Linguistik Indonesia, Februari 2026, 111-131 *Volume ke-44, No. 1*

ISSN cetak 0215-4846; ISSN online 2580-2429

WHEN LEXICAL DATA SPEAK: REASSESSING THE GENETIC CLASSIFICATION OF INDONESIA'S REGIONAL LANGUAGES

Ernanda[1], Ulil Amri[2]
Universitas Jambi[1,2]
ernanda@unja.ac.id[1], ulil.ludostrait@unja.ac.id[2]

Abstract

This study analyzes the lexical similarity of ten regional languages in Indonesia, namely Jambi Malay (**jax**), Kerinci (**kvr**), Minangkabau (**min**), Banjar (**bjn**), Mentawai (**mwv**), Sasak (**sas**), Javanese (**jav**), Toba (**tob**), Angkola (**akb**), and Mandailing (**btm**), along with Indonesian (**ind**) as a lingua franca; the genetic status of the regional languages; and the separation times among the regional languages. Data were collected through field observations at ten locations, with three informants per language, gathering 257 glosses from core (L1), nature (L2), general (L3), and cultural (L4) vocabulary. The analysis was conducted in three stages: first, synchronic lexical similarity was calculated using the Jaccard method; second, genetic relationships were analyzed through lexicostatistics based on L1 and L2; third, glottochronology was used to estimate language separation times among the regional languages. The results indicate that no language pairs share high similarity; most fall into the low-to-moderate similarity category. Lexicostatistical analysis reveals that core languages (**jax**, **kvr**, **min**, and **bjn)** and peripheral languages **(tob**, **akb**, and **btm)** form distinct genetic families, while **mwv** is the most lexically isolated language. Estimates of separation times indicate that core languages have a more recent lineage, while other languages such as **mwv**, **jav**, and **tob** show earlier divergence periods. These findings confirm that geographic proximity does not always correlate with linguistic relationships and suggest the need to revise the classification of Indonesian languages available in online databases, particularly the position of **mwv**, which should be reclassified as part of the Barrier Island language group rather than part of the Sumatran group. This study also highlights the importance of using primary data in language documentation to provide a more accurate map of linguistic evolution for regional languages in the archipelago.

Keywords: glottochronology, genetic classification, historical comparative linguistics, lexical similarity

Abstrak

*Penelitian ini menganalisis kesamaan leksikal pada 10 bahasa daerah di Indonesia, yaitu Melayu Jambi (**jax**), Kerinci (**kvr**), Minangkabau (**min**), Banjar (**bjn**), Mentawai (**mwv**), Sasak (**sas**), Jawa (**jav**), Toba (**tob**), Angkola (**akb**), dan Mandailing (**btm**), serta Bahasa Indonesia (*ind*) sebagai lingua franca; status genetik bahasa daerah; dan waktu pisah antar bahasa daerah. Data lapangan dikumpulkan pada 10 titik pengamatan dengan tiga orang informan untuk masing-masing bahasa daerah . Peneliti mengumpulkan 257 glosa yang terdiri dari kosakata inti (L1), alami (L2), umum (L3), dan budaya (L4). Analisis dilakukan dalam tiga tahap: pertama, menentukan kesamaan leksikal sinkronis yang dihitung melalui perhitungan Jaccard; kedua, menentukan hubungan genetik dengan metode leksikostatistik berdasarkan L1 dan L2; ketiga, glotokronologi diaplikasikan untuk memperkirakan waktu pisah antar bahasa daerah. Hasil penelitian menunjukkan bahwa tidak ada pasangan bahasa yang memiliki kemiripan tinggi; sebagian besar termasuk dalam kategori kemiripan rendah*

*hingga sedang. Analisis leksikostatistik mengungkapkan bahwa bahasa-bahasa kluster inti (**jax**, **kvr**, **min**, dan **bjn**) dan kluster periferal (**tob**, **akb**, dan **bjn**) merupakan bahasa-bahasa dengan genetik yang relatif dekat, sedangkan **mwv** merupakan bahasa yang paling terisolasi secara leksikal. Perkiraan waktu pisah menunjukkan bahwa bahasa-bahasa kluster inti dan periferal memiliki kekerabatan yang lebih dekat, sementara bahasa lain seperti **mwv**, **jav**, dan **tob** menunjukkan waktu pisah yang lebih awal. Temuan ini menegaskan bahwa kedekatan geografis tidak selalu berkorelasi dengan hubungan linguistik dan menunjukkan perlunya merevisi klasifikasi bahasa-bahasa daerah di Indonesia yang tersedia dalam basis data daring, khususnya posisi **mwv**, yang seharusnya masuk ke dalam kelompok bahasa kepulauan luar dan bukan bagian dari kelompok Bahasa Sumatera. Studi ini juga menyoroti pentingnya penggunaan data primer dalam dokumentasi bahasa untuk memberikan peta evolusi linguistik yang lebih akurat untuk bahasa-bahasa daerah di kepulauan ini.*

Kata kunci: *glotokronologi, klasifikasi genetik, linguistik historis komparatif, persamaan leksikal*

INTRODUCTION

Language documentation and research today are greatly assisted by several online databases of world languages such as Ethnologue (Eberhard et al., 2025) which classifies the taxonomy of language families and the Austronesian Basic Vocabulary Database/ABVD (Greenhill et al., 2025a) which provides linguistic data. However, these databases cannot be entirely relied upon as primary sources, as they generally depend on linguists, researchers, fieldworkers, and community contributors around the world who provide language data, updates, published works (books, articles, and reports), and unpublished sources to compile entries (Campbell & Grondona, 2008; Collin, 2010; Cornwell, 2019; Greenhill et al., 2008; Hammarström, 2015; Hugh, 2012; Paolillo & Das, 2006). In addition, some of the available data in these online databases are outdated; for instance, the Mentawai data date from 1992 (Greenhill et al., 2025b). Campbell & Grondona (2008) and Cornwell (2019) claim that Ethnologue is considered problematic because it is published by an institution outside of academic linguistic organizations. The curators of the ABVD themselves also noted that the database is potentially inconsistent in its orthography and phoneme representation, which can interfere with the interpretation of lexical forms (Greenhill et al., 2008). Given these limitations, it is crucial to directly analyze primary linguistic data. This is especially important for understanding language relationships such as those in the Indonesian Archipelago, which has a complex linguistic landscape.

The majority of regional languages in Indonesia are classified as Malayo-Polynesian (Adelaar, 2011; Adelaar & Schapper, 2024; Blust, 2013, 2015; Ross, 1996; Sneddon, 2003). Moreover, in Indonesia, areal proximity is frequently associated with indications of shared cultural and linguistic ancestry. Theoretically, geographic closeness increases the likelihood of linguistic convergence through long-term cultural contact (Casasanto, 2008; Cavalli-Sforza, 1997; Gudschinsky, 1956; Huisman et al., 2021). However, notable exceptions exist, such as the case of the Mentawai and Minangkabau languages, which are geographically close but historically and phonologically distinct (Billings & McDonnell, 2024; Edwards, 2015; Ermanto, 2025; Ermanto & Emidar, 2018). This phenomenon highlights the significance of the current study, which provides empirical evidence that areal proximity does not necessarily correspond to linguistic relatedness.

Previous studies have generally examined only two or three languages, and only in a limited scope. Some of these include studies of the Javanese-Sasak relationship (Mahriyuni et al., 2023), the Mentawai dialects (Budiono et al., 2023), the Banjar-Malay relationship (Afria et al., 2020; Wahab & Halin, 2021), the Patani-Minangkabau relationship (Nalee et al., 2020), the Mandailing-Toba relationship (Dewanti & Zainuddin, 2024), the Minangkabau-Banjar relationship (Amri et al., 2024), the Minangkabau-Batak Toba relationship (Amri, Sirait, et al., 2025), and the Kerinci-Jambi-Minangkabau relationship (Sholeha & Hendrokumoro, 2022). One of the few notable studies that has analyzed more than three Indonesian regional languages was conducted by Meliana et al. (2024), who traced the linguistic roots of Malay and Batak languages in Sumatra, namely the Rejang, Serawai, Lembak, Toba, Mandailing, and Nias languages.

The small number of languages examined in these studies limits our insight into the evolutionary dynamics of sound change. Without extensive comparative data, historical connections between languages often appear unclear due to gaps in their phonemic correspondences. The present study analyzes ten Indonesian regional languages: Jambi Malay (**jax**), Kerinci (**kvr**), Minangkabau (**min**), Banjar (**bjn**), Mentawai (**mwv**), Sasak (**sas**), Javanese (**jav**), Toba (**tob**), Angkola (**akb**), and Mandailing (**btm**), including Indonesian (**ind**)[i] as the *lingua franca* of the Indonesian archipelago. By analyzing multiple languages, the lexical and phonological correspondences across the dataset can be traced more accurately.

