Linguistik Indonesia, Februari 2026, 55-77 *Volume ke-44, No. 1*

ISSN cetak 0215-4846; ISSN online 2580-2429

THE EXPLORATION OF LANGUAGE EVOLUTION: A STUDY OF LINGUISTIC DIVERGENCE AND KINSHIP IN THE PROVINCES OF SUMATRA ISLAND

Riska Meliana[1], Irfa Luthfia Rahmani[2], Ana Yuliana[3], Rizqy Wafi[4]
Universitas Negeri Yogyakarta[1,3,4], Universitas Bengkulu[2]
riskameliana.2022@student.uny.ac.id[1], irfaluthfiar@unib.ac.id[2],
anayuliana.2022@student.uny.ac.id[3], rizqywafi.2021@student.uny.ac.id[4]

Abstract

This research investigates language evolution and linguistic kinship on the island of Sumatra. It also extends coverage to cognate languages in other regions. The languages studied include those spoken by various ethnic groups, namely Aceh, Gayo (Aceh), Batak Toba, Mandailing (North Sumatra), Rejang, Serawai (Bengkulu), Melayu Bangka, and Kayu Agung (Bangka Belitung). The main objectives are (1) to investigate how quantitative and qualitative analyses reveal kinship relationships between Acehnese (AT) and Gayo (GT) in Aceh, Batak Toba (BTT) and Batak Mandailing (BMT) in North Sumatra, Rejang (RT) and Serawai (ST) in Bengkulu, and Melayu Bangka (MBT) and Kayu Agung (KAT) in Bangka Belitung; (2) to identify and present empirical evidence to determine the divergence time for each language pair, and (3) to classify the studied languages into specific kinship groups and to identify the proportions of kinship relationships among languages in Aceh, North Sumatra, Bengkulu, and Bangka Belitung. This research used the lexicostatistical and glottochronological methods developed by Swadesh. Word kinship was evaluated using a list of 200 words. The results showed significant differences among the eight languages. The languages in North Sumatra Province and Bengkulu, for example, had a low similarity rate of 17%. The kinship percentage of local languages in Bengkulu and Bangka Belitung provinces averaged 50.5%. This places them in the "Language of Family" category, indicating a correlation in vocabulary despite variations in phonetic elements and dialects. Glottochronological calculations estimate the time of separation between the languages to range from 430 BC to 3,590 AD. This research makes a significant contribution and plays a vital role in supporting language documentation and preservation. It also helps to understand the social and cultural dynamics that influence language development in society.

Keywords: Glottochronology, Lexicostatistics, languages in Sumatra

Abstract

Penelitian ini menyelidiki evolusi bahasa dan kekerabatan linguistik di beberapa provinsi di Pulau Sumatra dan memperluas cakupan ke bahasa-bahasa serumpun di wilayah lain. Bahasa-bahasa yang diteliti meliputi bahasa-bahasa yang dituturkan oleh berbagai kelompok etnis, yaitu Aceh, Gayo (Aceh), Batak Toba, Batak Mandailing (Sumatera Utara), Rejang, Serawai (Bengkulu), Melayu Bangka, dan Kayu Agung (Bangka Belitung). Tujuan penelitian ini adalah untuk (1) mengetahui sejauh mana hasil analisis kuantitatif dan kualitatif dapat mengungkap hubungan kekerabatan antara bahasa Aceh (AT) dan Gayo (GT) di Aceh, bahasa Batak Toba (BTT) dan Batak Mandailing (BMT) di Sumatra Utara, bahasa Rejang (RT) dan Serawai (ST) di Bengkulu, serta bahasa Melayu Bangka (MBT) dan Kayu Agung (KAT) di Bangka Belitung; (2) menemukan dan menyajikan bukti empiris yang dapat digunakan untuk menentukan waktu perbedaan dari masing-masing pasangan

bahasa tersebut; (3) dan mengklasifikasikan bahasa-bahasa yang diteliti ke dalam kelompok-kelompok kekerabatan tertentu serta mengidentifikasi proporsi hubungan kekerabatan di antara bahasa-bahasa di Aceh, Sumatra Utara, Bengkulu, dan Bangka Belitung. Penelitian ini menggunakan metode leksikostatistik dan glotokronologi yang dikembangkan oleh Swadesh dengan mengevaluasi kekerabatan kata berdasarkan daftar 200 kata dasar. Hasil penelitian menunjukkan adanya perbedaan yang signifikan di antara kedelapan bahasa tersebut, terutama antara bahasa-bahasa di Provinsi Sumatera Utara dan Bengkulu yang memiliki tingkat kemiripan yang rendah, yaitu 17%. Persentase kekerabatan bahasa-bahasa daerah di Provinsi Bengkulu dan Bangka Belitung mencapai rata-rata 50,5%, menempatkan mereka dalam kategori "Bahasa Keluarga", yang mengindikasikan adanya korelasi kosakata meskipun terdapat perbedaan elemen fonetik dan dialek. Berdasarkan perhitungan glotokronologi, perkiraan waktu pemisahan antara bahasa-bahasa tersebut berkisar antara 430 SM hingga 3.590 Masehi. Penelitian ini memberikan kontribusi yang signifikan dan berperan penting dalam mendukung upaya dokumentasi dan pelestarian bahasa, serta membantu memahami dinamika sosial dan budaya yang mempengaruhi perkembangan bahasa di masyarakat.

Kata kunci: *Glotokronologi, Leksikostatistik, bahasa-bahasa di Sumatra*

INTRODUCTION

Verbal exchanges or interactions occur between members of society in social life. Language has a broad function as a method of social communication. Therefore, an intermediary tool called language is needed. In communicating, a person aims to convey meaning to others through language (Tantri et al., 2024). Language plays a vital role in understanding cultural heritage and diversity in a region and passing on knowledge from one generation to another (Mailani et al., 2022; Meliana et al., 2024). The movement of people from one region to another causes the language to become separated from the parent language or mother tongue because it adapts to social, natural, and environmental contacts where the community lives. The linguistic development of a language is inseparable from the kinship or similarity of one language with another (Istanti et al., 2020) . Language kinship is a collection of languages from a language group with similarities and the same development history (Setiawan, 2020). Thus, languages that are related or have similar vocabularies are the same proto-language (Meliana et al., 2024).

