Jurnal Pustaka Indonesia (JPI)
2024, Vol. 1, No. 1, 46 – 53
https://doi.org/ 10.62159/jpi.v1i1.392

OPEN ACCESS
Research Article

Quantifying Dialect Relatedness in Serawai (Bengkulu, Indonesia): A 200-Item Lexicostatistical Study

Dewi Ayu Lestari[1], Ali Akbarjono[2], Meddyan Heriadi[3]
[1,2,3] Universitas Islam Negeri Fatmawati Sukarno Bengkulu, Indonesia

ABSTRACT

This study quantifies the internal relatedness of the Serawai language in Bengkulu Province by comparing two named varieties Padang Capo ("o") and Puding ("au") and assessing whether they constitute dialects of a single language. Using a qualitative–quantitative lexicostatistical design, we elicited a complete 200-item Swadesh list from adult native speakers at both sites, normalised tokens, and applied conservative, rule-governed cognacy coding supplemented by analyst memos and a second-reader check; we also summarised recurrent phonological correspondences to interpret the numeric signal. Of 200 aligned items, 124 were cognate, yielding 62% lexical relatedness, with differences concentrated in limited, patterned vocalic and segmental alternations characteristic of dialectal separation. Interpreted against commonly used lexical-similarity bands and a standard basic-vocabulary retention model, the evidence situates the varieties' most recent common stage at roughly the last 1.0–1.3 millennia, a heuristic window consistent with high mutual intelligibility. The findings support classifying "o" and "au" as close dialects within one Serawai language and, methodologically, convert impressionistic labels into a replicable baseline (complete elicitation, explicit coding rules, qualitative correspondence notes, and uncertainty-aware reporting). Implications include treating Serawai as a single language for educational materials, orthographic guidance, and public communication, and using the documented workflow to expand coverage across additional villages and neighbouring Malayic varieties; future work should add formal correspondence tables with instrumental phonetics and integrate sociolinguistic profiling and phylogenetic modelling to refine internal subgrouping.

ARTICLE HISTORY
Received: 2 February 2024
Revised: 7 March 2024
Accepted: 8 April 2024

KEYWORDS
Bengkulu; Dialectology; Historical linguistics; Lexicostatistics.

CORRESPONDING AUTHOR
Dewi Ayu Lestari, Universitas Islam Negeri Fatmawati Sukarno Bengkulu, Indonesia. Email: dewiayulestari1310@gmail.com

Introduction

Documenting and comparing regional Austronesian varieties in Indonesia remains an urgent task for both linguistic vitality and historical classification within the Malayic subgroup in Sumatra. Recent scholarship shows that closely related varieties may diverge in phonology, morphophonology, and lexicon in ways that are typologically instructive and socially consequential, shaping local intelligibility, identity, and language policy (Adelaar, 2017; Ernanda, 2021; Lauder et al., 2021). Within Malayic, recurrent correspondences in final vowels and

consonants interact systematically with lexical identity across speech communities, a pattern that complicates dialect boundaries and underscores the need for transparent, replicable measures of relatedness (Adelaar, 2017; Ernanda, 2021; Lauder et al., 2021). Establishing such baselines is not only of theoretical interest; it also informs culturally responsive education and public communication in local languages during periods of rapid social change (Adelaar, 2017; Ernanda, 2021; Lauder et al., 2021).

To adjudicate fine-scale relationships among neighbouring isolects, researchers continue to rely on lexicostatistics operationalised through a Swadesh list and explicit cognacy rules as a practical starting point for estimating lexical proximity and for generating first-pass divergence chronologies under clearly stated assumptions. Methodological reviews emphasise pairing quantitative similarity measures with qualitative correspondence analysis so that percentage scores are interpretable in light of regular sound change (Bowern, 2018; Sagart et al., 2019; Zhang et al., 2020). Indonesian studies echo this combined strategy, showing that conservative cognacy coding over a 200-item Swadesh list yields reproducible results while correspondence tables capture phonological structure behind the numbers (Dardanila et al., 2018; Saddhono & Hartanto, 2021; Zhang et al., 2020). Although debates persist over retention-rate calibration and loanword handling, a carefully documented pipeline from elicitation to normalisation, coding, and reporting provides a transparent baseline that later work can refine with denser datasets and complementary models (Bowern, 2018; Sagart et al., 2019; Zhang et al., 2020).

