

Lexical and Phonological Relationship between Bulukumba Buginese and Toraja: A Historical-Comparative Study

Awal Wahyudi

Universitas Islam Negeri Alauddin Makassar
awal.wahyudi@uin-alauddin.ac.id

Abstract

This study examines the lexical and phonological relationship between Bulukumba Buginese (BB) and Toraja (TL), with attention to local variety, cognate criteria, and recurrent sound correspondences. The dataset comprises 200 Swadesh basic-vocabulary meanings collected in 2026 from six native speakers, three per variety, through interviews, note-taking, and audio recording. The analysis combines lexicostatistics, glottochronology, phonemic-correspondence identification, and comparison with Proto-Austronesian (PAN) forms retained in the dataset. Eighty-seven related pairs were identified: 36 identical pairs, 20 pairs differing by one phoneme, 20 pairs supported by phonemic correspondence, and 11 pairs based on phonetic similarity. The resulting cognacy rate is 43.5%; excluding the 11 phonetic-similarity pairs reduces it to 38.0%, which remains within the same “family” band of the legacy lexicostatistical classification used in the study. With a retention constant of 0.805, the initial glottochronological estimate is approximately 1,919 years before 2026. Recalculation of the stated standard-error formula gives 0.0351 and a model-dependent range of approximately 2,097–1,740 years before 2026 (roughly 71 BCE–286 CE). Six correspondence sets are identified: /ŋ~n/, /w~b/, /ə~a/, /r~d/, /ø~k/, and /o~e/, with unequal evidential strength. PAN-oriented comparisons also show lenition, fortition, aphaeresis, apocope, syncope, and prothesis. The findings support BB–TL historical relatedness while showing that neither cognacy percentages nor glottochronological dates alone establish genealogical subgrouping.

Keywords: *cognates, glottochronology, historical-comparative linguistics, lexicostatistics, sound correspondence*

INTRODUCTION

Historical-comparative linguistics establishes historical relatedness most securely through recurrent correspondences and etymologically defensible cognate sets rather than surface resemblance alone. Quantitative lexical methods can support that task, but their validity depends on the quality and transparency of cognate coding. Large-scale research has shown both the usefulness and the limits of automatic or quantitative word comparison: methods that exploit language-specific sound correspondences outperform simple similarity measures, while expert-annotated cognates remain an important benchmark (List, Greenhill,

& Gray, 2017; Rama, List, Wahle, & Jäger, 2018; Jäger, 2018, 2019; Batsuren, Bella, & Giunchiglia, 2022). This study therefore treats lexicostatistics as one layer of evidence and evaluates it together with recurrent phonological patterns.

Bulukumba Buginese (hereafter BB) and Toraja (hereafter TL) are Austronesian varieties spoken in South Sulawesi. Austronesian research demonstrates that lexical differentiation is shaped by historical transmission as well as geographic and social separation, making locality relevant to comparative interpretation (Padilla-Iglesias, Gjesfjeld, & Vinicius, 2020). Buginese itself has a distinctive phonetic system (Talaohu, 2019), and variety-specific work has documented Bulukumba Buginese as an identifiable local object of linguistic analysis rather than an abstract undifferentiated “Buginese” code (Wahyudi, 2023). In the same regency, Jannah, La Ino, and Zahrani (2024) found a 52% lexical relationship between Buginese and Coastal Konjo using 200 Swadesh meanings, further showing why South Sulawesi comparisons should specify the variety, locality, and coding procedure used.

The most direct benchmark is Mayangsari (2020), who reported 53% relatedness between Buginese and Toraja. A closer audit, however, shows that the earlier study began with 200 Swadesh meanings but excluded nine glosses, calculated 101 cognates from the remaining 191 items, and used 101/191 as the denominator for the 53% result. The present dataset counts 87 related pairs across all 200 meanings. Consequently, the difference between 53% and 43.5% cannot be attributed to dialect alone: the denominator, lexical source, elicitation, and treatment of near-cognates also differ. This issue mirrors a broader pattern in Indonesian historical-comparative work, where lexicostatistical outcomes vary with language sampling and cognate criteria (Afria, Izar, Prawolo, & Arezky, 2020; Sulistiyarini & Hendrokumoro, 2023; Fadhila, Saddhono, & Suwandi, 2024). The unresolved question is therefore not whether Buginese and Toraja have ever been compared, but what evidence emerges when Bulukumba Buginese is analyzed explicitly and the numerical result is checked against the phonological evidence in the same dataset.