The vocabulary of these various languages reveals notable similarities and differences. On the one hand, words with similar meanings and structures reflect historical relationships or mutual influences between the languages. On the other hand, some words are markedly different, indicating unique and independent developments of each language. For instance, the languages spoken in the provinces of Aceh and North Sumatra exhibit some similarities, such as the word for “ayam” or “chicken”, which is pronounced [manok] in Aceh and [manuk] in Toba and Mandailing. Another example can be seen in the languages spoken in the provinces of Bengkulu and Bangka Belitung. Some similarities are evident in the word for *bersih* ‘clean’, which is pronounced [bərsih] in Rejang and Kayu Agung and [bərsiah] in Serawai and Malay Bangka languages. In this study, not all languages spoken in Aceh, North Sumatra, Bengkulu, and Bangka Belitung could be included due to the limited number of sources and the broad scope of discourse.

The term *indigenous language* or *autochthonous language* originates from Greek which means “arising from the land itself.” It refers to languages that emerged, evolved, and have been historically transmitted across generations within a specific geographic area by its native

inhabitants for centuries. In contrast, the term *allochthonous language*, also derived from Greek meaning "originating from elsewhere," describes languages introduced into a region by migrant or incoming populations (Bastardas-Boada, 2018; Kazakova & Shakhnazaryan, 2020). Stephens (1976) distinguishes *indigenous* communities as groups "native to a particular area" with an indefinite duration of residence. In contrast, autochthonous communities are understood as groups with a strong and longstanding attachment to the place where they live (of-the-soil connection). Today, the terms *autochthonous* and *allochthonous* are preferred over *indigenous* and *immigrant* respectively (McLeod, 2009; Granić, 2025). Accordingly, this study consistently adopts the terms *autochthonous* and *allochthonous*.

Table 1 below demonstrates four autochthonous languages with three allochthonous languages, namely Aceh, Devayan, Gayo, and Sigulai in Aceh Province, and six autochthonous languages with three allochthonous languages, namely Nias and Batak languages, including Toba, Mandailing, Simalungun, Pakpak (Dairi), and Karo in North Sumatra Province. Bengkulu Province has three autochthonous languages and three allochthonous languages, namely Bengkulu, Enggano, and Rejang. Meanwhile, in the province of Bangka Belitung, there are only one autochthonous languages, namely Malay Bangka and Kayu Agung as their allochthonous languages that have been identified through the Language Mapping study in Indonesia by the Agency for Language Development and Cultivation (*Badan Pengembangan dan Pembinaan Bahasa*) (Mahsun et al., 2017).

Table 1. Language Identification in Aceh and North Sumatra Provinces

Aceh Province	**North Sumatra Province**	**Bengkulu Province**	**Bangka Belitung Province**
Autochthonous language			
Aceh language	Batak language	Bengkulu language	Melayu Bangka language
Devayan language	Nias language	Enggano language	
Gayo language		Rejang language	
Sigulai language			
Allochthonous language			
Batak language	Javanese language	Javanese language	Kayu Agung language
Javanese language	Melayu language	Minangkabau language	
Minangkabau language	Minangkabau language	Sunda language	

Source: Mahsun (2017)

Table 2 lists the language distribution areas in Aceh, North Sumatra, Bengkulu, and Bangka Belitung provinces. These tables cover specific regions and sub-districts where various languages are spoken. This information is valuable for understanding the region's linguistic diversity and can be used to promote efforts to preserve language and cultural diversity.

Table 2. Areas of Language Distribution in Aceh Province, North Sumatra Province, Bengkulu Province, and Bangka Belitung Province

Aceh Province	North Sumatra Province	Bengkulu Province	Bangka Belitung Province
Banda Aceh and Lhokseumawe City	Dolok Sanggul District, Humbang Hasundutan Regency	Tertik Village, Tebat Karai District, Kepahiang Regency.	Kimak Village, Merawang Subdistrict, Bangka Regency.
North Aceh Regency, Bener Meriah Regency	Panyabungan District and (Lumban Dolok Village) Siabu District, Mandailing Natal Regency	Selupu Rejang District, Rejang Lebong Regency	Sarang Mandi Village, Sungai Selan Subdistrict, Central Bangka Regency
Blangkejeren Subdistrict, Gayo Lues Regency	Siantar District, Simalungun Regency and Asahan Regency	Talangrasau Village, Lais District, North Bengkulu Regency	Tempilang Subdistrict, (Mayang Village) Simpang Teritip Subdistrict, (Ranggi Asam Village) Jebus Subdistrict, West Bangka Regency
Takengon City, Central Aceh Regency	Paran Julu Village, Sipirok District, South Tapanuli Regency	Tanjung Aur Village, Air Padang District, North Bengkulu Regency	Pangkalpinang City
		Talo District, Humbang Seluma Regency	

Source: Mahsun (2017)

In 2019, thirty-three languages were identified on the island of Sumatra. According to the Language Mapping Study in Indonesia conducted by The Agency for Language Development and Cultivation (Badan Pengembangan dan Pembinaan Bahasa), these languages are distributed across various provinces: 15 languages in Bengkulu Province, 9 languages in North Sumatra Province, 15 languages in Bengkulu Province, and 6 languages in the Bangka Belitung Islands Province (Mahsun et al., 2017). The figure illustrates the distribution of language areas that served as sampling sites in Aceh Province, North Sumatra Province, Bengkulu Province, and Bangka Belitung Province.

Figure 1 illustrates the distribution of language areas that served as sampling sites in Aceh, North Sumatra, Bengkulu, and Bangka Belitung Provinces. The map highlights the linguistic diversity within local communities. This diversity raises questions regarding the similarities and differences among the languages under study.

