確定出詞頻2次以上的音節486種。巴賽語具有若干在音韻類型學上值得關注的特徵：捲舌側面音（ɭ）、有聲齒齦側面擦音（ɮ）、軟顎鼻音（ŋ）、齦顎擦音（ʃ）等特殊音素，以及複雜的音節起始輔音叢（onset cluster）。486種的音節目錄規模，遠超夏威夷語（約60種）與日語（約100種），亦超過普通話（約400種，不計聲調），在世界語言中屬於中上等規模，在台灣南島語（台灣原住民語）中更顯示出較為複雜的音韻面貌。

關鍵詞： 巴賽語・平埔族・音節目錄・音韻類型學・南島語族・瀕危語言

---

1. 前言

巴賽語（Basay，又作巴賽、巴薩依）係台灣北部宜蘭平原至台北盆地東北側一帶的平埔族——巴賽族所使用的南島語系語言。17世紀荷蘭統治時期的文獻及清代史料中均有零散詞彙記錄，然而，學界普遍認為此語言的母語使用者在20世紀前半葉便已近乎消亡（李壬癸 1996；2000）。目前巴賽語被列為消亡語言，中央研究院語言學研究所等機構持續推動記錄保存與語言復振工作。

本文有兩個目標：其一，運用現存詞典資料計量提取巴賽語的音節目錄，同時明確呈現正字法與國際音標（IPA）的對應關係，對巴賽語音節體系加以描述；其二，將所得音節目錄的規模與世界各語言進行比較，以釐清巴賽語在音韻類型學上的位置。

本文所採用的計量提取方法，並非用以取代基於語法描述的語音學分析，而是以現存詞彙資料為基礎，對音節分布狀況進行描述性記錄——這對文獻資料有限的消亡語言而言，尤具補充性的學術價值。

---

2. 資料與方法

2.1 資料來源

本研究使用巴賽語詞典資料庫（`basay_dict.jsonl`，共3,324筆條目）。各條目包含巴賽語正字法形式、中日英三語譯文及來源代碼。來源代碼分為以下幾類：`B`（巴賽語固有詞彙）、`T`（台語借詞）、`M`（閩南語借詞）、`S`、`V`，以及`PAN`（南島語族原始語重建形式）。

標記為`PAN`的條目（960筆）予以排除。南島語族原始語重建形式屬於比較語言學上的抽象構擬，並非巴賽語現存詞彙的實際記錄，若納入分析將影響音節目錄的真實性。其餘2,364筆條目構成本文的分析語料。

2.2 正字法體系

巴賽語採用拉丁字母正字法。表1呈現正字法符號與IPA的對應關係。

表1　巴賽語正字法・IPA對照表

正字法	IPA	說明
`n'`	ŋ	軟顎鼻音
`s'`	ʃ	齦顎擦音
`l'`	ɭ	捲舌側面音
`z'`	ɮ	有聲齒齦側面擦音
`o'`	ə	中央中元音（央元音）
`'`（韻尾）	ʔ	聲門塞音（音節末位）
`q`	q	小舌／咽喉音（推定）
`ts`	ts	齒齦塞擦音
`ts'`	tʃ	齦顎塞擦音
`j`	j～dʒ	近音或塞擦音（依語境而定）

正字法	IPA	說明
`v`	v	有聲唇齒擦音

聲門塞音（`'`）作為前一音節的韻尾輔音出現，並非獨立的音節起始輔音（onset）。因此，以`'`起首的書寫形式（如`'a`、`'ul`）被視為前音節韻尾轉寫的附帶現象，予以排除，不計入音節目錄。

2.3 音節提取程序

音節提取按以下步驟進行：

1. 對各條目的巴賽語正字法形式進行清理，移除注記、括號及替代形式，僅保留主要形式。
2. 採用基於（C*）V（V?）（C?）模板的音節切分演算法，以元音字母（a, e, i, o, u及特殊元音字符）作為音節核心。若後接輔音後緊跟元音，則將該輔音劃歸下一音節的起始輔音。
3. 移除以空格、標點符號、數字、連字符或撇號（`'`）起首的形式（視為雜訊）。
4. 進一步排除詞頻為1的音節，以降低轉寫差異及借詞附帶現象的影響。
5. 保留詞頻2次以上的音節，構成最終目錄。