In two regional languages in Indonesia, for example, the lexeme for ‘black’ is realized as [ita] and [birəŋ], which seem to be derived from two different etymons. However, when the analysis is extended to multiple languages, the lexeme for ′black’ is found in several forms, i.e., [hitam], [ita], [hiraŋ], [birəŋ], [irəŋ], [biroŋ], [kəla], [kaljaŋ], [posu], [pusuʔ], [maposu], and [nalomlom]. With a more extensive dataset, we can see that several words share the same etymon; for instance, the series [hitam] ~ [ita] ~ [hiraŋ] ~ [birəŋ] ~ [irəŋ] ~ [biroŋ] exhibits clear phonemic correspondences, such as [h] ~ [b] ~ [ø]; [t] ~ [r]; [a] ~ [ə] ~ [o]; and [m] ~ [ŋ] > [ita] ~ [birəŋ]. Further complicating this analysis, other sets of forms [kəla] ~ [kaljaŋ], [posu] ~ [pusuʔ] ~ [maposu], and [nalomlom], which also denote the same meaning, demonstrate that the concept of ‘black’ is represented by multiple lexical variations. By incorporating data from multiple languages, this study is able to identify phonetic transitions and determine kinship relations more accurately than analyses based on only two languages.

These objectives are pursued through the application of lexicostatistics and glottochronology. Lexicostatistics serves as a quantitative instrument to determine the degree of language relatedness by calculating the percentage of cognate words. Scientifically, the use of this method is grounded in Gudschinsky’s (1956) assumptions: first, every language has a stable core vocabulary (including pronouns, numbers, and body parts) that is less prone to change than general vocabulary; thus, it is more suitable for tracing the relationships and history of languages; second, the core vocabulary remains at the same percentage every thousand years; third, the basic vocabulary is assumed to be lost at approximately the same rate across all languages; and fourth, the percentage of cognates is used to estimate the point at which two languages diverged from their parent language.

Regarding the stability of core vocabulary, in some languages, some core vocabulary items are represented by more than one etymon, e.g., the first-person singular pronoun in Indonesian is realized as both [aku] and [saja]. The classical lexicostatistical formula may

yield ambiguous results if we follow the binary rule of one etymon per gloss. To resolve this, this study integrates the lexicostatistic formula with Jaccard's statistical calculation (Bag et al., 2019; Kamiura & Sekine, 2023; Majumdar, 2025) to identify lexical similarities among modern languages.

In the subsequent stage, this study applies glottochronology to convert the percentage of cognates into estimates of language divergence periods (Kroeber, 1955). Assuming that basic vocabulary undergoes change at a constant average rate (Petroni & Serva, 2010), the resulting lexical distances are considered proportional to the time depth of separation between the languages.

Based on the foregoing discussion, the study aims to (1) determine the level of lexical similarity among eleven languages in Indonesia, (2) identify the genetic classification of ten regional languages based on their lexical distribution, and (3) calculate the estimated separation time of each regional language using glottochronological analysis. To achieve these objectives, the paper is organized as follows. The profiles of the languages studied are presented in the next section, followed by a section on research methodology. Finally, the results and discussion are presented, and conclusions are drawn in the last section.

LANGUAGE PROFILES

The Indonesian archipelago is home to rich linguistic diversity (Klamer, 2018; Zein, 2020), offering a remarkable data source for language documentation and linguistic inquiry. In particular, Indonesia is the point of origin for various Austronesian languages with complex kinship relationships. This research focuses on Indonesian (**ind**), the national language, alongside ten regional languages, i.e., Jambi Malay (**jax**), Kerinci (**kvr**), Minangkabau (**min**), Banjar (**bjn**), Mentawai (**mwv**), Javanese (**jav**), Sasak (**sas**), Toba (**tob**), Angkola (**akb**), and Mandailing (**btm**).

Indonesian (**ind**) functions as a unifying national language (*lingua franca*) for speakers of over 700 distinct regional languages across the country. Based on Blust's (2013) work, the inclusion of Indonesian is regarded as essential because of its historical role in facilitating cultural and lexical exchange. In this study, Jambi Malay (**jax**) refers to the language spoken in Jambi Province, excluding the varieties in Kerinci Regency and Sungai Penuh City (Ernanda et al., 2025). Previous linguistic analyses of the Jambi Malay language have been conducted by researchers (Anderbeck, 2008; Sholeha & Hendrokumoro, 2022), even though such analyses have been limited to two or three languages (Anderbeck, 2008; Sholeha, 2022). Therefore, this study takes Kumpeh as its observation point.

The Kerinci language (**kvr**) is spoken in Kerinci Regency and Sungai Penuh[ii] City, located within Jambi Province. The Kerinci language has been studied through various linguistic perspectives (Endriani et al., 2023, 2023; Ernanda, 2017, 2020, 2021; Ernanda & Yap, 2024; Fatria et al., 2023; Harmedianti et al., 2023; Steinhauer, 2018). Kerinci language varieties, especially those spoken in the city center, are currently being marginalized by more dominant languages, including Minangkabau (Ernanda, 2015, 2018; Wilymafidini, 2017). The Minangkabau language (**min**) in West Sumatra[iii] exhibits a wide geographic distribution due to the tradition of migration (Naim, 2013; Rahman, 2016), with an estimated seven million speakers (Crouch, 2009). A number of previous studies have examined the Minangkabau language (Amri, 2017, 2022; Amri et al., 2020, 2024, 2024; Amri, Husna, et al., 2025; Amri, Sirait, et al., 2025).

The Banjar language (**bjn**) is widely spoken as a mother tongue in South Kalimantan (Wahab & Halin, 2021), primarily in Hulu Sungai Utara, Hulu Sungai Selatan, Hulu Sungai Tengah Barito Kuala, Tapin, Kotabaru, Banjar Regency[iv], Balangan, Banjarmasin, Tanah Laut, and Tabalong, inter alia (Badan Pengembangan dan Pembinaan Bahasa, 2025). It is also spoken in Central and East Kalimantan (Kawi, 1991). Banjar comprises two main dialects, namely Pahuluan and Kuala (Farid, 2012). In contrast, the Mentawai language (**mwv**), spoken in West Sumatra, shows a more localized pattern of dialects. Previously Mentawai was classified into four dialects (Lenggang et al., 1978), but recent studies suggest that there are now only two main dialects: Siberut and Sipora Pagai (Budiono et al., 2023). A number of previous studies have examined the Mentawai language (Febrina, 2014; Pratiwi et al., 2025). Field observations for this study were carried out across four major islands: Siberut, Sipora, North Pagai, and South Pagai.

The Sasak language (**sas**) of Lombok [v] presents a complex case for linguistic classification, currently serving a community of approximately three million speakers (Bellwood, 2024). Although Sasak was traditionally classified within the Bali-Sasak-Sumbawa subgroup by Adelaar (2004), Burhanuddin et al. (2025) claim that the significant non-inherited lexical innovations were recently identified, arguing that it separated from its Austronesian ancestor at an earlier period.

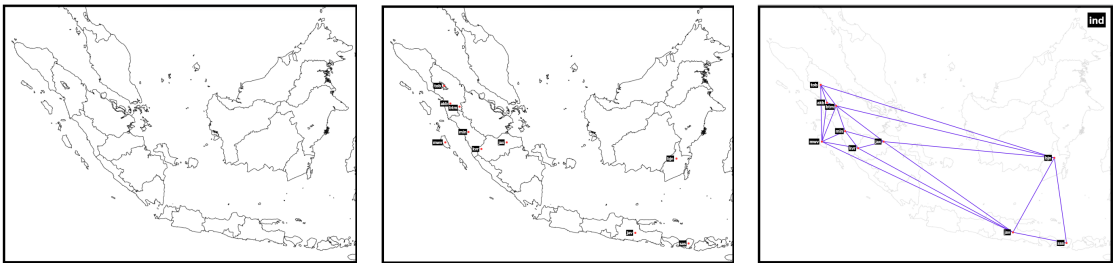
The Javanese language (**jav**) possesses a distinct sociolinguistic feature: its speech-level system, which includes *Ngoko*, *Madya*, and *Krama*. Javanese is also spoken by a very large population, roughly 90 million (Oakes, 2009). This language is spoken by the Javanese ethnic group, most of whom live in the provinces of Central Java, East Java, the Special Region of Yogyakarta, West Java, and Banten. In addition to Java, it is also widely distributed throughout East Kalimantan, South Kalimantan, Southeast Sulawesi, West Nusa Tenggara, Bali, Aceh, North Sumatra, the Riau Islands, Riau, Jambi, Bengkulu, and Lampung (Tim Pemetaan Bahasa, 2017). Beyond the Indonesian archipelago, Javanese has also spread to other countries, such as Suriname, Singapore, Malaysia, the Netherlands, and New Caledonia (Humaeni et al., 2011). For the purposes of this research, fieldwork was conducted in Jember.