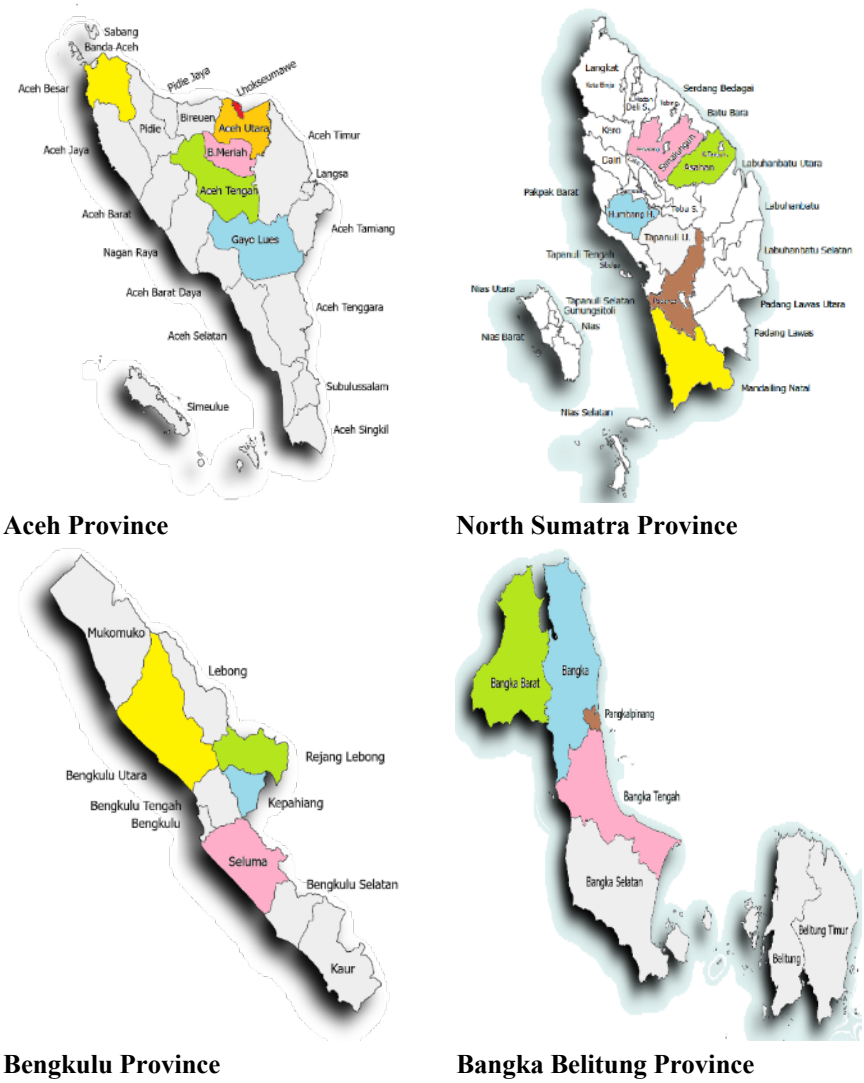

Fig 1. Customs Map of Sumatra Island
Source: https://paintmaps.com/

Previous research has generally been limited to comparing several languages in one or two provinces in Sumatra or analysing purely phonological aspects. However, there are many relevant studies for this research, including: (1) research by Zhang and Gong (2016) entitled "How Many Is Enough? -Statistical Principles for Lexicostatistics" found that smaller vocabulary lists, like the Swadesh 100-word list, are as effective as more extensive lists in lexicostatistics, making them practical for historical linguistic analysis, (2) Nefaa (2024), entitled "Genetic relatedness of Tunisian Sign Language and French Sign Language", explored the genetic relationship between Tunisian Sign Language (LST) and French Sign Language (LSF), uncovering significant lexical similarities through regional variations in sign usage were noted. Similarly, (3) Meliana et al. (2024), entitled "Tracing the Linguistic Roots of Malay and Batak Languages in Sumatra Island: A Historical Comparative Study", investigated the linguistic roots of Malay and Batak languages using lexicostatistical and glottochronological methods. It identifies an average kinship of 22.66% between Bengkulu and North Sumatra language pairs, classifying them as "Family stock".

This study aims to fill the gap of previous research by observing and analyzing the kinship relations of indigenous languages in four provinces on the island of Sumatra, namely Aceh, North Sumatra, Bengkulu, and Bangka Belitung. This research offers an update by focusing on the comparison and kinship relations of languages in the wider Sumatra region. Although previous studies on "Lexical Kinship Analysis using Lexicostatistical and Glotochronological Methods" have made significant contributions, the literature review shows a shortcoming in the coverage of the research area. Therefore, the research questions are formulated as follows:

1. To what extent do quantitative and qualitative analyses reveal the kinship relationships between the Acehnese (AT) and Gayo (GT) languages in Aceh, the Batak Toba (BTT) and Batak Mandailing (BMT) languages in North Sumatra, the Rejang (RT) and Serawai (ST) languages in Bengkulu, and the Melayu Bangka (MBT) and Kayu Agung (KAT) languages in Bangka Belitung?
2. What empirical evidence can be provided to determine the divergence time of these language pairs?
3. How can the languages be classified into groups, and what proportions of kinship relationships can be identified among the languages in Aceh, North Sumatra, Bengkulu, and Bangka Belitung?

LITERATURE REVIEW

Comparative Historical Linguistics

The comparative method, developed in the 19th century, played a crucial role in reconstructing proto-languages. Comparative-historical linguistics examines relationships between languages by comparing vocabulary and grammar to trace their historical development (Starostin, 1999). This branch of linguistics, or comparative linguistics, seeks to identify genetic links between languages and reconstruct shared ancestors (Zafar, 2023). It uses lexicostatistics to analyze vocabulary and glottochronology to estimate language divergence (Makkawaru & Hendrokumoro, 2022). By comparing structures, vocabulary, and phonetics, this field reveals connections between languages and tracks their evolution (Mahriyuni et al., 2023). Identifying linguistic similarities and differences provides insights into language development and relationships (Anayati et al., 2022).

The main goal of comparative linguistics is to establish genetic links between languages, indicating a common proto-language (Nixon, 2022). It also examines language changes, focusing on shifts within specific periods (Kumala & Lauder, 2021), and seeks to classify languages within various typologies (Reagan, 2021). According to Keraf (1952), this field identifies cognate languages by comparing elements that indicate kinship within a language family (A'laikum & Ermanto, 2023). The concept of "language family" has evolved, affecting how cognates are identified. The traditional view assumes regular sound changes in spoken languages (Nefaa, 2023). This approach analyzes cognate sets and shared traits, showing that linguistic similarities reflect systematic inheritances from a common proto-language (Dardanila et al., 2023). Comparative-historical methods are vital for reconstructing language histories, focusing on comparative-historical grammars and etymological dictionaries (Allamuratova, 2024).

Lexicostatistics and Glottochronology

In historical linguistics, lexicostatistics estimates the percentage of lexical cognates between languages, while glottochronology calculates the approximate date of language separation (Tao et al., 2023). These methods use vocabulary to infer historical language relationships (McMahon & McMahon, 2012). Developed by Swadesh (1954), Lexicostatistics analyzes essential vocabulary retention to determine genetic language relationships (Swadesh, 1955; Starostin, 2000). It compares shared cognates words with similar meanings and origins indicating closer linguistic ties (Zhang & Gong, 2016; Ntelu & Djou, 2017; Arokoyo & Lagunju, 2019; Rahmawati, 2022). This approach quickly assesses linguistic connections using small datasets (Yu et al., 2020).