The novelty of this study lies in three connected contributions. First, it isolates Bulukumba Buginese as the Buginese variety under analysis and positions the result against earlier, more general Buginese–Toraja estimates. Second, it does not stop at the cognacy percentage: the 43.5% estimate is evaluated together with six recurrent BB–TL phonemic correspondence sets and PAN-oriented sound changes. Third, it audits the internal arithmetic of the glottochronological procedure and adds a sensitivity check for the most judgment-

dependent cognate category. These steps address a central methodological concern in comparative work: transcription, coding, and recurrent phonological structure must remain inspectable if a lexical percentage is to carry historical meaning (Anderson et al., 2018; Macklin-Cordes, Bowern, & Round, 2021).

Against that background, this study asks: (1) what lexical relationship between BB and TL is obtained from the 200-item dataset; (2) what approximate divergence interval follows from the glottochronological model when its stated arithmetic is applied consistently; and (3) what recurrent phonemic correspondences and PAN-related sound changes are attested in the compared forms? The objectives are to quantify cognacy, test the robustness of the legacy classification to a stricter treatment of phonetic-similarity pairs, estimate a model-dependent temporal interval, and describe the principal phonological patterns. The intended contribution is a more auditable variety-level account of BB–TL relatedness rather than a stronger genealogical claim than the evidence can support.

METHOD

This research uses a combined qualitative-quantitative descriptive design within historical-comparative linguistics. The quantitative component calculates lexical cognacy with lexicostatistics and reproduces the study's glottochronological estimate. The qualitative component identifies recurrent phonemic correspondences and classifies attested sound changes. This combined design is common in recent Indonesian comparative studies because a percentage alone cannot establish how lexical similarities are phonologically structured (Aisyah & Widayati, 2022; Lailiyah & Wijayanti, 2022; Mahendra & Hendrokumoro, 2022; Mahriyuni, Pramuniati, & Maftuhah, 2023).

The primary data consist of 200 Swadesh basic-vocabulary meanings for BB and TL. Field data were obtained in the 2026 study through interviews, recording, and note-taking from six native speakers: three BB speakers and three TL speakers. The informants were native speakers residing in the research locations, aged 25–60 years, and able to participate in oral elicitation. Using three speakers per variety permits cross-checking of elicited lexical forms and reduces dependence on a single speaker's idiolect; however, the sample is not treated as statistically representative of all within-variety or dialectal variation. The source documentation identifies the comparison variety as Toraja (TL) but does not provide a

narrower subvariety label, so no more specific dialect designation is inferred here. It also does not document a formal aggregation rule for resolving any inter-speaker lexical variants; accordingly, the analysis reports the lexical forms retained in the dataset rather than reconstructing an unrecorded decision procedure. No additional demographic characteristics are inferred because they are not documented in the source data. The complete lexical list is retained as supplementary material so that individual cognate decisions can be rechecked.

For the Proto-Austronesian comparison, the analysis retains the PAN forms supplied in the source dataset and uses them as comparative reference forms rather than claiming a new reconstruction. A contemporary digital reference, the Austronesian Comparative Dictionary dataset, is cited to make future verification of individual PAN etyma possible (Blust, Trussel, & Smith, 2023). This distinction matters: the core empirical claim concerns relationships among the attested BB and TL forms, while PAN forms provide a diachronic interpretive layer.

Cognate coding follows the four categories used in the dataset: identical pairs, pairs differing by one phoneme, pairs supported by recurrent phonemic correspondence, and pairs judged to have close phonetic similarity. Comparable Indonesian studies employ related categories when operationalizing cognacy and correspondence (Erniati, 2021; Hadi & Kusumaningrum, 2020; Muhammad & Hendrokumoro, 2022; Sholeha & Hendrokumoro, 2022; Fadhila et al., 2024). The primary cognacy rate is $C = J/n \times 100$, where J is the number of related pairs and n is the number of compared meanings. With $J = 87$ and $n = 200$, $C = 43.5\%$. Because “phonetic similarity” is the least constrained of the four categories, a sensitivity calculation is also reported after removing those 11 pairs: $(87-11)/200 \times 100 = 38.0\%$. The traditional percentage bands reproduced in Table 1 are retained only as the classification convention used by this line of research, not as a substitute for formal phylogenetic subgrouping.