This study also incorporates three out of seven languages in North Sumatra: Toba (tob), Angkola (**akb**), and Mandailing (**btm**)[vi]. All three varieties belong to the Western Austronesian subgroup (Aritonang & Silalahi, 2022; Harvina et al., 2017). Toba is mainly spoken in North Tapanuli and Toba Samosir Regencies, with around four million speakers (Saragih & Mulyadi, 2020). Angkola is concentrated in South Tapanuli, specifically within Padang Sidimpuan and the surrounding areas of Batang Tua, Gunung Tua, and Sipirok (Armis et al., 2023). Meanwhile, Mandailing is predominantly spoken in Mandailing Natal Regency, North Sumatra (Dalimunthe, 2018). Previous studies on these languages include diverse scholarly contributions, such as Percival's (1981) Toba grammar, an analysis of sound changes in Proto-Austronesian across Batak isolects (Siregar et al., 2022), research on Angkola morphology (Nasution, 2024), and research on the diachronic relationship of Mandailing (Batubara, 2025), among others.

METHODOLOGY

This study extracted linguistic data from ten observation points: Kumpeh District (**jax**), Sungai Penuh (**kvr**), Padang Panjang (**min**), Teluk Selong Ulu District (**bjn**), Muara Siberut (**mwv**), Jember (**jav**), Kopang District (**sas**), Ambarita District (**tob**), Padang Sidimpuan (**akb**), and Mandailing Natal (**btm**). Additionally, lexical items for Indonesian (**ind**) were sourced from the

Indonesian Online Big Dictionary (KBBI Online). The map of the observation points is illustrated in the figure below.



a. Base map b. Observation points c. Observation distances

Figure 1. Maps of observation points and distances

Data for this study were collected from February to December 2025. Each isolect studied involved three informants: one main informant and two supporting informants. The criteria for selecting informants were adapted from Mahsun (1995), Nadra and Reniwati (2023), and Kisyani-Laksono and Savitri (2009), who describe the characteristics of informants as follows: 1) aged between 20–60 years; 2) born in the research area, along with their spouses and parents; 3) having relatively low education; 4) belonging to a lower-middle social status with limited mobility; 5) holding a job with minimal human interaction; 6) able to communicate with the researchers; 7) being proud of their regional language; and 8) having no speech or mental disorders. The data consist of audio recordings of 257 glosses in every language; which were transcribed and categorized into four types: core vocabulary (L1 = 81 glosses), nature vocabulary (L2 = 45 glosses), general vocabulary (L3 = 110 glosses), and cultural vocabulary (L4 = 21 glosses).

To address the research problem, calculations were performed in three stages. The initial stage involved calculating the similarity of eleven modern languages spoken today (synchronically). The analysis involved a synchronic analysis of lexical overlap among the eleven languages, forming *55 unique language pairs*[vii], to identify the general etymological density by analyzing a wide array of vocabulary (257 glosses; L1-L4). This stage established a baseline for how these languages interact on a surface level (synchronically), accounting for both inherited traits (L1 and L2) and potential lexical borrowings (L3 and L4) resulting from areal proximity and contact. The formula used in this stage of analysis is a modification and adaptation of Jaccard's calculation (Bag et al., 2019; Kamiura & Sekine, 2023; Majumdar, 2025) as follows:

$$J_{(A,B)} = \frac{A \cap B}{A \cup B} \times 100\%$$

$J_{(A,B)}$: Similarity percentage of languages A and B, derived from 257 glosses

$A \cap B$: Number of cognates (words derived from the same etymons) shared by languages A and B (257 glosses)

$A \cup B$: Number of etymons present in languages A and B, from 257 glosses

The similarity percentages of the first stage were categorized into three levels: low, mid, and high similarity by following the classification in Table 1.

Table 1. Level of lexical similarities (synchronically)

Percentage	Level of similarity
81-100%	High
31-80%	Mid
≤30%	Low

The second stage involves analyzing comparative historical data to determine the lexical similarity and subgrouping among ten regional languages in Indonesia. This analysis is based on the percentage of cognates identified as inherited traits from L1 and L2 (126 glosses). This stage was conducted to isolate genetic relationships from external influences. The categories L1 and L2 represent concepts least susceptible to cultural change or lexical borrowing (i.e., pronouns, body parts, and fundamental natural phenomena). In this stage, *45 language pairs*[viii] were analyzed to determine the percentage of cognates through lexicostatistical calculation, as follows:

$$C_{(A,B)} = \frac{a \cap b}{a \cup b} \times 100\%$$

$C_{(A,B)}$: Lexical similarity percentage of languages A and B, derived from core vocabulary (L1 + L2 = 126 glosses)

$a \cap b$: Number of cognates (lexical items originating from the same etymons) shared by languages A and B (L1 + L2 = 126 glosses)

$a \cup b$: Number of etymons present in languages A and B, from core vocabulary (L1 + L2 = 126 glosses)

The similarity percentages of the lexicostatistical calculation provide a statistical basis for grouping the languages into distinct linguistic taxonomies, as follows:

Table 2. Language classification based on lexicostatistics
(Tim Pemetaan Bahasa, 2018)

Percentage	Category
81-100%	Same language
37-80%	Same family
12-36%	Same stock
4-11%	Same microphyllum
1-3%	Same mesophyllum
≤1%	Same macrophyllum

The final stage of the analysis was conducted to estimate the separation time of each language from its parent language, based on glottochronological analysis. The cognate percentages obtained from the second stage are utilized to determine the time depth of divergence. This facilitates an estimation of both the degree of relatedness and the historical timeline of linguistic separation. The results provide a temporal map[ix] that aligns linguistic evolution with the historical migration patterns in the Indonesian archipelago.

The glottochronology calculation used in this stage is conducted to provide a diachronic perspective on the data by applying a constant rate of change to the 126 core glosses, with a retention constant of 0.805. The calculation follows the formula:

$$t = \frac{\log(C)}{2\log(r)} \times 1000$$

t : Time depth in years
C : The percentage of shared cognates
r : The retention constant (0.805)

The result of the analysis is presented in the permutation forms (Matrices 1a–b) with the exclusion of Bahasa Indonesia (**ind**) in the second and third stages of analysis, because the language is considered the *lingua franca* of the Indonesian archipelago and is constantly evolving by adopting and borrowing various lexicons from both regional and international languages.

Jax	1									
kvr	2	11								
min	3	12	20							
bjn	4	13	21	28						
mwv	5	14	22	29	35					
sas	6	15	23	30	36	41				
jav	7	16	24	31	37	42	46			
tob	8	17	25	32	38	43	47	50		
akb	9	18	26	33	39	44	48	51	53	
btm	10	19	27	34	40	45	49	52	54	55
	ind	**jax**	**kvr**	**min**	**bjn**	**mwv**	**sas**	**jav**	**tob**	**akb**

a. Permutation form for stage 1

kvr	1								
min	2	10							
bjn	3	11	18						
mwv	4	12	19	25					
sas	5	13	20	26	31				
jav	6	14	21	27	32	36			
tob	7	15	22	28	33	37	40		
akb	8	16	23	29	34	38	41	43	
btm	9	17	24	30	35	39	42	44	45
	jax	**kvr**	**min**	**bjn**	**Mwv**	**sas**	**jav**	**tob**	**akb**

b. Permutation form for stage 2 and 3

Matrix 1. *Permutation forms*

RESULTS AND ANALYSES

This section presents the results and the analysis conducted in three stages while simultaneously answering the research questions. First, this analysis evaluates the degree of similarity among the regional languages under study, based on shared etymons and cognates. In the second stage, the genetic relationships of the regional languages are determined. Finally, this study estimates the approximate times of language separation. The findings are discussed below.