Most lexicostatistical studies use pairwise comparisons to calculate the proportion of shared cognates and construct linguistic trees, estimating the time depth of relationships (Grant, 2010; Gapur et al., 2018; Parmini et al., 2023; Parmini, 2024). A higher percentage of cognates signals a more recent divergence from a common ancestor (Reagan, 2021; Nefaa, 2023). Although lexicostatistics faces criticism for its assumptions and limitations, it remains valuable for understanding language evolution (Troike, 1969; Onuoha & Esther, 2020; Harianto et al., 2021). Glottochronology, also developed by Swadesh, measures the rate of cognate replacement to estimate the time of language divergence (Ratcliffe, 2020; Petroni & Serva, 2011; Zulham et al., 2022). Modern methods incorporate computational tools to improve accuracy and objectivity (Bast, 2015; Rama & Wichmann, 2020; Rozov, 2022). Despite criticisms, glottochronology provides insights into genealogical relationships and language divergence (Muñoz, 2018).

METHODS

Research Design and Informants

This research applies an inductive thinking design approach, in which various linguistic phenomena found in the field are analyzed using relevant theories and methods to achieve the research objectives (Liu, 2016; Ghanbar et al., 2024). The research methods include qualitative and quantitative approaches for a more comprehensive understanding. The data in this study was collected from four provinces strategically selected to ensure a broad scope of representation and increase the reliability of the findings. The locations were chosen for their diverse social, cultural, and linguistic characteristics, allowing for a more holistic and in-depth study. This reduces sample bias and improves the generalizability of the findings (Sovacool et al., 2018; Gonzales, 2024; Lim, 2024).

The Aceh language is spoken by Jalan Hutan Kota, Tibang residents in Syiah Kuala District, Banda Aceh City. Meanwhile, the Gayo language is used by the community living in Jalan Lebe Kader, Bebesen District, Central Aceh Regency. In North Sumatra, the Batak languages of Toba and Mandailing are spoken by Batak tribes in various locations across Medan. Similarly, in Bengkulu Province, the Rejang and Serawai languages are still actively used in daily communication by native speakers residing in Selebar District, Bengkulu City. In the Bangka Belitung Islands Province, the Kayu Agung language is spoken in Kimak Village, Merawang District, Bangka Regency. At the same time, Melayu Bangka serves as the primary language for most residents in Bangka Regency.

The informants in this study consisted of 41 participants representing eight regional languages from the provinces of Aceh, North Sumatra, Bengkulu, and Bangka Belitung. Each

language group consisted of five to six native speakers, with ages ranging from 30 to 75 years old. Both male and female informants were included to ensure variation in speech data. Of the total participants, 19 were male and 22 were female, allowing for a balanced representation of gender-based linguistic variations. All participants were native speakers who actively used their respective languages in daily communication. Meanwhile, supporting informants are individuals who play a role in providing additional information related to the use of the language. The requirements for determining informants: 1) male or female; 2) aged between 30-75 years (not senile); 3) a native of an area where the language is actively used in daily communication; 4) having knowledge of the language; 5) understanding Bahasa Indonesia, and 6) physically and mentally healthy. Table 3 presents the data of the informants.

Table 3. Informant Data

Native Speakers	Region of Origin	Number of Informants	Gender (M/F)	Age Range (Years)
Aceh (AT)	Banda Aceh, Aceh Utara, and Lhokseumawe	5	5/0	30–59
Gayo (GT)	Bener Meriah, Aceh Tengah, Takengon, and Blangkejeren	6	4/2	30–54
Batak Toba (BTT)	Dolok Sanggul, Kisaran, and Siantar	5	2/3	51–75
Batak Mandailing (BMT)	Panyabungan, Lumban Dolok, and Paranjulu	5	2/3	31–60
Rejang (RT)	Tertik, Talangrasau, Selupu Rejang, and Tanjung Aur	5	1/4	33–47
Serawai (ST)	Talo and Seluma	5	3/2	31–64
Melayu Bangka (MBT)	Mayang, Ranggi Asam, Tempilang, and Pangkal Pinang	5	1/4	38–52
Kayu Agung (KAT)	Sarangmandi and Kimak	5	3/2	31–65

Data Sources and Data Collection

This study employed interview and note-taking techniques as the primary methods of data collection. In-depth interviews were conducted with informants to obtain accurate and authentic linguistic data from native speakers. The interview process was systematically designed based on a pre-established set of questions, focusing on eliciting basic vocabulary, phonological forms, and variations in language use within daily communication. Informants were selected according to specific research criteria, with careful consideration of details such as location, timing, recording methods, and participant consent to ensure data validity Virella & Woulfin, 2024). In addition, note-taking techniques were used to document and organize interview results systematically, ensuring accuracy and consistency before analysis (Williams & McWilliams, 2024; Oksanen, 2024). In the linguistic context, this study collected 200 basic vocabulary items based on the Swadesh list through direct and telephone interviews with native speakers.

The main limitation of this study lies in the data collection process, which was conducted with several informants through telephone interviews. Due to the geographical distance and wide distribution of informants across various provinces, the researcher was unable to conduct face-to-face interviews at every location. This limitation may have affected the depth of interaction and

the accuracy of the phonetic data obtained, as direct observation and acoustic verification could not be performed. Therefore, future research is recommended to conduct on-site data collection or utilize video-based interviews to obtain more accurate data and enable more in-depth linguistic analysis.

Data Analysis

This study applied qualitative and quantitative data analysis methods (Schoonenboom, 2023). Lexicostatistics prioritizes statistical vocabulary observations to determine language classification based on lexical similarities and differences (Zhang & Gong, 2016: Rakgogo & Mandende, 2023). In contrast, glottochronology focuses on calculating the temporal depth or age of related languages to support grouping within the scope of historical linguistics (McMahon & McMahon, 2012; Makkawaru & Hendrokumoro, 2022). Genetic links between languages indicate a common origin or proto-language, with the primary goal of comparative linguistics being to establish language families, reconstruct proto-languages, and identify linguistic changes that resulted in documented language forms (Zafar, 2023; Zafar, 2024).

This research procedure includes the following stages: (1) Creation of a data table containing 200 Swadesh vocabularies; (2) Data collection from interviewees (native speakers) through interviews to obtain relevant data; (3) Analysis and grouping of language pairs based on the data obtained, which includes: a) word pairs with identical correspondence, b) word pairs with phonemic correspondence, c) phonetically similar word pairs, and d) word pairs that differ in one phoneme; 4) Gloss calculation using lexicostatistics and glottochronology methods; 5) Error term calculation to evaluate the accuracy of the data obtained during the translation and data collection process; and 6) Similar vocabulary tagging based on kinship system classification, including language (dialect), family, stock, micro phylum, meso phylum, and macro phylum (Swadesh, 1952; Swadesh, 1955).