---

3. 音節目錄描述

3.1 整體統計

上述程序共確定486種詞頻2次以上的音節。詞頻分布如表2所示。

表2　各詞頻區間的音節種數

詞頻區間	種數	占比
高頻（50次以上，★）	27種	5.6%
中頻（10至49次，☆）	113種	23.3%
低頻（2至9次）	346種	71.2%
合計	**486種**	**100%**

高頻音節幾乎全為簡單的CV型（la, ma, sa, ta, se, ka, pa等），與南島語族中普遍存在的CV音節偏好一致。

3.2 依起始輔音（Onset）分布

表3呈現各起始輔音類別的音節種數。

表3　主要起始輔音別音節種數（詞頻2次以上）

起始輔音	IPA	音節種數	代表音節（高頻排列）
∅（零起始輔音）	—	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
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拼法）
z	z	22	za, zu, zo, zan
z’	ʒ	5	ʒa, ʒu, ʒian, ʒaz
輔音叢（ml’, mn, kn, tm等）	各異	計28	ml’a, kna, tma等

/l/起始類別擁有最多音節種數（40種），/s/次之（49種）。冠狀音（coronal）的優勢反映了南島語族中冠狀音廣泛參與構詞過程（如前綴、中綴）的共同傾向。

3.3 音節結構類型

本研究所確認的主要音節結構模板如下：

- **V**：a, i, u等（單元音）
- **CV**：la, ma, sa, ba等（最多見，基本類型）
- **CVC**：lan, man, tan, bun等
- **CVV**：lau, mai, tiu等（雙元音核）
- **CVVC**：laan, maan等（長元音＋韻尾）
- **CCV**：kna, tma, ml’a, sja等（起始輔音叢）
- **CCVC**：knat, mnan, tmat等

CV型最為普遍，CCV及以上複雜結構的詞頻相對較低，與音節結構有標性等級序列（Blevins 1995；Maddieson 2006）的類型學預測一致。

3.4 類型學上有標的音素

以下四個音素是巴賽語有別於大多數南島語族語言及台灣南島語的顯著特徵。

捲舌側面音（ɭ，正字法`l'`）： 捲舌音在台灣南島語分支中並不常見。巴賽語中ɭ作為獨立音素發揮功能，構成14種音節（ɭa, ɭi, ɭai, ɭal, ɭe等）。

有聲齒齦側面擦音（ɮ，正字法`z'`）： 此音素在世界語言中極為罕見，IPA設有專門符號。巴賽語中確認有5種音節以ɮ起首：ɮa, ɮu, ɮian, ɮaz, ɮə。

有聲唇齒擦音（v）： /v/雖見於部分菲律賓語言，在台灣南島語分支中較為罕見，巴賽語中此音素的存在被視為一大特徵（李壬癸 1996）。共記錄24種以/v/起首的音節。

複雜起始輔音叢（ml’, mn, kn, tm等）： ml’a、kna、tma、mnan等形式與波里尼西亞語言的典型CV音節形成鮮明對比，顯示出南島馬來・波里尼西亞語群中較為保守的音節音位配列（phonotactics）。

---

4. 類型學比較

4.1 音節目錄規模的跨語言比較

音節種數的計算因聲調、元音長短是否計入而差異顯著。本文表4採用不區分聲調及長短的音節形式種數作為比較基準。

表4　主要語言音節目錄規模比較

語言	音節種數（約）	最大音節結構	備注
夏威夷語	約60種	CV, V	8個輔音、5個元音
日語	約100種	CV（＋撥音、促音）	以拍（mora）為基本單位
普通話	約400種	CVC（韻尾僅n/ŋ）	含聲調約1,300種
巴賽語（本研究）	**486種**	**CCVC**	詞頻2次以上；含輔音叢
德語	約2,700種	CCCCVCCCCC	豐富的派生構詞
英語	約10,000至15,000種	CCCVCCCC	詞彙量龐大；借詞眾多
泰盧固語	約12,000種	複雜	達羅毗荼語系；黏著語

由此可見，巴賽語的486種音節，明顯超過具有嚴格音位配列限制的聲調語言（夏威夷語、日語、普通話），但遠低於允許複雜輔音叢的屈折語（英語、德語）。巴賽語在跨語言分布中屬於中等偏上的規模。