The glottochronological calculation uses $w = \log(c)/(2 \log(r))$, where w is estimated separation time in millennia, c is the cognate proportion, and r is the assumed retention constant. The worked calculation uses $r = 0.805$, so this value is applied consistently. The standard error is recalculated from the formula $S = \sqrt{[C(1-C)/n]}$. With $C = 0.435$ and $n = 200$, $S = 0.0351$; therefore $C1 = C + S = 0.4701$. The resulting range follows the legacy glottochronological error procedure adopted in the study and is reported only as a model-dependent heuristic range; it is not a conventional 95% statistical confidence interval. This

caution is necessary because modern computational historical linguistics explicitly separates evidence for relatedness from assumptions used in time calibration and phylogenetic inference (Jäger, 2018, 2019; Rama et al., 2018).

For qualitative analysis, BB and TL word pairs are aligned and inspected for recurrent segmental relationships in comparable positions or phonological environments. A proposed correspondence receives greater evidential weight when it recurs across several lexical items; two-item sets are therefore described more cautiously than larger sets. Sound changes relative to the PAN forms are described using the labels retained from the source analysis—lenition, fortition, aphaeresis, apocope, syncope, and prothesis—where those labels are supported by the listed forms (Rizqi & Widayati, 2021; Utamayana, Budasi, & Ramendra, 2025). These labels are used descriptively rather than as evidence for exceptionless historical laws. Standardized phonetic transcription is critical to this kind of comparison because inconsistent symbol systems can create artificial similarities or differences (Anderson et al., 2018).

Table 1. Legacy Lexicostatistical Classification Used in the Study

Level	Cognate Words (%)	Approx. Separation (centuries)
Language	100 – 81	0 – 5
Family	81–36	5–25
Stock	36–12	25–50
Microphylum	12–4	50 –75
Mesophylum	4 – 1	75 –100
Macrophylum	< 1%	>100

FINDINGS AND DISCUSSION

The analysis identifies 87 related pairs among 200 meanings: 36 identical pairs, 20 pairs differing by one phoneme, 20 pairs supported by recurrent phonemic correspondence, and 11 pairs categorized by phonetic similarity. Results are interpreted at two levels. The first is the numerical proportion of related basic vocabulary. The second is the structure of the evidence: whether similarities recur in phonologically definable environments and whether the result remains broadly stable when the least constrained cognate category is removed.

Lexicostatistical Relationship

The primary calculation is $87/200 \times 100 = 43.5\%$. Under the legacy classification in Table 1, 36–81% is labelled “family,” so BB and TL fall within that category. The classification is not entirely dependent on the 11 phonetic-similarity judgments. Excluding those items produces $76/200 = 38.0\%$, which still lies above the 36% lower boundary of the

same legacy band. This sensitivity check does not prove subgrouping, but it shows that the categorical placement is not created solely by the most subjective cognate class. The more defensible interpretation is that the dataset contains substantial basic-vocabulary relatedness that warrants examination of recurrent correspondences rather than a claim based on percentage alone.

The comparison with Mayangsari (2020) is especially informative. Her 53% estimate is based on 101 cognates among 191 counted meanings after nine of the original 200 glosses were excluded, whereas the present study uses 87 related pairs over the full 200-item denominator. Thus, 53% versus 43.5% is not an apples-to-apples dialect comparison. It combines a possible variety effect with differences in denominator and cognate coding. Jannah et al. (2024) likewise reported 52% for Buginese–Coastal Konjo in Bulukumba, while other Indonesian studies report substantially different percentages for other pairings under similar lexicostatistical frameworks (Erniati, 2021; Aisyah & Widayati, 2022; Sulistiyarini & Hendrokumoro, 2023). These comparisons support a narrower conclusion: the 43.5% estimate is a property of this BB–TL dataset and coding protocol, not a universal constant for every Buginese–Toraja wordlist.

$$C = J/n \times 100 = 87/200 \times 100 = 43.5\%$$

Glottochronological Estimate and Arithmetic Correction

With $c = 0.435$ and $r = 0.805$, the glottochronological equation yields $w = 1.9188$ millennia, or approximately 1,919 years before the 2026 research year. The standard-error value stated in the earlier calculation was 0.031; applying the formula actually provided in the study, $S = \sqrt{[(0.435 \times 0.565)/200]}$, yields 0.0351. The adjusted cognate proportion is therefore $C1 = 0.4701$ rather than 0.466. Reporting this correction is necessary because the separation interval is directly sensitive to the cognate proportion inserted into the logarithmic model.