Stage 1: The degree of similarity among eleven languages in Indonesia

The degree of similarity among the 55 language pairs was analyzed through a systematic comparison of eleven consecutive languages based on three main indicators: the number of shared etymons (Matrix 2a), the number of confirmed cognates (Matrix 2b), and the percentage of lexical similarity calculated using a Jaccard-based statistical approach (Matrix 2c). Through this step-by-step analysis, this study attempts to provide an empirical overview of the distribution and degree of lexical closeness among the languages studied. The lexical dataset consists of 257 glosses yielding 1.210 etymons from eleven languages: **ind**, **jax**, **kvr**, **min**, **bjn**, **mwv**, **sas**, **jav**, **tob**, **akb**, and **btm**.

	ind	jax	Kvr	min	bjn	mwv	sas	jav	tob	Akb
jax	366									
kvr	414	389								
min	473	449	462							
bjn	442	423	443	502						
mwv	556	529	542	605	513					
sas	479	448	474	537	451	506				
jav	504	475	502	560	477	534	468			
tob	503	476	489	550	468	497	473	490		
akb	469	442	455	516	454	490	455	474	359	
btm	460	432	450	507	452	490	453	471	367	314

a. list of etymons

	ind	jax	kvr	min	bjn	Mwv	sas	jav	tob	akb
jax	283									
kvr	252	248								
min	253	248	252							
bjn	192	182	179	180						
mwv	69	67	71	68	68					
sas	146	148	139	136	130	66				
jav	131	131	121	123	114	48	114			
tob	108	106	110	109	99	61	85	78		
akb	135	133	137	136	106	61	96	87	178	
btm	141	140	139	142	105	58	95	87	167	213

b. list of cognates

	ind	jax	kvr	min	bjn	mwv	sas	jav	tob	akb
jax	77,32%									
kvr	60,87%	63,75%								
min	53,49%	55,23%	54,55%							
bjn	43,44%	43,03%	40,41%	35,86%						
mwv	12,41%	12,67%	13,10%	11,24%	13,26%					
sas	30,48%	33,04%	29,32%	25,33%	28,82%	13,04%				
jav	25,99%	27,58%	24,10%	21,96%	23,90%	8,99%	24,36%			
tob	21,47%	22,27%	22,49%	19,82%	21,15%	12,27%	17,97%	15,92%		
akb	28,78%	30,09%	30,11%	26,36%	23,35%	12,45%	21,10%	18,35%	49,58%	
btm	30,65%	32,41%	30,89%	28,01%	23,23%	11,84%	20,97%	18,47%	45,50%	67,83%

c. Percentage of lexical similarity

Matrix 2. *The degree of similarity of eleven languages*

Matrix 2a offers a broad view of lexical items that can be traced to the same etymological sources across language pairs. The relatively high values in several pairings suggest substantial lexical similarity, particularly among languages that are historically intertwined or geographically close. A more restricted pattern is observed in Matrix 2b, where only lexical items with phonological correspondence and stable meaning are retained. The lower values in Matrix 2b appear in all language pairs. This indicates that each language is unique. In other words, no two languages present a complete set of identical etymons[x] in portraying 257 glosses within and across the same meaning.

Matrix 2c shows the percentage of lexical similarity calculated through a Jaccard-based statistical approach, which normalizes shared cognates against the total lexicon. This stage brings the comparison onto a proportional scale, allowing differences in vocabulary size (total etymons) and cognates in certain pairs to be taken into account. The results show that most language pairs are at low similarity levels, while only a small number reach the mid-range category. As a whole, Matrices 2a–c describe the lexical similarity among the eleven languages, ranging from broad etymological overlap to confirmed cognancy, and measure the synchronic similarity of the 257 analyzed glosses.

Lexical similarity among the eleven languages was categorized using three indicators: (a) the number of shared etymons, (b) the number of cognates, and (c) similarity percentages calculated through a Jaccard-based lexicostatistical procedure. For the synchronic analysis, the

similarity percentages were grouped into three levels: high (81–100%), mid (31–80%), and low (≤30%), as presented in Matrix 3.

jax	Mid									
kvr	Mid	Mid								
min	Mid	Mid	Mid							
bjn	Mid	Mid	Mid	Mid						
mwv	Low	Low	Low	Low	Low					
sas	Low	Mid	Low	Low	Low	Low				
jav	Low	Low	Low	Low	Low	Low	Low			
tob	Low	Low	Low	Low	Low	Low	Low	Low		
akb	Low	Low	Low	Low	Low	Low	Low	Low	Mid	
btm	Low	Mid	Low	Low	Low	Low	Low	Low	Mid	Mid
	ind	**jax**	**kvr**	**min**	**bjn**	**mwv**	**sas**	**jav**	**Tob**	**akb**

Matrix 3. *Level of similarity*

Matrix 3 shows that no language pair attains a high similarity level. None of the pairs display lexical overlap comparable to nearly identical varieties. Mid-level similarity is observed in a limited set of language pairs, primarily among varieties that are geographically and historically proximate. Seven language pairs (**ind-jax**, **ind-kvr**, **ind-min**, **jax-kvr**, **jax-min**, **kvr-min**, and **akb-btm**) show relatively high similarities (over 50%), indicating that these pairs exhibit substantial numbers of shared etymons and cognates. The higher similarities may be attributed to the possibility of lexical inheritance which is further examined in the second stage of the analysis[xi].

In contrast, low-level similarity is observed in most of the pairwise comparisons. This pattern is especially prominent in comparisons involving languages such as **mwv**, **sas**, **jav**, and **tob** when compared with other members of the dataset. These languages consistently show smaller cognate counts, which result in low similarity percentages despite some shared etymological material. Such patterns indicate a greater degree of lexical divergence, likely reflecting earlier separation, sustained innovation, or contact-driven lexical replacement. Nevertheless, the similarity levels of all language pairs remain below the high category.

Stage 2: Genetic status of ten languages based on lexical distribution

This stage of analysis examines lexical kinship among ten regional languages (excluding **ind**). The lexical dataset consists of 126 glosses (L1 and L2) which yield 572 different etymons (from ten regional languages). The selection of L1 and L2 is based on their higher retentivity level because they consist of pronouns, demonstratives, body parts, numbers, and basic nature terms that are universally found in all languages. Therefore, this set of 126 glosses was carefully chosen for its relative resistance to borrowing and semantic replacement, which is usually present in general and cultural vocabulary. Matrices 4a–c display (a) total shared etymons, (b) confirmed cognates, and (c) proportional similarity (in percentages).

	jax	kvr	min	bjn	mwv	sas	jav	tob	akb
kvr	182								
min	208	211							
bjn	206	208	232						
mwv	253	250	279	253					
sas	221	225	253	219	239				
jav	238	244	262	237	258	235			
tob	237	233	262	231	239	229	239		
akb	220	215	241	224	238	225	238	172	
btm	214	212	235	222	238	222	233	172	160

a. list of etymons

	jax	kvr	min	bjn	mwv	sas	jav	tob	akb
kvr	123								
min	122	118							
bjn	99	96	97						
mwv	43	45	41	42					
sas	72	67	64	73	44				
jav	61	54	61	61	31	51			
tob	52	55	51	57	40	47	43		
akb	67	71	70	62	39	49	42	98	
btm	69	70	72	60	35	48	43	98	104

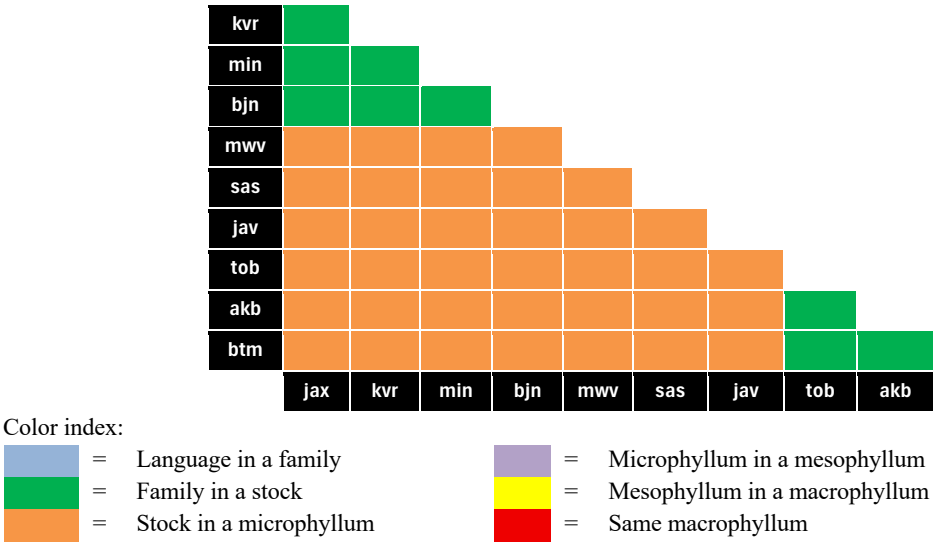
b. list of cognates

	jax	kvr	min	bjn	mwv	sas	jav	tob	akb
kvr	67,58%								
min	58,65%	55,92%							
bjn	48,06%	46,15%	41,81%						
mwv	17,00%	18,00%	14,70%	16,60%					
sas	32,58%	29,78%	25,30%	33,33%	18,41%				
jav	25,63%	22,13%	23,28%	25,74%	12,02%	21,70%			
tob	21,94%	23,61%	19,47%	24,68%	16,74%	20,52%	17,99%		
akb	30,45%	33,02%	29,05%	27,68%	16,39%	21,78%	17,65%	56,98%	
btm	32,24%	33,02%	30,64%	27,03%	14,71%	21,62%	18,45%	56,98%	65,00%

c. Percentage of lexical kinship (Lexicostatistics)

Matrix 4. *Etymons, cognates, and percentage of lexical kinship of ten regional languages*

The percentage of lexical kinship is calculated by dividing the number cognates by the total number of etymons in each language pair. The calculation is based on an adapted version of lexicostatistical formula, where each unique etymon is treated as an individual value. These percentages represent the proportion of shared inherited vocabulary within the restricted L1 and L2 lexical domains. To interpret these values, the classification thresholds proposed by Tim Pemetaan Bahasa (2018) are applied.