Against this backdrop, the Serawai language of Bengkulu Province commonly described by speakers as comprising two named varieties, the “o” and the “au” dialects offers a compelling test bed for a combined dialectological–lexicostatistical approach. Descriptive accounts and neighbouring Malayic comparisons suggest recurrent vowel correspondences and segmental alternations at morpheme edges that plausibly index historical splits while maintaining high mutual intelligibility (Adelaar, 2017; Ernanda, 2021; Lauder et al., 2021). Yet claims about the Serawai pair remain largely impressionistic without a percentage-based similarity measure anchored in field data from distinct communities, accompanied by a qualitative inventory of regular correspondences that justify cognacy decisions. In parallel, Indonesian applications of dialectometry and lexicostatistics in other Sumatran groupings demonstrate that site-balanced sampling and explicit coding can move debates from anecdote to replicable metrics (Dardanila et al., 2018; Saddhono & Hartanto, 2021; Zhang et al., 2020).

This article addresses these needs for Serawai by applying a 200-item Swadesh instrument to two field sites representing the named poles Padang Capo (Seluma; “o”) and Puding (South Bengkulu; “au”) and by reporting (i) percentage cognacy under conservative, rule-based coding; (ii) a coarse divergence estimate based on established retention constants with uncertainty bounds; and (iii) qualitative correspondences that structure the cognacy space. In doing so, we contribute a replicable, data-driven baseline for Serawai’s intralanguage stratification and situate the variety more precisely within the Malayic continuum. The research gap we address is threefold: the 2018–2021 literature lacks a Serawai-specific, single-study report that (a) quantifies “o”–“au” lexical similarity with a 200-item list; (b) inventories regular segmental correspondences that underpin cognacy; and (c) models an approximate time depth under transparent assumptions. Accordingly, our objective is to provide a methodologically explicit, field-anchored account of Serawai relatedness that can be extended with larger samples,

phonological correspondences, and sociolinguistic variables in future work (Bowern, 2018; Sagart et al., 2019; Zhang et al., 2020).

Methods

This study adopted a qualitative–quantitative lexicostatistical design to gauge lexical relatedness between two named varieties of Serawai commonly referred to as the "o" and "au" dialects represented by field sites in Padang Capo (Seluma) and Puding (South Bengkulu). Data were elicited with a 200-item Swadesh basic-vocabulary instrument administered to native adult speakers with continuous residence since childhood and no reported speech or hearing impairments. Local community leaders assisted in identifying consultants, sessions were conducted in quiet settings, and follow-up visits were scheduled when clarification was required.

Each elicited item was first recorded in practical orthography and, where needed, supported by narrow International Phonetic Alphabet (IPA) notes to enable transparent cross-site comparison. Tokens were normalised to minimise orthographic noise for example, enforcing consistent final-vowel marking and then aligned item-by-item by Swadesh gloss. For every alignment we added memo fields documenting semantic scope, putative etymology, and any elicitation notes likely to affect cognacy decisions.

Cognacy was judged using conservative, rule-based criteria. Two forms were coded as cognate only when regular segmental correspondences could be posited and the core meaning remained stable across sites. Obvious loans (e.g., high-frequency Indonesian/Malay borrowings) and items exhibiting clear semantic drift were flagged and excluded from the cognate count, though retained in the supplementary list for transparency. Ambiguous cases were provisionally marked and revisited after a first coding pass; in all such instances, consistency with emergent correspondence sets was prioritised over isolated surface similarity.

To enhance analytic trustworthiness, we maintained an audit trail comprising the original elicitation sheets, a versioned codebook of decision rules (with positive and negative exemplars), and line-by-line analyst memos. A second analyst independently reviewed a subset of items; disagreements were resolved through negotiated consensus rather than majority vote, preserving phonological plausibility in a small-N setting. Quantitative tabulation was performed in a reproducible spreadsheet with cross-checks for entry errors, and coverage rates are reported alongside the results to indicate the effective denominator after exclusions.

The primary outcome was lexical relatedness expressed as the percentage of cognate items over the 200-word list, computed after excluding missing or excluded items to avoid denominator inflation. For a coarse time-depth inference, we converted the observed cognacy proportion into an approximate divergence estimate using a standard retention-rate model for basic vocabulary, adopting a conservative per-millennium retention constant $L = 0.81$ and bracketing uncertainty by ± 0.02 to reflect calibration debates. We report the central estimate with a simple sensitivity interval rather than point claims, and we complement the numeric outputs with qualitative summaries of recurrent sound correspondences (e.g., vowel alternations at morpheme edges) that validate cognacy assignments. Because the research involved non-intervention linguistic elicitation with adult volunteers and no personal identifiers, formal human-subjects review was not required. Participation was voluntary, informed orally, and

consultants could decline any item. All materials necessary for replication the 200-item aligned list, correspondence notes, and the coding rulebook are available from the author upon reasonable request.