Using $C1 = 0.4701$ with the same retention constant produces an adjusted estimate of approximately 1,740 years before 2026. The difference from the initial estimate is approximately 179 years. Following the interval procedure used in the study, the resulting range is approximately 2,097–1,740 years before 2026, corresponding roughly to 71 BCE–286 CE. These calendar equivalents are presented only to communicate scale; they are not independent archaeological or historical dates.

The methodological implication is more important than the small numerical shift. Glottochronology embeds a constant-rate assumption, while contemporary phylogenetic work models lexical change and time with more explicit uncertainty and language-family structure (Jäger, 2018, 2019). Austronesian lexical turnover can also be associated with geographic and social isolation (Padilla-Iglesias et al., 2020). The interval should therefore be treated as a heuristic estimate conditional on the chosen retention constant, not as a direct date for population separation, migration, or ethnogenesis.

$$w = \log(0.435) / [2 \log(0.805)] = 1.9188 \approx 1,919 \text{ years before 2026}$$

$$S = \sqrt{[(0.435 \times 0.565)/200]} = 0.0351; C_1 = 0.435 + 0.0351 = 0.4701$$

Corrected interval $\approx 2,097\text{--}1,740$ years before 2026 (approximately 71 BCE–286 CE)

Phonemic Correspondences

Six recurrent phonemic correspondence sets are identified in the dataset: /ŋ~n/, /w~b/, /ə~a/, /r~d/, /ø~k/, and /o~e/. Their evidential strength is unequal because the supporting sets range from two to seven lexical items. The seven-item /ə~a/ set and the three-item word-final /ŋ~n/ set provide stronger recurrence than the two-item sets, while all proposed patterns remain provisional until tested on broader lexical and dialectal samples. This hierarchy of evidence is important: historical comparison depends on repeated relationships in comparable environments, not merely on phonetic resemblance (Hadi & Kusumaningrum, 2020; List et al., 2017; Macklin-Cordes et al., 2021).

Correspondence /ŋ~n/

Three lexical items show BB final /ŋ/ corresponding to TL /n/: ‘star’, ‘push’, and ‘tree’. The recurring word-final environment strengthens this pattern relative to a one-off resemblance. The original rule can therefore be represented descriptively as BB /ŋ/ ~ TL /n/ / __# for the attested forms.

Table 2. BB–TL Correspondence /ŋ~n/

Swadesh No.	Gloss	BB	TL	Rule
37	star	[bintaŋ]	[bintœn]	/ŋ-n/-#
66	push	[sorɔŋ]	[surun]	/ŋ-n/-#
159	tree	[pɔŋ]	[pɔhɔn]	/ŋ-n/-#

Correspondence /w~b/

Two items, ‘moon’ and ‘night’, show BB /w/ corresponding to TL /b/ near the beginning of the word or penultimate syllable. The pattern is suggestive but limited by the

small number of examples. It should therefore be retained as an attested correspondence set, not generalized beyond the available lexical evidence.

Table 3. BB–TL Correspondence /w~b/

Swadesh No.	Gloss	BB	TL	Rule
39	moon	[wuləŋ]	[bulan]	/w-b/#-
137	night	[wənni]	[bɔŋi]	/w-b/#-

Correspondence /ə~a/

Seven lexical items support /ə~a/, making it the numerically strongest correspondence set in this analysis. The examples include ‘wash’, ‘inside’, ‘four’, ‘say’, ‘name’, ‘squeeze’, and ‘know’. The correspondence occurs in both penultimate and final-syllable environments, often before a consonant. Its recurrence across multiple meanings provides stronger qualitative support than the two-item correspondence sets.