Table 4 presents the classification of kinship systems in languages based on temporal depth and percentage of cognates, as determined by lexicostatistical and glottochronological calculations. This classification categorizes languages into different levels, ranging from dialects with a temporal depth of 0-5 centuries and a cognate percentage of 81-100% to macro-phylums with a temporal depth of over 100 centuries and a cognate percentage of 1% to less than 1%. This classification system aids in understanding the relationships between languages at various levels of relatedness and temporal depth. It provides a framework for categorizing languages based on their historical connections and the percentage of cognates they share. Researchers can utilize this classification to study language evolution, historical linguistics, and language preservation efforts.

Table 4. Classification of Language Kinship Systems

Language Level	Time-depth (in centuries)	Cognate percentage %
Dialect	0-5	81-100%
Language Family	5-25	36-81%
Language Stock	25-50	12-36%
Micro Phylum	50-75	4-12%
Meso Phylum	75-100	1-4%
Macro Phylum	100- and above	1 – less than 1%

Source: Makkawaru & Hendrokumoro, 2022

This study used a method to calculate the percentage of linguistic kinship based on the formula proposed by Morris Swadesh (1954). The method includes using the lexicostatistical formula to calculate the percentage of kinship between languages and the glottochronology formula to estimate the time of separation between languages. The glottochronology formula applied is what Keraf proposed in the 1952. The detailed analysis procedure is explained as follows.

$$C = \frac{Vt}{Vd} \times 100\ \%$$

Description:
C : Percentage of language kinship
Vt : Dependent variable
Vd : Basic variable

$$W = \frac{log.C}{2\ log.r}$$

Description:
W : Separation time (*time in depth*)
C : Percentage of language kinship
r : Retention in 1.000 years, retention 80,5% rounded to 81%
log: Logarithm

RESULT AND DISCUSSION

The discussion of this research begins by displaying a list of 200 basic Swadesh vocabularies consisting of language groupings in the provinces of Aceh, North Sumatra, Bengkulu, and Bangka Belitung with two pairs of languages that dominate in the province, such as GT, AT, BTT, BMT, RT, ST, KAT, and MBT. It shows the percentage of linguistic kinship among the eight languages. To get the percentage of these languages using the lexicostatistical method, looking for similarities or similarities in lexicon words both in form and meaning, which are described in the table as follows.

Table 5. Language Classification Based on Calculation Result

Language/Tribe Name	Kinship Word	Cognate Percentage %	Kinship Relationship
Gayo (GT) – Aceh (AT)	33	16,5%	Language Stock
Toba (BTT) – Mandailing (BMT)	98	49%	Language Family
Rejang (RT) – Serawai (ST)	61	30,5%	Language Stock
Kayu Agung (KAT) – Melayu Bangka (MBT)	125	62,55%	Language Family

Table 5 shows the grouping of languages based on the list of 200 basic Swadesh vocabularies among several pairs of languages that dominate in the province, namely GT, AT, BTT, BMT, RT, ST, KAT, and MBT. The languages are paired according to their neighboring provinces, such as Bengkulu Province - Aceh, Bengkulu Province - North Sumatra, Bengkulu Province - Bangka Belitung, Aceh Province - North Sumatra, Aceh Province - Bangka Belitung, and North Sumatra Province - Bangka Belitung. This shows the percentage of kinship of the eight languages based on pairs of provinces, to get the percentage of these languages using the lexicostatistical method, looking for similarities or similarities in lexicon words both in form and meaning described in Table 6.

Table 6. Language Classification Based on Calculation Result

Language/Tribe Name	Kinship Word	Cognate Percentage %	Kinship Relationship
Aceh – North Sumatra	44	22%	Language Stock
Aceh – Bengkulu	61	30,5%	Language Stock
Aceh – Bangka Belitung	53	26,5%	Language Stock
North Sumatra – Bengkulu	34	17%	Language Stock
Bengkulu – Bangka Belitung	101	50,5%	Language Family
North Sumatra – Bangka Belitung	46	23%	Language Stock

Time-Depth (W1)

In the 1950s, the theory of lexicostatistics evolved into the science of glottochronology by proposing a mathematical formula to determine the time when two languages separated. Based on the percentage of the core vocabulary of culturally independent words (Zafar, 2023; Meliana et al., 2024; Hendrokumoro et al., 2024). After determining the percentage of kinship between the two languages using the lexicostatistics formula, the next step is to estimate their time of separation using the glottochronology formula, as explained below.

a. Time of separation between languages in the provinces of Aceh - Bengkulu, North Sumatra - Bengkulu, and Bengkulu - Bangka Belitung.

Aceh - North Sumatra

$$W = \frac{log.C}{2\ log.r}$$
$$= \frac{log.0{,}22}{2\ log.81}$$
$$= \frac{-0{,}657}{-0{,}183}$$
$$= 3{,}590\ (x\ 1000)$$
$$= 3.590\ \text{tahun}$$

Aceh - Bengkulu

$$W = \frac{log.C}{2\ log.r}$$
$$= \frac{log.\ 0{,}305}{2\ log.81}$$
$$= \frac{-0{,}515}{-0{,}183}$$
$$= 2{,}814\ (x\ 1000)$$
$$= 2814\ \text{tahun}$$

Aceh - Bangka Belitung

$$W = \frac{log.C}{2\ log.r}$$
$$= \frac{log.0{,}265}{2\ log.81}$$
$$= \frac{-0{,}576}{-0{,}183}$$
$$= 3{,}147\ (x\ 1000)$$
$$= 3.147\ \text{tahun}$$

The calculation of the separation time is multiplied by 1000 so that the results of the calculation of the initial separation time in the languages of the Provinces of Aceh - Bengkulu, North Sumatra - Bengkulu, and Bengkulu - Bangka Belitung are shown in the previous calculation. In other words, the calculation of the initial separation time of the languages of the four pairs of provinces can be stated as follows: (a) The languages in Aceh and Bengkulu Provinces are estimated to have become one language around 2814 years, North Sumatra and Bengkulu around 3433 years, and Bengkulu and Bangka Belitung around 1617 years ago; (b) The languages in Aceh and Bengkulu Provinces are estimated to have cognate words around the 790th century BC, North Sumatra and Bengkulu around the 1409th century BC, and the languages in Bengkulu and Bangka Belitung Provinces around the 409th century AD (calculated in 2024).

b. Time of separation between languages in the provinces of Aceh - North Sumatra, Aceh - Bangka Belitung, dan North Sumatra - Bangka Belitung