Results and Discussion

Elicitation coverage and data quality

We elicited the full 200-item Swadesh list from both Serawai field sites Padang Capo ("o") and Puding ("au") with complete coverage and no missing items. Tokens were normalised and, where needed, annotated with narrow IPA notes to secure transparent cross-site comparison. Ambiguities were logged in analyst memos and adjudicated through negotiated consensus; a second analyst's spot checks confirmed category stability in the final coding sheet. Because loan items and clear semantic drifts were excluded by design from the cognate set, the resulting percentages reflect conservative, rule-governed judgments.

Lexical relatedness (cognacy yield)

Out of 200 aligned lexical items, 124 were judged cognate across the two sites, yielding 62% lexical relatedness. This magnitude is consistent with dialectal differentiation within a single language rather than a split into distinct languages and coheres with local reports of high mutual intelligibility.

Table 1. Lexical relatedness between Serawai "o" (Padang Capo) and "au" (Puding) based on a 200-item Swadesh list

Item count	Cognate items	Cognate %	Interpretation
200	**124**	**62%**	Close dialects within one language

Note: A hypothetical reclassification of ±3 borderline items would shift the cognacy percentage by only ±1.5 points, leaving the qualitative interpretation unchanged.

To contextualise the 62% lexical relatedness, we situate the Serawai pair within commonly used lexical-similarity bands; as shown in Table 2, the "o–au" pair falls into the Family (close dialects**)** range.

Table 2. *Lexical-Similarity Bands Used for Interpretation and Placement of the Serawai "o–au" Pair*

Level	Lexical Similarity Band (%)	Position of o-au
Language	100-81	
Family (close dialects)	81-36	✓
Stock	36-12	
Microphylum	12-4	
Mesophylum	4-1	
Macrophylum	1 to < 1	

Qualitative correspondence patterns

Qualitative inspection shows that differences concentrate in a limited set of vocalic alternations and stable segmental correspondences characteristic of dialectal (rather than genealogical) separation. Alternations typically occur at morpheme edges and in salient syllables, while core lexical skeletons remain recognisably parallel. These patterns substantiate the conservative cognacy decisions and explain how high overall similarity coexists with audible variety-specific colouring in everyday speech.

Time-depth Estime and Sensitivity

Using a standard basic-vocabulary retention model for closely related varieties, the observed cognacy of 62% corresponds to an estimated divergence a little over one millennium. With a commonly used per-millennium retention constant of 0.81, the estimate is ≈ 1,130 years before present. Because retention calibration varies across studies, we report a simple sensitivity range; across plausible constants from 0.79 to 0.83, the estimate spans ≈ 1,010 to 1,280 years. The rank-order inference is stable across this range: the two varieties remain close kin within Serawai.

Table 3. Sensitivity of divergence time to the retention constant (two-branch model; observed cognacy 62%)

Retention per millennium	Estimated years since split
0.79	1,014
0.80	1,071
0.81	1,134
0.82	1,204
0.83	1,283

Summary of main findings

The Serawai varieties sampled in Padang Capo (“o”) and Puding (“au”) exhibit high lexical affinity (62%) alongside limited, patterned phonological alternation, supporting their status as dialects of a single language. A coarse time-depth estimate situates their most recent common stage within the last 1.0–1.3 millennia. Together, the quantitative percentage and qualitative correspondences provide a replicable baseline for future classification, inviting extensions with multi-site sampling, explicit correspondence tables, and sociolinguistic profiling.

Discussion

The 62% lexical relatedness observed between the Serawai “o” (Padang Capo) and “au” (Puding) varieties is consistent with regional findings that geographically contiguous Malayic isolects preserve high basic-vocabulary overlap while differing through patterned segmental alternations, a profile typical of dialectal rather than language-level separation. Comparable outcomes emerge when lexicostatistics is paired with qualitative correspondence analysis in Sumatra: studies that juxtapose neighboring Malayic varieties (e.g., Mandailing–Malay) or cross-cluster pairs (e.g., Gayo–Mandailing) likewise report “close-kin” classifications once loans are controlled and cognacy rules are applied conservatively (Adelaar, 2018; Ogloblin, 2018; Steinhauer, 2018). Beyond Sumatra, dialectometric work on Javanese isolects demonstrates that