Table 4. BB–TL Correspondence /ə~a/

Swadesh No.	Gloss	BB	TL	Rule
49	wash	[səsaʔ]	[sasaʔ]	/ə-a/-C
59	inside	[rilaləŋ]	[lan]	/ə-a/-C
70	four	[əpaʔ]	[aʔpaʔ]	/ə-a/-C
110	say/speak	[makkaʔda]	[maʔkada]	/ə-a/-C
146	name	[asəŋ]	[saŋa]	/ə-a/-C
155	squeeze	[mapəraʔ]	[maʔpara]	/ə-a/-C
176	know	[isəŋ]	[isan]	/ə-a/-C

Correspondence /r~d/

The items ‘leaf’ and ‘eat’ show BB /r/ corresponding to TL /d/ before a vowel. Although only two examples are available, the environment is consistent within the dataset. Additional lexical material would be needed to determine whether the pattern is productive across a larger vocabulary.

Table 5. BB–TL Correspondence /r~d/

Swadesh No.	Gloss	BB	TL	Rule
55	leaf	[raun]	[daun]	/r-d/-V
136	eat	[manre]	[mande]	/r-d/-V

Correspondence /ø~k/

The forms for ‘skin’ and ‘louse’ display zero in BB where TL has initial /k/. In comparative notation this is represented as /ø~k/ word-initially. Because zero correspondences are particularly sensitive to morphological segmentation and lexical history, the interpretation should remain tied to the specific forms shown in Table 6.

Table 6. BB–TL Correspondence /ø~k/

Swadesh No.	Gloss	PAN	BB	TL	Rule
118	skin	*kuliC	[uliʔ]	[kuliʔ]	/ø-k/#-
120	louse	*qatimela	[utu]	[kutu]	/ø-k/#-

Correspondence /o~e/

Four items show a BB /o/ and TL /e/ relationship in forms associated with an adjacent /k/: ‘when’, ‘bite’, ‘if’, and ‘dry’. The four examples therefore suggest an /o~e/ correspondence associated with /k/. The recurrence is sufficient to retain the pattern as an attested correspondence set, but the present dataset is too limited to establish the precise conditioning environment as a categorical historical rule.

Table 7. BB–TL Correspondence /o~e/

Swadesh No.	Gloss	BB	TL	Rule
35	when	[nakko]	[iake]	/o-e/k-
77	bite	[okkoʔ]	[kekeʔ]	/o-e/k-
105	if	[narɛkko]	[yake]	/o-e/k-
114	dry	[marakko]	[maranjke]	/o-e/k-

Sound Changes from Proto-Austronesian Forms

The PAN comparison adds a diachronic layer to the BB–TL analysis. The purpose is not to reconstruct Proto-Austronesian anew, but to describe how selected BB and TL forms relate to the PAN forms already supplied in the dataset. The Austronesian Comparative Dictionary dataset provides a current digital reference for checking individual etyma (Blust et al., 2023). Studies of PAN reflexes in Indonesian varieties likewise show that inherited forms may exhibit systematic replacement, deletion, and addition rather than surface identity (Rizqi & Widayati, 2021; Utamayana et al., 2025).

Lenition and fortition. The source analysis classifies BB *b > w and *k > ʔ, and TL *v > w, as lenition. The examples include PAN *‘abu’ > BB [awu], PAN *anak > BB [anaʔ], and PAN *danav > TL [danaw]. It also labels TL *ə > a in the forms for ‘four’, ‘egg’, and ‘three’ as fortition. The latter label is retained here for consistency with the source analysis and is treated descriptively rather than as an independently established general strengthening rule. Because a sound change is stronger when supported by multiple etyma and a clear environment, the single or small-number examples should not be generalized beyond the evidence shown in the tables.

Sound deletion. The data contain three positional types. Aphaeresis is represented by initial *k > ø in the BB form for ‘see’; apocope by final *h > ø in forms such as ‘fruit’,

‘white’, ‘earth/soil’, and ‘breast’; and syncope by intervocalic deletion, including *v > ø/V–V in ‘two’ and *h > ø/V–V in ‘year’ and ‘old’. These cases illustrate why cognacy cannot be evaluated by raw form identity alone: historically related forms may diverge through recurrent segment loss. Comparable PAN-reflex studies also emphasize positional sound change as a key interpretive layer (Rizqi & Widayati, 2021; Utamayana et al., 2025).