Color index:

= Language in a family
= Family in a stock
= Stock in a microphyllum
= Microphyllum in a mesophyllum
= Mesophyllum in a macrophyllum
= Same macrophyllum

Matrix 5. *Language kinship based on L1 & L2*

Based on the lexicostatistical percentages in Matrix 5, the overall language status represented in the dataset can be summarized as follows: (a) no language pairs reach the 81–100% range, therefore, none of the varieties can be classified as the same language; (b) a limited number of pairings fall within the 37–80 percent range, indicating relationships at the level of the same family; (c) the majority of pairwise comparisons cluster within the 12–36% interval, which denotes membership in the same stock and suggests a shared but relatively distant genetic origin; and (d) no language pair reaches below 12%, therefore none of the language pairs correspond to the same microphyllum or mesophyllum categories. Overall, the data indicate that the language varieties under study are predominantly related at the stock level, with stronger family-level ties occurring only among a limited subset of varieties (i.e., **jax**, **kvr**, **min**, and **bjn** are languages in one family; **tob**, **akb**, and **btm** are also languages in one family).

Stage 3: Estimated separation time of ten regional languages

Estimating the separation times of ten Indonesia's regional languages in this study was conducted by utilizing the results of the genetic analysis in the previous subsection. The separation times were calculated using the standard glottochronology formula with a retention constant of 0.805. The results of the calculations (in years) can be seen in Matrix 6.

	jax	kvr	min	bjn	mwv	sas	jav	tob	akb
kvr	903,18								
min	1229,79	1339,65							
bjn	1689,06	1782,26	2010,08						
mwv	4085,02	3952,73	4420,30	4139,26					
sas	2585,13	2792,38	3168,34	2532,38	3900,81				
jav	3138,12	3476,47	3359,57	3128,41	4884,38	3521,60			
tob	3496,37	3327,84	3772,29	3225,64	4120,51	3650,25	3953,81		
akb	2740,58	2553,92	2849,76	2960,89	4169,20	3513,57	3998,38	1296,67	
btm	2609,04	2554,23	2726,71	3015,80	4418,64	3530,16	3895,20	1296,67	992,99

Matrix 6. *The results of glottochronology calculation of ten regional languages*

The results of the glottochronology calculations of 45 language pairs in Matrix 6 reveal a complex landscape of linguistic relationships, where the separation values suggest a clear hierarchy of kinship. At the heart of the most closely related group are **jax** and **kvr**, which exhibit the lowest divergence value in the entire study (903.18 years), signaling a very recent common ancestor or a high degree of mutual intelligibility. This *green cluster* extends to **akb** and **btm** (992.99 years), indicating another tightly knit pair that likely forms a distinct southern or peripheral branch alongside **tob**.

In contrast, **mwv** emerges as the most significant linguistic outlier. It consistently yields the highest divergence values across language pairs, peaking at approximately 4884 years in its relationship with **jav**. This suggests that **mwv** is the most ancient or isolated lineage in the set, having separated from the other varieties several millennia ago. While most of the matrix fluctuates between moderate distances (the 2000 to 3000 range), the relationship between **mwv** and nearly every other variety pushes the data into the *red zone* or the highest divergence category, highlighting a deep historical rift.

The results also point toward a primary grouping of **jax**, **kvr**, **min**, and **bjn**, which maintain relatively low internal distances (mostly under 2000 years), suggesting they represent a core family group. However, a *middle gap* is evident when comparing this core group to varieties like **jav** or **tob**, where values jump significantly. Ultimately, the results of glottochronological analysis tell a story of a relatively unified core (**jax**/**kvr**) that stands in dramatic opposition to the highly divergent and unique evolution of **mwv**.

DISCUSSION

This study shows that regional languages in Indonesia exhibit varying lexical similarities, hierarchical genetic status, and temporal divergence. The lexical analysis reveals that each language has its own distinctive features, despite some shared words. No language pair is highly similar. It indicates the absence of identical varieties, while moderate similarity suggests historical proximity or contact between languages (Bowern, 2013).

The use of core and nature vocabulary (L1 & L2) strengthens the reliability of genetic classification, as such vocabulary is stable and rarely borrowed from other languages, according to the principles of classical lexicostatistics (Gudschinsky, 1956). Our findings also suggest that core vocabulary in L1 and L2 tends to retain both its form and meaning. For example, the lexemes for 'tongue'[xii] have a high degree of similarity across all the variations sampled in this study. Our glottochronology results indicate that languages within a family have been separated more recently, while most language pairs are at the stock level, indicating a more ancient common origin. The estimated separation times confirm that the degree of lexical similarity and genetic kinship align with the history of language divergence.

The core language group (**jax**, **kvr**, **min**, **bjn**) separated more recently and still retains moderate lexical similarity, while the peripheral languages (**mwv**, **jav**, **tob**) separated earlier and exhibit low similarity. This pattern suggests early separation, the development of their own vocabulary, and possible geographic or social isolation (Padilla-Iglesias et al., 2020; Schreier, 2009). More specifically, an isolated language such as **mwv** has experienced a longer period of separation and shows high lexical divergence, suggesting a developmental trajectory distinct from other languages. Capell (1982) in Nothofer (1986) asserts that the isolects under study are included in the inland Indonesian type while **mwv** belongs to the Barrier Island or Oceanic type. These findings imply that these languages have followed different evolutionary and developmental paths.

According to online databases (Ethnologue, ABVD, and Glottologue), the core language group is categorized under Nuclear Malayic (**bjn** in the East Borneo Malay sub-group; **jax**, **kvr**, and **min** are under Northern Sumatra Malay); the peripheral language group (**tob**, **akb**, **btm**) and **mwv** are under Sumatran (**tob**, **akb**, and **btm** in the Batakic subgroup; **mwv** is classified as a distinct language); whilst the other two languages, **jav** and **sas** are classified under different groups (the Javanic group and the Bali Sasak-Sumbawa group respectively).

What is striking in this study is that the findings diverge from the online databases. The online databases classify **mwv**, **tob**, **akb**, and **btm** under the Sumatran group, while the results of this study show the opposite. mwv and the other three languages share lower lexical similarities (**tob** 16,47%, **akb** 17,65%, **btm** 14,71%) than non-Sumatran languages (**sas** 18,41%, **kvr** 18%, **jax** 17%). A possible explanation is that the online databases used different corpora or secondary

sources in determining linguistic classification. This study, on the other hand, draws primary empirical data from informants through fieldwork.

Based on the results of this study, we propose that **tob**, **akb**, and **btm** belong to the Sumatran group, whilst **mwv** belongs the Barrier Island languages and should not be classified into the Sumatran group as claimed by the online databases. In this categorization, we align with the classification proposed by Capell (1982) and Nothofer (1986).

CONCLUSION

This study aims to determine the level of similarity among eleven languages in Indonesia (**ind**, **jax**, **krv**, **min**, **bjn**, **mwv**, **sas**, **jav**, **tob**, **akb**, and **btm**); to identify the genetic status of ten Indonesia's regional languages (**jax**, **krv**, **min**, **bjn**, **mwv**, **sas**, **jav**, **tob**, **akb**, and **btm**) based on their lexical distribution; and to calculate the estimated separation times of these 10 regional languages. The data for this study are based on 257 glosses, categorized into four types: core vocabulary (L1 = 81 glosses), nature vocabulary (L2 = 45 glosses), general vocabulary (L3 = 110 glosses), and cultural vocabulary (L4 = 21 glosses).

In determining the level of similarity, the eleven languages in Indonesia are paired into 55 unique pairs. The calculation is based on an adaptation of lexicostatistical and Jaccard's methods on the entire set of data (257 glosses; L1-L4). The results show that none of the pairs fall into the high-similarity category. Mid-level similarity appears in 15 language pairs (**ind-jax**, **ind-kvr**, **ind-min**, **ind-bjn**, **jax-kvr**, **jax-min**, **jax-btn**, **jax-sas**, **jax-btm**, **kvr-min**, **kvr-bjn**, **min-bjn**, **tob-akb**, **tob-btm**, and **akb-btm**). Low-level similarity is especially prominent in comparisons involving languages such as **mwv**, **sas**, **jav**, and **tob**. In identifying the genetic status of the ten regional languages, the results show they are categorized into: a) **jax**, **kvr**, **min**, and **bjn** are from families within the same stock, b) **tob**, **akb**, and **btm** are also from families within the same stock; and c) **sas**, **jav**, and **mwv** as languages from distinct stocks. Regarding the estimated separation times, the results indicate that the most recent language pair to separate is **jax-kvr** (apx. 903.18 years ago) and the earliest pair to separate is **mwv** and **jav** (apx. 4.884,38 years ago).