contiguous speech zones can be quantitatively distinguished yet remain mutually intelligible, reinforcing the value of replicable metrics over impressionistic labels (Saddhono & Hartanto, 2021). Descriptive and historical analyses further emphasize that Malay/Indonesian varieties often diverge at the level of morphophonology especially in vowel behavior at morpheme edges without wholesale lexical replacement, a pattern that aligns with our qualitative correspondence notes for Serawai (Ernanda, 2021; Hoogervorst, 2018; Shin, 2021). Methodologically, the study's workflow full 200-item elicitation, rule-governed cognacy coding, and uncertainty-aware reporting accords with best-practice recommendations that numerical similarity should be interpreted alongside regular sound correspondences and with transparent assumptions about retention (Dardanila et al., 2018; Saddhono & Hartanto, 2021; Zhang et al., 2020).

The study's novelty lies in providing a Serawai-specific, site-balanced baseline that (i) quantifies "o–au" relatedness with a complete Swadesh instrument; (ii) justifies cognacy through an analyst-audited codebook plus qualitative correspondence summaries; and (iii) reports a divergence window using sensitivity to plausible retention rates rather than a single-point date. Within the 2018–2021 literature, Serawai's internal split had not been treated with this combined percentage-plus-correspondence protocol grounded in field data from two distinct communities, so our contribution moves the discussion from anecdotal claims to a documented evidence stack (Adelaar, 2021; Jeszenszky et al., 2021; Yannuar, 2020).

The implications are both applied and theoretical. For language maintenance and planning, a conservative 62% relatedness supports treating "o" and "au" as dialects of one Serawai language in educational materials, public communication, and orthographic guidance, resonating with broader calls to empower local languages during periods of social disruption. For research, the baseline facilitates finer-grained mapping of the Serawai continuum and enables triangulation with phylogenetic and dialectometric models increasingly used to embed micro-level similarity within macro-level histories (Dunn, 2019; Jeszenszky et al., 2021; Klamer, 2019). Operationally, the pipeline we document elicitation, normalization, rule-based coding, and sensitivity-aware interpretation offers a reusable template for expanding coverage to additional Serawai villages and for comparative work across Bengkulu's Malayic neighbors.

Several limitations bound interpretation. The analysis draws on two sites and a single elicitation window, so intra-dialect heterogeneity and interspeaker variation may be underrepresented. Lexicostatistics focuses on basic vocabulary and can miss structural innovations in morphology and phonology; while we mitigate this by summarizing correspondence patterns, future work should add formal sound-correspondence tables and instrumental phonetics. Cognacy coding retains a degree of subjectivity even under conservative rules and double review, and divergence estimates depend on debated retention constants; our sensitivity range is intended to communicate order of magnitude rather than a precise chronology. Extending the sample, integrating sociolinguistic profiling, and exploring Bayesian phylogenetic models to combine lexical counts with correspondence constraints would strengthen the external validity of Serawai classification (Hoffmann et al., 2021; Neureiter et al., 2021; Suchard et al., 2018)

Conclusion

This study provides a replicable, field-anchored account of Serawai internal differentiation by showing that the Padang Capo ("o") and Puding ("au") varieties share 62% basic-lexicon cognacy and differ primarily through limited, patterned vocalic and segmental alternations evidence consistent with close dialects within a single language rather than separate languages. Interpreted against standard lexical-similarity bands and a conservative retention model, the findings place the varieties' most recent common stage at roughly the last 1.0–1.3 millennia, a heuristic timeframe that contextualizes their high mutual intelligibility without claiming precise chronology. Beyond the specific figures, the contribution lies in converting impressionistic labels into transparent metrics complete 200-item elicitation, rule-governed cognacy coding, qualitative correspondence notes, and uncertainty-aware reporting thereby establishing a baseline for subsequent classification of the Serawai continuum. Practically, the results support treating "o" and "au" as a single language for educational materials, orthographic guidance, and public communication, while theoretically they motivate expanded comparative work across Bengkulu's Malayic neighbors. Future extensions should broaden site coverage, formalize sound-correspondence tables with instrumental phonetics, and integrate sociolinguistic profiling and phylogenetic modeling to refine internal subgrouping; nevertheless, the present analysis already furnishes a robust platform for evidence-based language description and maintenance.