North Sumatra - Bengkulu
$W = \frac{log.C}{2\ log.r}$
$= \frac{log.\ 0{,}17}{2\ log.81}$
$= \frac{-\ 0{,}769}{-\ 0{,}183}$
= 3,433 (x 1000)
= 3433 years

North Sumatra - Bengkulu
$W = \frac{log.C}{2\ log.r}$
$= \frac{log.\ 0{,}17}{2\ log.81}$
$= \frac{-\ 0{,}769}{-\ 0{,}183}$
= 3,433 (x 1000)
= 3433 years

North Sumatra - Bangka Belitung
$W = \frac{log.C}{2\ log.r}$
$= \frac{log.0{,}23}{2\ log.81}$
$= \frac{-\ 0{,}638}{-\ 0{,}183}$
= 3,486 (x 1000)
= 3.486 years

The calculation of the separation time is multiplied by 1000 so that the results of the initial calculation of the separation time in the languages of Aceh - North Sumatra, Aceh - Bangka Belitung, and North Sumatra - Bangka Belitung Provinces are shown in the previous calculation. In other words, the calculation of the initial separation time in the languages of the four pairs of provinces can be stated as follows: (a) The languages in the provinces of Aceh and North Sumatra are estimated to become one language around 3590 years, Aceh and Bangka Belitung around 3147 years, and North Sumatra and Bangka Belitung around 3486 years ago; (b) The languages of Aceh and North Sumatra are estimated to have separated cognate words around 1566 BC, Aceh and Bangka Belitung around 1123 BC, and North Sumatra and Bangka Belitung around 1462 BC (calculated in 2024).

Range of Time-Depth Error and Updated Cognate Percentage

The previous calculation could have provided a precise estimate of the year of separation of the two languages. Therefore, more specific calculations are needed to avoid the errors that occurred in the initial calculations. Advanced statistical techniques are still required to achieve more accurate results. This technique calculates the error period. The next step is to find the error range to determine a more accurate separation time using the formula below. Calculating error ranges between languages in Aceh - Bengkulu, North Sumatra - Bengkulu, and Bengkulu - Bangka Belitung Provinces.

Aceh – Bengkulu
Error periods = $W_1 - W_2$
= 2.814 – 2.579
= 235 years

North Sumatra – Bengkulu
Error periods = $W_1 - W_2$
= 3.433 – 3.863
= - 430 years

Bengkulu – Bangka Belitung
Error periods = $W_1 - W_2$
= 1.617 – 1.459
= 158 years

a. Calculating error ranges between languages in Aceh - North Sumatra, Aceh - Bangka Belitung, dan North Sumatra - Bangka Belitung

Aceh - North Sumatra
Error periods = $W_1 - W_2$
= 3.590 – 3.295
= 295 years

Aceh - Bangka Belitung
Error periods = $W_1 - W_2$
= 3.147 – 2.885
= 262 years

North Sumatra - Bangka Belitung
Error periods = $W_1 - W_2$
= 3.486 – 3.453
= 33 years

b. The results of the calculation of the error range to determine a more appropriate separation time concluded at age between the languages in Aceh-Bengkulu, North Sumatra-Bengkulu, Bengkulu-Bangka Belitung, Aceh-North Sumatra, Aceh-Banga Belitung, and North Sumatra-Bangka Belitung provinces can be expressed as follows:
   1. It is estimated that a single language has been formed into one language between the languages of Aceh-Bengkulu about (2,814 - 2,579), North Sumatra-Bengkulu (3,433 - 3,863), Bengkulu-Bangka Belitung (1,617 - 1,459), Aceh-North Sumatra (3,590 - 3,295), Aceh-Banga Belitung (3,147 - 2,885), and North Sumatra-Bangka Belitung (3,486 - 3,453) years ago.
   2. Estimated language family between the languages of Aceh-BKL about (2,814 - 2,579), North Sumatra-Bengkulu (3,433 - 3,863), Bengkulu-Bangka Belitung (1,617 - 1,459), Aceh-North Sumatra (3,590 - 3,295), Aceh-Banga Belitung (3,147 - 2,885), and North Sumatra-Bangka Belitung (3,486 - 3,453) years ago.
   3. It is estimated that it began to separate from Proto language in the four pairs of languages in two provinces around Aceh-Bengkulu around (555-790) BC, North Sumatra-Bengkulu (1271-1839) BC, Bengkulu-Bangka Belitung (407-565) AD, Aceh-North Sumatra (1271-1566) BC, Aceh-Banga Belitung (861-1123) BC, and North Sumatra-Bangka Belitung (1429-1462) BC.

Cognate Percentage Among Aceh, North Sumatra, Bengkulu, and Bangka Belitung Provinces

After being calculated using the lexicostatistical formula to determine the percentage of kinship between languages in Aceh and North Sumatra Province, the results obtained show that there are 44 pairs of words (22%) that fall into the “Language stock” category. In the comparison between Aceh and Bengkulu Province, the percentage of language kinship shows that there are 61 pairs of words (30.5%) that fall into the “Language stock” category. The results obtained from the comparison between Aceh and Bangka Belitung Province show that there are 53 pairs of words (26.5%) that fall into the “Language stock” category. These word pairs are identically correlated, phonetically compatible, and differ in only one phoneme. Furthermore, in the comparison between Bengkulu and North Sumatra Province, the percentage of language kinship shows 34 pairs of words (17%) that fall into the “Language stock” category. The results of the percentage of kinship between Bengkulu and Bangka Belitung Province show 101 pairs of words (50.5%) that fall into the “Language family” category. In the comparison between North Sumatra and Bangka Belitung Province, the percentage of language kinship shows 46 pairs of words (23%) that fall into the “Language stock” category, with word pairs that correlate identically and only differ by one phoneme. The total of 200 gloss pairs can be seen in Table X, which consists of several categories or kinship criteria.

Identically Correlated Word Pairs

Identically corresponding or correlated word pairs are two words that have the same or very similar meaning relationships and substitute each other in certain contexts. It indicates that the basic vocabulary being compared is built from the same phoneme elements and contains the same meaning. In the kinship comparison between Aceh Province and North Sumatra, which includes the languages of GT, AT, BTT, and BMT, 13-word pairs (6.5%) were found. Aceh and Bengkulu

provinces, involving GT, AT, RT, and ST, found 21 pairs of words (10.5%). The following is an example of identically related word pairs based on language pairs in the province, as shown in the following Table 7.