This study also critiques the language classification in online databases for asymmetrically categorizing languages based on inconsistent secondary data. One notable finding is that this research disagrees with the classification of **mwv** into the same group as **tob**, **akb**, and **btm** (Sumatran or Northern Sumatran). The results of the present analysis show that the similarity and genetic status of **mwv**, when paired with these three languages, are lower than those of the other analyzed languages. This study proposes that **mwv** is better categorized into a different group such as the Indonesian Barrier Island group proposed by Capell (1982) in Nothofer (1986).

Lastly, this study has limitations in that the corpora are limited to 257 glosses, and the total number of analyzed languages is limited to eleven languages (the *lingua franca* **ind** and ten regional languages). Future researchers are encouraged to enrich the dataset, as Indonesia possesses more than 700 regional languages. Another limitation is that this analysis focuses only on lexicostatistical adaptation using Jaccard's calculation and classical glottochronology formula. It is recommended that future research broaden the analysis to include *Cosine Similarity* and *Levenshtein Distance* for a more rigorous and comprehensive examination within the historical comparative linguistic approach.

NOTE

We would like to thank our primary and secondary informants for their contributions in providing a reliable linguistic dataset and their valuable explanations regarding our unstoppable curiosities about the languages and their cultural intactness. Our thanks also go to the *Linguistik Indonesia* editorial team for their valuable feedback on an earlier version of this paper.

REFERENCES

Adelaar, A. (2011). *Austronesian linguistics*. Oxford University Press. 10.1093/obo/9780199772810-0055

Adelaar, A., & Himmelmann, N. (2004). *The Austronesian languages of Asia and Madagascar*. Routledge.

https://www.taylorfrancis.com/books/mono/10.4324/9780203821121/austronesian-languages-asia-madagascar-alexander-adelaar-nikolaus-himmelmann

Adelaar, A., & Schapper, A. (2024). *The Oxford guide to the Malayo-Polynesian languages of Southeast Asia*. Oxford University Press. https://global.oup.com/academic/product/the-oxford-guide-to-the-malayo-polynesian-languages-of-southeast-asia-9780198807353

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. https://doi.org/10.26499/mm.v18i1.2330

Amri, U. (2017). *Identifikasi fonologis dan leksikal Bahasa Minangkabau isolek Nagari Pariangan* [Master Thesis, Universitas Andalas]. http://scholar.unand.ac.id/56808/

Amri, U. (2022). Variasi fonologis fonem vokal Bahasa Minangkabau isolek Nagari Pariangan. *Islamic Manuscript of Linguistics and Humanity*, *4*(1), 89–107. https://ejournal.uinib.ac.id/jurnal/index.php/imlah/article/download/4554/2761

Amri, U., Husna, L., Kartika Putri, A., Zubaidah, Z., & Pratiwi, A. (2025). Fonem konsonan, vokal, dan diftong dalam bahasa Minangkabau dan bahasa Korea: Kajian linguistik kontrastif. *Puitika*, *1*(1), 83–102. https://doi.org/10.25077/puitika.v21i1.689

Amri, U., Nadra, N., & Yusdi, M. (2020). Variasi leksikal bahasa Minangkabau di Nagari Tuo Pariangan. *Nusantara: Jurnal Ilmu Pengetahuan Sosial*, *7*(1), 52–78. http://dx.doi.org/10.31604/jips.v7i1.2020.52-78

Amri, U., Putra, Y. M., Putri, A. K., Triandana, A., & Fitriah, S. (2024). A comparative analysis of lexical variation of verbs in Minangkabau and Banjar languages: Historical comparative linguistic study. *Vivid: Journal of Language and Literature*, *13*(2), 185–193. https://doi.org/10.25077/vj.13.2.185-193.2024

Amri, U., Sirait, J. V., & Dewi, H. (2025). Kajian lintas bahasa variasi leksikal peralatan rumah tangga pada isolek Minangkabau dan Batak Toba. *Prosiding Seminar Nasional Humaniora*, *4*, 193–202. https://conference.unja.ac.id/SNH/article/view/391

Anderbeck, K. R. (2008). *Malay dialects of the Batanghari river basin (Jambi, Sumatra)*. SIL International. https://www.sil.org/resources/publications/entry/9245

Aritonang, I. Y., & Silalahi, D. A. (2022). Perubahan bunyi bahasa Proto-Austronesia ke dalam bahasa Batak dialek Toba. *Talenta Conference Series: Local Wisdom, Social, and Arts (LWSA)*, *5*(1), 106–111. https://doi.org/10.32734/lwsa.v5i1.1331

Armis, M. K., Harahap, A. I., & Syarfina, T. (2023). Analisis prosodi kajian fonetik akustik pada Bahasa Batak Angkola. *Fon: Jurnal Pendidikan Bahasa dan Sastra Indonesia*, *19*(1), 158–165. https://doi.org/10.25134/fon.v19i1.6878

Badan Pengembangan dan Pembinaan Bahasa. (2025). *Persebaran bahasa daerah berdasarkan provinsi* [Indonesian Government]. Data Pokok Kebahasaan dan Kesastraan. https://dapobas.dikdasmen.go.id/home?show=isidata&id=195

Bag, S., Kumar, S. K., & Tiwari, M. K. (2019). An efficient recommendation generation using relevant Jaccard similarity. *Information Sciences*, *483*, 53–64. https://doi.org/10.1016/j.ins.2019.01.023

Batubara, M. H. (2025). The position of bahasa Mandailing within the linguistic affiliation of nusantara languages: A systematic literature review. *Seunebok Lada: Jurnal ilmu-ilmu Sejarah, Sosial, Budaya dan Pendidikan*, *12*(2), 552–565. https://doi.org/10.33059/jsnbl.v12i2.12933

Bellwood, P. (2024). The origins and spread of agriculture in the Indo-Pacific region: Gradualism and diffusion or revolution and colonization? In D. R. Harris (Ed.), *The origins and spread of agriculture and pastoralism in Eurasia* (pp. 465–498). Routledge. https://www.taylorfrancis.com/chapters/edit/10.4324/9780203020906-6/origins-spread-agriculture-david-harris

Billings, B., & McDonnell, B. (2024). Sumatran. *Oceanic Linguistics*, *63*(1), 112–174. https://doi.org/10.1353/ol.2024.a928205

Blust, R. A. (2013). *The Austronesian languages* (Revised Edition). Asia-Pacific Linguistics Research School of Pacific and Asian Studies of The Australian National University. http://hdl.handle.net/1885/10191

Blust, R. A. (2015). 35 Southeast Asian islands and Oceanic Austronesian linguistic history. In P. Bellwood (Ed.), *The global history of human migration:* (pp. 276–283). Willey Blackwell. https://www.researchgate.net/publication/313988992_35_Southeast_Asian_islands_and_Oceania_Austronesian_linguistic_history

Bowern, C. (2013). Relatedness as a factor in language contact. *Journal of Language Contact*, *6*(2), 411–432. https://doi.org/10.1163/19552629-00602010

Budiono, S., Novita, R., & Syarfina, T. (2023). Mentawai language variations in the Mentawai Islands Regency, West Sumatra Province. *Jurnal Arbitrer*, *10*(1), 8–18. https://doi.org/10.25077/ar.10.1.8-18.2023

Burhanuddin, B., Melani, B. Z., & Saharudin, S. (2025). Austronesian's traces in Sasak: Historical linguistics study. *Jurnal Arbitrer*, *12*(2), 238–258. https://doi.org/10.25077/ar.12.2.238-258.2025

Campbell, L., & Grondona, V. (2008). Ethnologue: Languages of the world. *Language*, *84*(3), 636–641. https://doi.org/10.1353/lan.0.0054

Capell, A. (1982). Bezirkssprachen im gebiet des UAN. *Gava': Studies in Austronesian Languages and Cultures Dedicated to Hans Kähler*, 1–14. https://www.semanticscholar.org/paper/GAVA%CA%BF-%3A-studies-in-Austronesian-languages-and-%3A-to Carle/570d3c46c810ff9afd0dd2d1239ceca9826e0e69

Casasanto, D. (2008). Similarity and proximity: When does close in space mean close in mind? *Memory & Cognition*, *36*(6), 1047–1056. https://doi.org/10.3758/MC.36.6.1047

Cavalli-Sforza, L. L. (1997). Genes, peoples, and languages. *Proceedings of the National Academy of Sciences*, *94*(15), 7719–7724. https://doi.org/10.1073/pnas.94.15.7719

Collin, R. O. (2010). Ethnologue. *Ethnopolitics*, *9*(3–4), 425–432. https://doi.org/10.1080/17449057.2010.502305