Reference

Adelaar, A. (2017). Dialects of Malay/Indonesian. In *The Handbook of Dialectology* (pp. 571–581). Wiley. https://doi.org/10.1002/9781118827628.ch36

Adelaar, A. (2018). Regular sound change; The evidence of a single example. *Wacana, 19*(2), 408. https://doi.org/10.17510/wacana.v19i2.703

Adelaar, A. (2021). South Borneo as an ancient Sprachbund area. *Wacana, 22*(1), 81. https://doi.org/10.17510/wacana.v22i1.963

Bowern, C. (2018). Computational Phylogenetics. *Annual Review of Linguistics, 4*(1), 281–296. https://doi.org/10.1146/annurev-linguistics-011516-034142

Dardanila, ., Mulyadi, ., & Tantawi, I. (2018). Lexicostatistics of Gayo Language with Mandailing Language. *Proceedings of the International Conference of Science, Technology, Engineering, Environmental and Ramification Researches*, 1199–1203. https://doi.org/10.5220/0010069511991203

Dunn, J. (2019). Global Syntactic Variation in Seven Languages: Toward a Computational Dialectology. *Frontiers in Artificial Intelligence, 2*. https://doi.org/10.3389/frai.2019.00015

Ernanda, E. (2021). Some notes on the Semerap dialect of Kerinci and its historical development. *Wacana, 22*(1), 58. https://doi.org/10.17510/wacana.v22i1.978

Hoffmann, K., Bouckaert, R., Greenhill, S. J., & Kühnert, D. (2021). Bayesian phylogenetic analysis of linguistic data using BEAST. *Journal of Language Evolution, 6*(2), 119–135. https://doi.org/10.1093/jole/lzab005

Hoogervorst, T. G. (2018). Utterance-final particles in Klang Valley Malay. *Wacana, 19*(2), 291. https://doi.org/10.17510/wacana.v19i2.704

Jeszenszky, P., Steiner, C., & Leemann, A. (2021). Reduction of Survey Sites in Dialectology: A New Methodology Based on Clustering. *Frontiers in Artificial Intelligence, 4*. https://doi.org/10.3389/frai.2021.642505

Klamer, M. (2019). The dispersal of Austronesian languages in Island South East Asia: Current findings and debates. *Language and Linguistics Compass, 13*(4). https://doi.org/10.1111/lnc3.12325

Lauder, A. F., Lauder, M. R., & Kiftiawati, K. (2021). Preserving and empowering local languages amidst the Covid-19 pandemic; Lessons from East Kalimantan. *Wacana, 22*(2), 439. https://doi.org/10.17510/wacana.v22i2.1006

Neureiter, N., Ranacher, P., van Gijn, R., Bickel, B., & Weibel, R. (2021). Can Bayesian phylogeography reconstruct migrations and expansions in linguistic evolution? *Royal Society Open Science, 8*(1), 201079. https://doi.org/10.1098/rsos.201079

Ogloblin, A. (2018). Notes on structural distinctions in Malay dialects. *Wacana, 19*(2), 327. https://doi.org/10.17510/wacana.v19i2.706

Saddhono, K., & Hartanto, W. (2021). A dialect geography in Yogyakarta-Surakarta isolect in Wedi District: An examination of permutation and phonological dialectometry as an endeavor to preserve Javanese language in Indonesia. *Heliyon, 7*(7), e07660. https://doi.org/10.1016/j.heliyon.2021.e07660

Sagart, L., Jacques, G., Lai, Y., Ryder, R. J., Thouzeau, V., Greenhill, S. J., & List, J.-M. (2019). Dated language phylogenies shed light on the ancestry of Sino-Tibetan. *Proceedings of the National Academy of Sciences, 116*(21), 10317–10322. https://doi.org/10.1073/pnas.1817972116

Shin, C. (2021). Iban as a koine language in Sarawak. *Wacana, 22*(1), 102. https://doi.org/10.17510/wacana.v22i1.985

Steinhauer, H. (2018). Sound-changes and loanwords in Sungai Penuh Kerinci. *Wacana, 19*(2), 375. https://doi.org/10.17510/wacana.v19i2.708

Suchard, M. A., Lemey, P., Baele, G., Ayres, D. L., Drummond, A. J., & Rambaut, A. (2018). Bayesian phylogenetic and phylodynamic data integration using BEAST 1.10. *Virus Evolution, 4*(1). https://doi.org/10.1093/ve/vey016

Yannuar, N. (2020). Bòsò Walikan Malangan; Structure and development of a Javanese reversed language. *Wacana, 21*(1), 168. https://doi.org/10.17510/wacana.v21i1.879

Zhang, H., Ji, T., Pagel, M., & Mace, R. (2020). Dated phylogeny suggests early Neolithic origin of Sino-Tibetan languages. *Scientific Reports, 10*(1), 20792. https://doi.org/10.1038/s41598-020-77404-4