Sound addition. One prothesis example appears in the dataset: PAN *‘inum’ ‘drink’ corresponds to BB [minuŋ], represented as *ø > m / #___. A single token is insufficient to establish a general historical rule, so the pattern is reported as an attested lexical change rather than an exceptionless law. This distinction between an observation and a generalized correspondence is maintained throughout the discussion.

Table 8. Strong and Weak Sound Pairs Used in the Original Analysis

Stronger Sound	Weaker Sound
p	b
p	f
f	h
x	h
b	w
v	w
a	ə
d	l
s	r
k	ʔ

Table 9. Lenition

Swadesh No.	Gloss	PAN	BB	TL	Rule
1	ash	*‘abu’	[awu]	-	*b > w
6	child	* anak	[anaʔ]	-	*k > ʔ
52	lake	*danav	-	[danaw]	*v > w

Table 10. Fortition

Swadesh No.	Gloss	PAN	BB	TL	Rule
70	four	*ə(m)pat	-	[aʔpaʔ]	*ə > a
185	egg	*/t/eluR	-	[talɔʔ]	*ə > a
191	three	*təlu'	-	[talu]	*ə > a

Table 11. Aphaeresis

Swadesh No.	Gloss	PAN	BB	TL	Rule
130	see	*kita'	[ita]	-	*k > ø/ #-

Table 12. Apocope

Swadesh No.	Gloss	PAN	BB	TL	Rule
38	fruit	*bu'ah	[bua]	[bua]	*h > ø/-#
163	white	*putih	[pute]	-	*h > ø/-#
180	earth/soil	*tanah	[tana]	-	*h > ø/-#
188	breast	*susuh	[susu]	[susu]	*h > ø/-#

Table 13. Syncope

Swadesh No.	Gloss	PAN	BB	TL	Rule
67	two	*duva'	[dua]	[dua]	*v > ø/V-V
177	year	*tahun	[taun]	[taun]	*h > ø/V-V
196	old	*tuha'	[toa]	[tua]	*h > ø/V-V

Table 14. Prothesis

Swadesh No.	Gloss	PAN	BB	TL	Rule
143	drink	*'inum	[minuŋ]	-	*ø > m /#-

Interpretation, Comparison with Previous Research, and Limitations

Taken together, the evidence is stronger than a bare 43.5% figure but narrower than a complete subgrouping demonstration. The lexical result remains in the same legacy category when the 11 phonetic-similarity items are removed, yielding 38.0%. More importantly, several similarities can be organized into recurrent correspondence sets, particularly /ə~a/ and word-final /ŋ~n/. The PAN-oriented comparisons add plausible patterns of segmental replacement, deletion, and addition, including changes described in the source analysis as lenition and fortition. These converging layers support historical relatedness, while the small correspondence sets and manual coding prevent stronger claims about the precise branching position of BB and TL.

Four limitations are material. First, although three speakers per variety were sampled, the speaker sample remains small and is sufficient for lexical cross-checking rather than for estimating the full range of within-variety or dialectal variation. Second, the source documentation does not provide a narrower Toraja subvariety label or a formal rule for resolving any inter-speaker lexical variants; the analysis therefore avoids inferring information not recorded in the dataset. Third, cognate coding is manual, and the “phonetic similarity” category is less constrained than recurrent correspondence; the 38.0% sensitivity result partly addresses, but does not eliminate, that issue. Fourth, phonemic correspondences are based on a 200-item list and several sets contain only two examples. Contemporary

cognate research shows that explicit annotation standards, reproducible transcription, and correspondence-aware methods reduce the risk of mistaking chance similarity for inheritance (List et al., 2017; Rama et al., 2018; Batsuren et al., 2022; Anderson et al., 2018). These limitations should therefore be treated as design constraints, not as reasons to discard the observed patterns.

Future research should increase the number of speakers and localities beyond the three-speaker-per-variety design, separate uncertain cognates before calculating primary and sensitivity estimates, use independent coding by more than one analyst, and standardize phonetic transcription. Wider comparison across Buginese, Toraja, Makassar, Konjo, and other South Sulawesi varieties would test whether the six correspondence sets are local, contact-induced, or inherited more deeply. A machine-readable dataset could then support correspondence-aware cognate detection and explicit phylogenetic modeling, approaches that are increasingly feasible in computational historical linguistics (Jäger, 2018, 2019; Rama et al., 2018). Such work would extend the present study without using 43.5% or the glottochronological interval as a predetermined target.