Table 7. Identical Word Pairs Between Aceh-North Sumatra and Aceh-Bengkulu Province

Gloss	Aceh-North Sumatra Province Language Pairs				Aceh-Bengkulu Province Language Pairs			
	AT	GT	BTT	BMT	AT	GT	RT	ST
Garam	[sira]		[sira]	[sira]				
Asap					[asap]	[asap]		[asap]

The comparison between Aceh and Bangka Belitung Province, which includes the languages of GT, AT, KAT, and MBT, shows 31 pairs of words (15.5%). Furthermore, in the kinship comparison between North Sumatra and Bengkulu Province, which involves BTT, BMT, RT, and ST, there are 17 pairs of words (8,5%). The following is an example of identically related word pairs based on language pairs in the province, as shown in Table 8.

Table 8. Identical Word Pairs Between Aceh-Bangka Belitung and North Sumatra-Bengkulu Province

Gloss	Aceh-Bangka Belitung Province Language Pairs				North Sumatra-Bengkulu Province Language Pairs			
	AT	GT	KAT	MBT	BTT	BMT	RT	ST
Sepatu	[səpatu]	[səpatu]	[səpatu]	[səpatu]				
Tipis					[tipis]	[tipis]	[tipis]	[tipis]

The comparison between Bengkulu and Bangka Belitung provinces, which include RT, ST, KAT, and MBT, 47 pairs of words (23.5%) were found. A comparison between North Sumatra and Bangka Belitung Province, which includes the languages of BTT, BMT, KAT, and MBT, shows 17 pairs of words (8,5%). The following is an example of identically related word pairs based on language pairs in the province, as shown in Table 9.

Table 9. Identical Word Pairs Between Bengkulu-Bangka Belitung provinces and North Sumatra-Bangka Belitung Province

Gloss	Bengkulu-Bangka Belitung Province Language Pairs				North Sumatra- Bangka Belitung Province Language Pairs			
	RT	ST	KAT	MBT	BTT	BMT	KAT	MBT
Busuk	[busuk]		[busuk]		[busuk]	[busuk]	[busuk]	
Tirai	[tabir]	[tabir]	[tabir]	[tabir]	[tabir]	[tabir]	[tabir]	[tabir]

Phonetically Similar Word-Pair Correspondences

Phonetically similar word pairs are two words that sound almost the same or very similar, although they may have different meanings. These words sound similar when spoken, so they can often be confused. This study identifies several word pairs with phonetic similarities among various tribal languages in Sumatra. The analysis shows that there are 19 word pairs (9.5%) that are related between AT, GT, BTT, and BMT. In addition, 28 pairs of words (14%) were found to be related between AT, GT, RT, and ST. Then, there are 19 pairs of words (9.5%) that are related

between AT, GT, KAT, and MBT. A total of 23 pairs of words (11.5%) were found to have phonetic similarities between RT, ST, BTT, and BMT. The analysis also showed 40 pairs of words (20%) that are related between RT, ST, KAT, and MBT. Finally, there are 20 pairs of words (10%) that are related between BTT, BMT, KAT, and MBT. These findings highlight the existence of significant phonetic similarities among the various tribal languages, which may contribute to further understanding of the linguistic relationships and language history of the region.

Table 10. Phonetically Similar Word Pairs Between Aceh and North Sumatra Province

Gloss	Phoneme Correspondence	Language Pairs 1		Language Pairs 2	
		AT	GT	BTT	BMT
Angin	/e/ ↔ /i/	[aɳen]			[aɳin]
Kulit		[kulet]	[kulit]		[kulit]
Ayam	/o/ ↔ /u/	[manok]		[manuk]	[manuk]
Bunga	/ə/ ↔ /a/		[buɳe]	[buɳa]	[buɳa]
Menutup			[mənutup]	[manutup]	[manutup]

Table 10 presents several pairs of words between Aceh and North Sumatra Provinces, highlighting phonetically similar correspondences in four languages. Several pairs of the words have different phonemes /e/ to /i/, for example, in the words [aɳen] and [kulet] in AT, which correspond to BMT [aɳin] and [kulit]. Furthermore, there is a difference in the phoneme /o/ to /u/ at the end of the word in AT language, namely [manok], [batok], [sibok], which corresponds to BTT and BMT language into [manuk], [batuk], [sibuk]. In other cases, there is also a difference in the phoneme /ə/ ↔ /a/, for example, [buɳe], [mənutup], and [kəreta], which change into [buɳa], [manutup], [kareta].

Table 11. Phonetically Similar Word Pairs Between Bengkulu Province and Bangka Belitung Province

Gloss	Phoneme Correspondence	Language Pairs 1		Language Pairs 2	
		RT	ST	KAT	MBT
Angin	/i/ ↔ /e/	[aɳin]	[aɳin]	[aɳen]	
Tajam	/a/ ↔ /ə/	[tajəm]	[tajam]	[tajəm]	[tajəm]
Sabar	/a/ ↔ /e/	[sabar]	[sabar]	[saber]	
Aku	/o/ ↔ /ə/	[ko]		[kə]	

Table 11 demonstrates that the gloss “wind” or *angin*, the final vowel /i/ in RT and ST [aɳin] corresponds to /e/ in KAT [aɳen]. Furthermore, the vowel /a/ alternates with schwa /ə/ in the gloss “sharp” or *tajam*, showing variation across RT, ST, KAT, and MBT such as [tajəm] into [tajəm]. Similarly, in “patience” or *sabar,* the final vowel /a/ in RT and ST corresponds to /e/ in KAT such as [sabar] into [saber]. Finally, the pronoun “I” or *aku* demonstrate a correspondence between /o/ in RT and schwa /ə/ in KAT and MBT such as [ko] into [kə]. These patterns indicate systematic vowel shifts, particularly involving /i/, /a/, and /o/, which tend to correspond with /e/ or schwa / ə/ in other dialects.

Repeated Phoneme Correspondences in Word Pairs

Word pairs associated with repeated phoneme correspondences usually refer to words that have similar or repeated sound patterns. Word pairs associated with repeated phoneme correspondences usually refer to words that have similar or repeated sound patterns. In the analysis of repeated word pairs that show kinship, GT, AT, BTT, and BMT languages only have seven pairs of words (3.5%). Meanwhile, for GT, AT, RT, and ST, 17 pairs of words (8.5%) were found to be related due to regular phonetic correspondence. In GT, AT, KAT, and MBT, there are 17 pairs of words (8.5%) that are phonetically related. Whereas in RT, ST, BTT, and BMT, 15 pairs of words (7.5%) with similar relationships were found. RT and ST languages show 22 pairs of words (11%) that correspond phonetically. Finally, in BTT, BMT, KAT, and MBT, there are only 13 pairs of words (6.5%) that are related because they correspond regularly.