4.2 音素目錄與音節複雜性的關係

Fenk-Oczlon & Fenk（2021）基於18個語系61種語言的分析，證明音素目錄規模與每音節平均音素數之間存在顯著正相關。Maddieson（2006）同樣利用WALS（語言結構世界圖集）資料指出，輔音目錄較大的語言其音節結構往往也較為複雜。

巴賽語的情況正好印證了這一普遍規律：包含ɭ、ɮ、ʔ（韻尾）、q、tʃ、v等類型學有標音素的龐大輔音目錄，與容許複雜起始輔音叢的寬鬆音節音位配列相互作用，共同催生了接近500種的音節目錄。因此，巴賽語可視為音素目錄豐富性與音節複雜性正相關這一類型學規律的具體例證。

4.3 在台灣南島語分支中的位置

台灣南島語（福爾摩沙語群）代表南島語族最早期的分支，被普遍認為保留了最多的古語特徵（Blust 1999）。在台灣原住民語中，泰雅語、排灣語以其相對複雜的音節結構著稱。巴賽語同樣具有此傾向，但捲舌側面音（ɭ）與有聲齒齦側面擦音（ɮ）的共存，在現有台灣南島語文獻中尚未見於其他語言，顯示巴賽語在音韻上具有獨特性。

4.4 使用人口與音韻複雜性

Lupyan & Dale（2010）及Fenk-Oczlon & Fenk（2021）指出，使用人口較少的語言往往保有或發展出更為複雜的音韻體系，部分原因在於小型社群受到的語言接觸壓力及成人二語習得簡化效應較弱。巴賽語作為已消亡的小型原住民社群語言，其複雜音韻體系似乎維持至有記錄的最後階段，正是這一假說的極端例證。

---

5. 方法論說明

本研究的計量方法存在若干值得明確說明的局限性。

第一，音節切分演算法以純粹的音位配列為基礎，未考量形態邊界。對於形態結構複雜的巴賽語而言，詞綴可能在音節內部形成切分點，導致CCV及CCVC結構的數量被高估。

第二，分析結果依賴原始資料庫的轉寫一致性。若不同田野調查者或轉寫規範被應用於不同條目，同一音素可能以不同的正字法形式呈現，造成音節種數的重複計算或遺漏。排除出現頻率為1的音節，是為部分緩解這一問題所採取的措施。

第三，本研究確定的486種音節，係「詞典記錄詞彙中觀察到的音節種數」，而非「巴賽語音韻體系在理論上允許的全部音節種數」。前者是後者的子集，受詞彙覆蓋率的限制。

基於以上局限性，本文的描述應被理解為基於現存最佳資料的暫行音節目錄。

---

6. 結論

本文利用巴賽語詞典資料庫，以計量方法提取出詞頻2次以上的音節486種，並就其音韻特徵與類型學意義進行了論述。主要研究發現如下：

1. 巴賽語的486種音節目錄，超過夏威夷語（約60種）、日語（約100種）及不計聲調的普通話（約400種），在跨語言分布中屬於中上等規模。
2. 捲舌側面音（ɭ）、有聲齒齦側面擦音（ɮ）、有聲唇齒擦音（v），以及複雜起始輔音叢（ml’, kn, tm等），是音節目錄規模擴大的主要驅動因素。
3. 這些特徵與「較大的輔音目錄與較複雜的音節結構共現」這一跨語言規律一致，並確立了巴賽語在台灣南島語分支中音韻較為複雜的語言地位。
4. 基於詞典資料的計量音節提取，可作為描述記錄不足及已消亡語言音韻體系的有效輔助方法。

未來研究方向包括：針對巴賽語的拍（mora）及韻律結構進行分析；與凱達格蘭語（Ketagalan）、噶瑪蘭語（Kavalan）等相鄰平埔族語言進行系統性比較；以及在現有音頻錄音資料的基礎上，以聲學語音學方法補充正字法分析。

參考文獻

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

李壬癸 (1996). 《宜蘭縣南島民族與語言》. 宜蘭縣：宜蘭縣政府。

李壬癸 (2000). 台灣南島語言的語音符號系統. 台北：文鶴出版.

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

中央研究院語言學研究所（編）. 巴賽語詞典資料庫（`basay_dict.jsonl`）. 台北：中央研究院.

資料提取與分析以Python進行。詞典資料庫由中央研究院公開提供。