CONCLUSION

This study finds 87 related pairs among 200 Swadesh meanings in Bulukumba Buginese and Toraja, giving a primary cognacy rate of 43.5%. Under the legacy lexicostatistical scale used in the study, this falls within the “family” category. A stricter sensitivity calculation that excludes 11 pairs supported only by phonetic similarity yields 38.0%, which remains within the same legacy band. The dataset also contains six correspondence sets, /ŋ~n/, /w~b/, /ə~a/, /r~d/, /ø~k/, and /o~e/, with unequal evidential strength, together with PAN-related examples of segmental replacement, deletion, and addition. The recalculated standard error is 0.0351, producing a model-dependent glottochronological range of approximately 2,097–1,740 years before 2026, or roughly 71 BCE–286 CE; this range is heuristic rather than a conventional statistical confidence interval. The central conclusion is therefore evidence-based but limited: BB and TL show substantial lexical and phonological historical relatedness in this dataset, while the numerical percentage and date estimate do not by themselves establish a precise genealogical subgrouping. The six-informant design strengthens lexical cross-checking, but broader speaker and locality

sampling, stricter cognate annotation, and wider South Sulawesi comparison are still required for a stronger genealogical claim.

REFERENCES

- Afria, R., Izar, J., Prawolo, I. S., & Arezky, B. (2020). Relasi bahasa Melayu Riau, Bugis, dan Banjar: Kajian linguistik historis komparatif. *Medan Makna: Jurnal Ilmu Kebahasaan dan Kesastraan*, 18(1), 94–106. doi:10.26499/mm.v18i1.2330
- Aisyah, S., & Widayati, D. (2022). Hubungan kekerabatan bahasa Pesisir Pasar, Kampung, dan Sorkam (Kajian linguistik historis komparatif). *Aksara: Jurnal Ilmu Pendidikan Nonformal*, 8(3), 2367–2376. doi:10.37905/aksara.8.3.2367-2376.2022
- Anderson, C., Tresoldi, T., Chacon, T., Fehn, A.-M., Walworth, M., Forkel, R., & List, J.-M. (2018). A cross-linguistic database of phonetic transcription systems. *Yearbook of the Poznan Linguistic Meeting*, 4, 21–53. doi:10.2478/yplm-2018-0002
- Batsuren, K., Bella, G., & Giunchiglia, F. (2022). A large and evolving cognate database. *Language Resources and Evaluation*, 56, 165–189. doi:10.1007/s10579-021-09544-6
- Blust, R., Trussel, S., & Smith, A. D. (2023). CLDF dataset derived from Blust’s “Austronesian Comparative Dictionary” (Version 1.2) [Data set]. Zenodo. doi:10.5281/zenodo.7741197
- Erniati. (2021). Klasifikasi leksikostatistik bahasa Banggai dan bahasa Hoti di Kabupaten Seram Bagian Timur. *Totobuang*, 9(2), 239–255. doi:10.26499/ttbng.v9i2.333
- Fadhila, A., Saddhono, K., & Suwandi, S. (2024). Kekerabatan bahasa Teringin dan bahasa Kampung Baru Kabupaten Kotawaringin Barat: Kajian linguistik historis komparatif. *Bahtera Indonesia: Jurnal Penelitian Bahasa dan Sastra Indonesia*, 9(2), 604–612. doi:10.31943/bi.v9i2.775
- Hadi, N. R. P., & Kusumaningrum, S. (2020). Korespondensi fonemis bahasa Ogan dan bahasa Bangka: Kajian linguistik historis komparatif. *Jurnal Culture*, 7(2), 191–203. doi:10.53873/culture.v7i2.222
- Jäger, G. (2018). Global-scale phylogenetic linguistic inference from lexical resources. *Scientific Data*, 5, 180189. doi:10.1038/sdata.2018.189
- Jäger, G. (2019). Computational historical linguistics. *Theoretical Linguistics*, 45(3–4), 151–182. doi:10.1515/tl-2019-0011
- Jannah, W., La Ino, & Zahrani. (2024). Kekerabatan bahasa Bugis dan bahasa Konjo Pesisir di Kabupaten Bulukumba (Linguistik bandingan historis). *Cakrawala Listra: Jurnal Kajian Sastra, Bahasa, dan Budaya Indonesia*, 7(1), 47–59. doi:10.33772/cakrawalalistra.v7i1.2724