Cornwell, S. E. (2019). *Language classification in the Ethnologue and its consequences*. Proceedings of the Annual Conference of CAIS/Actes du congrès annuel de l'ACSI. https://doi.org/10.29173/cais1104

Crouch, S. E. (2009). *Voice and verb morphology in Minangkabau, a language of West Sumatra, Indonesia* [Master Thesis, The University of Western Australia]. https://pure.mpg.de/rest/items/item_886558_2/component/file_886556/content

Dalimunthe, S. (2018). Hubungan kekerabatan bahasa Batak Mandailing dan bahasa Tanah Ulu (Suatu kajian linguistik historis komparatif). *Medan Makna: Jurnal Ilmu Kebahasaan Dan Kesastraan*, *16*(1), 84–91. https://doi.org/10.26499/mm.v16i1.2276

Dewanti, R., & Zainuddin. (2024). Kinship relationship between Mandailing and Toba languages: A comparative historical linguistic study. *JALC: Journal of Applied Linguistic and Studies of Cultural*, *2*(2), 23–28.
https://jurnal.rahiscendekiaindonesia.co.id/index.php/jalc/article/view/529

Eberhard, D. M., Gary F., S., & Fennig, C. D. (2025). *Ethnologue* [Online Encyclopedia]. Ethnologue. https://www.ethnologue.com/

Edwards, O. (2015). The position of Enggano within Austronesian. *Oceanic Linguistics*, *54*(1), 54–109. https://doi.org/10.1353/ol.2015.0001

Endriani, H., Ernanda, & Afria, R. (2023). Alih kode dialek Kecamatan Danau Kerinci dengan bahasa Korea: Studi kasus pada penggemar budaya Korea. *Kajian Linguistik Dan Sastra*, *2*(3), 293–304. https://doi.org/10.22437/kalistra.v2i3.24358

Ermanto. (2025). *Linguistik historis komparatif: Teori dan praktik penentuan kekerabatan bahasa di dunia*. PT. Raja Grafindo Persada.
https://www.rajagrafindo.co.id/produk/linguistik-historis-komparatif-teori-dan-praktik-penentuan-kekerabatan-bahasa-di-dunia-prof-dr-ermanto-s-pd-m-hum/

Ermanto, & Emidar. (2018). *Perbandingan bahasa Minangkabau, Kerinci, dan Mentawai: Suatu tinjauan linguistik historis komparatif*. Universitas Negeri Padang Press. https://www.researchgate.net/publication/328344663_Perbandingan_Bahasa_Minangkabau_Kerinci_dan_Mentawai_Suatu_Tinjauan_Linguistik_Historis_Komparatif

Ernanda. (2015). Phrasal alternation in the Pondok Tinggi dialect of Kerinci: An intergenerational analysis. *Wacana*, *16*(2), 355–382. https://doi.org/10.17510/wacana.v16i2.382

Ernanda. (2017). Phrasal alternation in Kerinci. *Wacana*, *18*(3), 791–812. https://doi.org/10.17510/wacana.v18i3.637

Ernanda. (2018). Pemilihan bahasa dan sikap bahasa pada masyarakat Pondok Tinggi Kerinci. *Titian: Jurnal Ilmu Humaniora*, *2*(2), 193–211.
https://doi.org/10.22437/titian.v2i02.6087

Ernanda. (2020). The referential uses of demonstratives in Kerinci Malay, Indonesia. *Arbitrer*, *7*(2), 118–127. https://doi.org/10.25077/ar.7.2.118-127.2020

Ernanda. (2021). Some notes on the Semerap dialect of Kerinci and its historical development. *Wacana, Journal of the Humanities of Indonesia*, *22*(1), 4. https://doi.org/10.17510/wacana.v22i1.978

Ernanda, Ekarina, & Arief, N. (2025). Ornamental replication in multilingual Duano speakers. *WORD*, *71*(3), 131–156. https://doi.org/10.1080/00437956.2025.2540183

Ernanda, & Yap, F. H. (2024). Phrasal alternation and Kerinci demonstrative (i) neh: Implications for spatial and socio-interactional deixis. *Journal of Pragmatics*, *222*, 40–59. https://doi.org/10.1016/j.pragma.2023.12.002

Farid, R. N. (2012). Bahasa Banjar: Its varieties and characteristics (A conceptual description of Bahasa Banjar in sociolinguistics point of view). *Language Maintenance and Shift II*, *2*, 517–521. https://www.academia.edu/93387049/Bahasa_Banjar_Its_Varieties_and_Characteristics_A_Conceptualdescription_of_Bahasa_Banjar_in_Sociolinguistics_Point_of_View_

Fatria, M., Ernanda, & Afria, R. (2023). Analisis relasi makna sinonim dan antonim bahasa Kerinci dialek Tebing Tinggi Kecamatan Danau Kerinci. *Kajian Linguistik dan Sastra*, *2*(2), 114–121. https://doi.org/10.22437/kalistra.v2i2.23184

Febrina, R. (2014). *Geografi dialek bahasa Mentawai di Kecamatan Siberut Selatan* [Master Thesis, Universitas Andalas]. http://scholar.unand.ac.id/7938/

Greenhill, S. J., Blust, R., & Gray, R. (2025a). *Austronesian basic vocabulary database*. Austronesian Basic Vocabulary Database. https://abvd.eva.mpg.de/austronesian/

Greenhill, S. J., Blust, R., & Gray, R. D. (2008). The Austronesian basic vocabulary database: From bioinformatics to lexomics. *Evolutionary Bioinformatics*, *4*, EBO-S893. https://doi.org/10.4137/EBO.S893

Greenhill, S. J., Blust, R., & Gray, R. D. (2025b). *Basic vocabulary database: Mentawai*. Austronesian Basic Vocabulary Database.
https://abvd.eva.mpg.de/austronesian/search.php?type=language&query=mentawai

Gudschinsky, S. C. (1956). The ABC's of lexicostatistics (glottochronology). *Word*, *12*(2), 175–210. https://doi.org/10.1080/00437956.1956.11659599

Hammarström, H. (2015). Ethnologue 16/17/18th editions: A comprehensive review. *Language*, *91*(3), 723–737. https://doi.org/10.1353/lan.2015.0038

Harmedianti, H., Ernanda, & Afria, R. (2023). Variasi leksikal bahasa Kerinci isolek desa-desa di Kecamatan Depati Tujuh Kabupaten Kerinci: Kajian dialektologi. *Jurnal Kalistra: Kajian Bahasa dan Sastra*, *1*(3), 257–270. https://doi.org/10.22437/kalistra.v1i3.20307

Harvina, H., Fariani, F., Putra, D. K., Simanjuntak, H., & Sihotang, D. (2017). *Daliha na tolu pada masyarakat Batak Toba di Kota Medan*. Balai Pelestarian dan Nilai Budaya Aceh. https://repositori.kemendikdasmen.go.id/24438/

Hugh, V. (2012). *Ethnologue: The linguistic straw-man*. The Journeyler. https://hugh.thejourneyler.org/2012/ethnologue-the-linguistic-straw-man/

Huisman, J. L., Franco, K., & van Hout, R. (2021). Linking linguistic and geographic distance in four semantic domains: Computational geo-analyses of internal and external factors in a dialect continuum. *Frontiers in Artificial Intelligence*, *4*, 668–035. https://doi.org/10.3389/frai.2021.668035

Humaeni, A., Ulumi, H. F. B., & Heryatun, Y. (2011). *Peta bahasa masyarakat Banten*. Laboratorium Bantenologi IAIN Sultan Maulana Hasanuddin. https://repository.uinbanten.ac.id/4238/1/Peta%20Bahasa.pdf

Kamiura, M., & Sekine, R. (2023). Jaccard matrix for nonlinear filter statistics. *SICE Journal of Control, Measurement, and System Integration*, *16*(1), 152–163. https://doi.org/10.1080/18824889.2023.2194169

Kawi, D. (1991). *Bahasa Banjar: Dialek dan subdialeknya* [Doctoral Dissertation, Universitas Indonesia]. https://lontar.ui.ac.id/detail?id=83540

Kisyani, L., & Savitri, A. D. (2009). *Dialektologi*. Unesa University Press.

Klamer, M. (2018). Documenting the linguistic diversity of Indonesia: Time is running out. In Santri. E. P. Djahimo (Ed.), *Revitalization of local languages as the pillar of pluralism* (pp. 1–10). Satya Wacana University Press. https://www.researchgate.net/publication/363832819_ISBN_Proceedings-_International_Conference_on_Local_Languages_Revitalization_on_Local_Languages_as_the_Pillar_of_Pluralism

Kroeber, A. L. (1955). Linguistic time depth results so far and their meaning. *International Journal of American Linguistics*, *21*(2), 91–104. https://doi.org/10.1086/464318

Lenggang, Z., Nio, B. K. H., Ansyar, M., Zainil, & Adam, S. (1978). *Bahasa Mentawai*. Pusat Pembinaan dan Pengembangan Bahasa, Departemen Pendidikan dan Kebudayaan. https://repositori.kemendikdasmen.go.id/2366/

Mahriyuni, Isda Pramuniati, & Rizky Ainun Maftuhah. (2023). Lexicostatistics of Javanese and Sasak languages: Comparative historical linguistic studies. *Mimbar Ilmu*, *28*(1), 124–130. https://doi.org/10.23887/mi.v28i1.59797

Mahsun. (1995). *Dialektologi diakronis: Sebuah pengantar*. Gadjah Mada University Press.