- Lailiyah, N., & Wijayanti, F. I. (2022). Kekerabatan bahasa Jawa, Bali dan Bima: Perspektif linguistik historis komparatif. *Linguistik Indonesia*, 40(2), 327–345. doi:10.26499/li.v40i2.308
- List, J.-M., Greenhill, S. J., & Gray, R. D. (2017). The potential of automatic word comparison for historical linguistics. *PLOS ONE*, 12(1), e0170046. doi:10.1371/journal.pone.0170046
- Macklin-Cordes, J. L., Bowern, C., & Round, E. R. (2021). Phylogenetic signal in phonotactics. *Diachronica*, 38(2), 210–258. doi:10.1075/dia.20004.mac
- Mahendra, D., & Hendrokumoro. (2022). Relasi kekerabatan bahasa Sasak dan bahasa Banjar. *Deiksis*, 14(2), 125–141. doi:10.30998/deiksis.v14i2.10316
- Mahriyuni, Pramuniati, I., & Maftuhah, R. A. (2023). Lexicostatistics of Javanese and Sasak languages: Comparative historical linguistic studies. *Mimbar Ilmu*, 28(1), 124–130. doi:10.23887/mi.v28i1.59797
- Mayangsari, D. (2020). Leksikostatistik bahasa Bugis dan bahasa Toraja (Lexicostatistic of Bugis language and Toraja language). *Jalabahasa*, 16(1), 83–96. doi:10.36567/jalabahasa.v16i1.471
- Muhammad, S. R., & Hendrokumoro. (2022). Hubungan kekerabatan bahasa Aceh dan bahasa Gayo: Kajian linguistik historis komparatif. *Metahumaniora*, 12(2), 171–180. doi:10.24198/metahumaniora.v12i2.40891
- Padilla-Iglesias, C., Gjesfjeld, E., & Vinicius, L. (2020). Geographical and social isolation drive the evolution of Austronesian languages. *PLOS ONE*, 15(12), e0243171. doi:10.1371/journal.pone.0243171
- Rama, T., List, J.-M., Wahle, J., & Jäger, G. (2018). Are automatic methods for cognate detection good enough for phylogenetic reconstruction in historical linguistics? In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)* (pp. 393–400). Association for Computational Linguistics. doi:10.18653/v1/N18-2063
- Rizqi, F. A., & Widayati, D. (2021). Perubahan bunyi bahasa Proto Austronesia ke dalam bahasa Jawa dialek Sumatera (Kajian linguistik historis komparatif). *Kulturistik: Jurnal Bahasa dan Budaya*, 5(2), 29–35. doi:10.22225/kulturistik.5.2.3380
- Sholeha, M., & Hendrokumoro. (2022). Kekerabatan bahasa Kerinci, Melayu Jambi, dan Minangkabau. *Diglosia: Jurnal Kajian Bahasa, Sastra, dan Pengajarannya*, 5(2), 399–420. doi:10.30872/diglosia.v5i2.404
- Sulistiyarini, S., & Hendrokumoro. (2023). Hubungan kekerabatan bahasa Jawa, Sunda, dan Makassar: Kajian linguistik historis komparatif. *Madah: Jurnal Bahasa dan Sastra*, 14(2), 187–202. doi:10.31503/madah.v14i2.633

- Talaohu, A. R. (2019). The phonetics system in Buginese language. *KLAUSA (Kajian Linguistik, Pembelajaran Bahasa, dan Sastra)*, 3(1), 23–44. doi:10.33479/klaus.v3i01.146
- Utamayana, I. W. Y., Budasi, I. G., & Ramendra, D. P. (2025). Reflections of Proto-Austronesian in the Teta dialect. *International Journal of Language and Literature*, 8(1), 21–30. doi:10.23887/ijll.v8i1.67505
- Wahyudi, A. (2023). Struktur kalimat tunggal bahasa Bugis dialek Bulukumba. *Jurnal Sarjana Ilmu Budaya*, 3(2), 1–13. doi:10.12259/jsib.v3i2.25543