Majumdar, D. (2025). *Introduction to lexical similarity*. Language Technology and Data Analysis Laboratory. https://ladal.edu.au/tutorials/lexsim/lexsim.html#jaccard-similarity

Meliana, R., Manalu, M. M. S., & Triyono, S. (2024). Tracing the linguistic roots of Malay and Batak languages in Sumatra Island: A historical comparative study. *OKARA: Jurnal Bahasa dan Sastra*, *18*(1), 142–164. https://doi.org/10.19105/ojbs.v18i1.12865

Nadra & Reniwati. (2023). *Dialektologi teori dan metode* (2nd ed.). Textium. https://grahailmu.id/textium/produk/dialektologi-edisi-2-teori-dan-metode/

Naim, M. (2013). *Merantau pola migrasi suku Minangkabau* (3rd ed.). PT Raja Grafindo Persada. https://www.scribd.com/document/851271920/Merantau-Pola-Migrasi-Suku-Minangkabau

Nalee, M. A., Nadra, N., & Yusdi, M. (2020). Hubungan kekerabatan bahasa Melayu Patani dengan bahasa Minangkabau. *Madah: Jurnal Bahasa dan Sastra*, *11*(1), 43–56. http://dx.doi.org/10.31503/madah.v11i1.225

Nasution, H. S. (2024). Comparative analysis of word formation and particle of language on Angkola Barumun and Angkola Tapanuli Selatan language: Written text taken from WA script. *Indonesian Journal of Education, Social Sciences and Research (IJESSR)*, *5*(2), 17–28. https://doi.org/10.30596/ijessr.v5i2.20399

Nothofer, B. (1986). The barrier island languages in the Austronesian language family. In P. Geraghty, L. Carrington, & S. A. Wurm (Eds.), *FOCAL II: Papers from the Fourth International Conference on Austronesian Linguistics* (Vol. 94, pp. 87–109). Department of Linguistics Research School of Pacific Studies The Australian National University. https://openresearch-repository.anu.edu.au/bitstreams/749ab386-9a3e-49d8-bd0e-7929cec4c069/download

Oakes, M. P. (2009). Javanese. In B. Comrie (Ed.), *The world's major languages* (pp. 830–843). Routledge. https://doi.org/10.4324/9780203301524

Padilla-Iglesias, C., Gjesfjeld, E., & Vinicius, L. (2020). Geographical and social isolation drive the evolution of Austronesian languages. *PLOS ONE*, *15*(12), e0243171. https://doi.org/10.1371/journal.pone.0243171

Paolillo, J. C., & Das, A. (2006). *Evaluating language statistics: The Ethnologue and beyond* [Contract report for UNESCO Institute for Statistics]. UNESCO Institute for Statistics. https://www.academia.edu/download/92975/UNESCO_report_Paolillo_Das.pdf

Percival, W. K. (1981). In *A grammar of the urbanised Toba-Batak of Medan* (Vol. 76). Departement of Linguistics, Research School of Pacific Studies, The Australian National University.

https://openresearch-repository.anu.edu.au/server/api/core/bitstreams/27ab9c6e-f1e3-4d17-8c8e-03d6b7423a2e/content

Petroni, F., & Serva, M. (2010). Measures of lexical distance between languages. *Physica A: Statistical Mechanics and Its Applications*, *389*(11), 2280–2283. https://doi.org/10.1016/j.physa.2010.02.004

Pratiwi, A., Revita, I., Fauzanna, W., Ghaniyyah, M., & Amri, U. (2025). Speech acts in Minangkabau language during commercial transactions in Mentawai's traditional market: A case study in Pasar Raya Muara Siberut. *Andalas International Journal of Socio-Humanities*, *7*(1), 31–40. https://doi.org/10.25077/aijosh.v7i1.80

Rahman, H. (2016). 'Merantau'—An informal entrepreneurial learning pattern in the culture of Minangkabau tribe in Indonesia. *DeReMa (Development Research of Management): Jurnal Manajemen*, *11*(1), 15–34. https://doi.org/10.19166/derema.v11i1.186

Ross, M. (1996). On the origin of the term "Malayo-Polynesian." *Oceanic Linguistics*, *35*(1), 143–145.

Saragih, E. L. L. & Mulyadi. (2020). Pola pembentukan konstruksi verba serial dalam bahasa Batak Toba (Teori X-Bar). *GERAM (Gerakan Aktif Menulis)*, *8*(1), 1–8. https://doi.org/10.25299/geram.2020.4432

Schreier, D. (2009). Language in isolation, and its implications for variation and change. *Language and Linguistics Compass*, *3*(2), 682–699. https://doi.org/10.1111/j.1749-818X.2009.00130.x

Sholeha, M. (2022). Kekerabatan bahasa Melayu Jambi dan Melayu Palembang. *Kabastra: Kajian Bahasa Dan Sastra*, *2*(1). https://doi.org/10.31002/kabastra.v2i1

Sholeha, M., & Hendrokumoro, H. (2022). Kekerabatan bahasa Kerinci, Melayu Jambi, dan Minangkabau. *Diglosia: Jurnal Kajian Bahasa, Sastra, Dan Pengajarannya*, *5*(2), 399–420. https://doi.org/10.30872/diglosia.v5i2.404

Siregar, E. D., Ernanda, & Afria, R. (2022). Perubahan bunyi bahasa Proto Austronesia (PAN) pada bahasa Karo, bahasa Toba, bahasa Pakpak, bahasa Simalungun, bahasa Mandailing dan bahasa Angkola: Kajian linguistik historis komparatif dan fonologi. *Kalistra: Kajian Linguistik dan Sastra*, *1*(2), 116. https://doi.org/10.22437/kalistra.v1i2.20294

Sneddon, J. (2003). *The Indonesian language*. University of New South Wales Press. http://ndl.ethernet.edu.et/bitstream/123456789/2877/1/80.pdf.pdf

Steinhauer, H. (2018). Sound-changes and loanwords in Sungai Penuh Kerinci. *Wacana, Journal of the Humanities of Indonesia*, *19*(2), 5. https://doi.org/10.17510/wacana.v19i2.708

Tim Pemetaan Bahasa. (2017). *Bahasa dan peta bahasa di Indonesia*. Kementerian Pendidikan dan Kebudayaan.

https://repositori.kemendikdasmen.go.id/7191/1/Peta%20Bahasa%202017.compressed-min%20%28pdf.io%29.pdf

Tim Pemetaan Bahasa. (2018). *Pedoman penelitian pemetaan bahasa*. Pusat Pengembangan dan Pelindungan Bahasa dan Sastra, Kemdikbud. http://repositori.kemdikbud.go.id/id/eprint/22496

Wahab, M. K. A., & Halin, A. K. C. (2021). Leksikostatistik dan glotokronologi antara bahasa Banjar dengan bahasa Melayu: Kajian linguistik sejarah dan Perbandingan. *Jurnal Kesidang*, *6*(1), 44–61.
https://www.unimel.edu.my/journal/index.php/JK/article/view/969

Wilymafidini, O. (2017). An analysis of the dominance of Minang dialect in Kerinci society. *Inovish Journal*, *2*(2), 63–78.
https://ejournal.polbeng.ac.id/index.php/IJ/article/view/234

Zein, S. (2020). *Language policy in superdiverse Indonesia* (1st Edition). Routledge. https://doi.org/10.4324/9780429019739

---

[i] Language codes are based on ISO 639-3

[ii] The study was carried out in Sungai Penuh

[iii] The observation point is in Padang Panjang

[iv] The data collection site is situated in Teluk Selong Ulu, Banjar Regency

[v] The observation point is in Kopang, Central Lombok Regency

[vi] The observation points for this study were Ambarita for Toba, Padang Sidimpuan for Angkola, and Mandailing Natal for Mandailing.

[vii] See *Matrix 1a*

[viii] See *Matrix 1b*

[ix] The map only has ten regional languages analyzed (not every language in the area)

[x] $A \cup B = (A+B) - (A \cap B)$; $A \neq B$

[xi] Next subsection

[xii] ind, jax, and min = [lidah]; kvr = [lidwah]; bjn and jav = [ilat]; mwv = [lila]; sas = [elaʔ]; tob, akb, and btm = [dila]