Table 12. Repeated Phoneme Word Pairs Between Bengkulu-Bangka Belitung Province and Sumatra Utara-Bangka Belitung Province

Gloss	Bengkulu-Bangka Belitung Province Language Pairs				North Sumatra-Bangka Belitung Province Language Pairs			
	RT	ST	KAT	MT	BTT	BMT	KAT	MBT
Debu	[dəbu]	[dəbu]	[dəbu]	[dəbu]				
Sibuk	[sibuk]	[sibuk]	[sibuk]	[sibuk]				
Tulang	[tulaŋ]	[tulaŋ]	[tulaŋ]	[tulaŋ]				
Batuk					[batuk]	[batuk]	[batuk]	
Manis					[manis]	[manis]	[manis]	[manis]
Minum					[minum]	[minum]		[minum]

Word Pairs with One Phoneme Different

One-phoneme word pairs are two words that have only one sound or phoneme difference in their composition. These words are often used in word games or to show a significant change in meaning by changing only one phoneme. In the vocabulary comparison between AT, GT, BTT, and BMT, 22 pairs of words (11%) were obtained. Vocabulary comparison between AT, GT, RT, and ST resulted in 39 pairs of words (19.5%). The vocabulary comparison between AT, GT, KAT, and MBT shows 31 pairs of words (15.5%). Furthermore, the comparison of language vocabulary between BTT, BMT, RT, and ST resulted in 23 pairs of words (11.5%). The comparison between RT, ST, KAT, and MBT shows 65 pairs of words (32.5%). Finally, vocabulary comparison between BTT, BMT, KAT, and MBT resulted in 35 pairs of words (17.5%) that are related because they only differ by one phoneme. Examples of these word pairs can be seen in the following table.

Table 13. Different Phoneme Word Pairs Between Aceh-North Sumatera Province and Aceh-Bengkulu Province

Gloss	Aceh-North Sumatera Province Language Pairs				Aceh-Bengkulu Province Language Pairs			
	AT	GT	BTT	BMT	GT	AT	RT	ST
Ayam	[manok]		[manuk]	[manuk]				
Batuk	[batok]	[matuk]		[batuk]				
Mencuci						[nəsah]		[nyəsah]
Orang tua						[jema tuə]		[jəmo tuo]
Sempit					[səmpət]	[empət]	[səpit]	[səmpit]

Based on the table above, it can be observed that there is a phonemic variation in the gloss "chicken" or *ayam* in AT; this word is pronounced as [manok], while in BTT and BMT, it is pronounced as [manuk]. This phonetic difference lies in the use of the phoneme /o/ in AT and the phoneme /u/ in BTT and BMT. Furthermore, the gloss "cough" or *batuk* in AT is pronounced as [batok], using the phoneme /o/. Meanwhile, in GT, this gloss is pronounced as [matuk], with the initial phoneme changing from /b/ to /m/. Furthermore, the gloss "wash" or *mencuci* is pronounced as [nəsah] in GT and [nyəsah] in ST. This phonetic difference lies in the use of the phoneme /y/ in the second position in [nyəsah] in ST, which is absent in [nəsah] in GT. A similar phenomenon is also seen in the gloss "narrow" or *sempit*, which is pronounced as [empət] in GT and [səpit] in RT, with the difference in the initial phoneme /e/ in GT and /s/ in RT. On the gloss "parents" or *orangtua*, in GT, it is pronounced as [jema tuə], while in ST, it is pronounced as [jəmo tuo]. This phoneme variation is seen in the use of the phonemes /ə/ in GT and /o/ in ST. Examples of these word pairs can be seen in the following table.

Unrelated Word Pairs

Word pairs that do not have kinship or phoneme similarity and do not meet the four criteria for percentage kinship in the two languages being compared provide important insights in the study of comparative linguistics. In this study, it was found that in addition to related vocabulary, there is also unrelated vocabulary in the comparison between Aceh and North Sumatra Province. The languages compared in this study include AT, GT, BTT, and BMT, with a total of 33 pairs of unrelated words (16.5%). Furthermore, the comparison of unrelated vocabulary between Aceh and Bengkulu Province includes AT, GT, RT, and ST, with 139 pairs of unrelated words (69.5%) or about 70%. For the comparison between Aceh and Bangka Belitung Province, involving the languages of AT, GT, KAT, and MT, 147 unrelated word pairs (73.5%) were found. Furthermore, in the comparison between North Sumatra Province and Bengkulu Province, involving the languages of BTT, BMT, RT, and ST, 166 pairs of words (83%) were found that have no kinship relationship. The comparison of unrelated vocabulary between Bengkulu and Bangka Belitung Province, which includes the languages of RT, ST, KAT, and MT, shows 99 pairs of words (49.5%) that are unrelated. Finally, in a comparison between North Sumatra and Bangka Belitung Province, involving the languages of BTT, BMT, KAT, and MT, 154 pairs of words (77%) were found to be unrelated. Examples of these unrelated word pairs can be seen in the tables presented below.

Tabel 14. Unrelated Word Pairs Between Aceh-North Sumatra Province and Aceh and Bengkulu Province

Gloss	Aceh-North Sumatra Province Language Pairs				Aceh-Bengkulu Province Language Pairs			
	AT	GT	BTT	BMT	AT	GT	RT	ST
Bangun	[bəudoh]	[uwət]	[duɳo]	[ɳot]				
Buruk	[brok]	[kotek]	[roa]	[jat]				
Lihat	[əu]	[əɳon]	[bereŋ]	[ligin]				
Besar					[rayəuk]	[kul]	[lai/Lay]	[bəsak]
Bohong					[səumulət]	[cogah]	[ŋəkə]	[pəmbuo'ŋ]
Banyak bicara					[cubreə]	[caroŋ]	[reŋeh]	[buas]

CONCLUSION

This study uncovers patterns of kinship and the dynamics of linguistic evolution among local languages in Aceh, North Sumatra, Bengkulu, and Bangka Belitung, using lexicostatistics and glottochronology. The analysis of 200 basic vocabulary items shows that the degrees of kinship among the language pairs vary considerably: Acehnese - Gayo at 16.5%, Rejang - Serawai at 30.5%, Toba - Mandailing at 49%, and Kayu Agung - Bangka Malay at 62.55%. These percentages place the Acehnese - Gayo and Rejang - Serawai pairs in the language stock category, indicating relatively distant relationships, while the Toba - Mandailing and Kayu Agung - Bangka Malay pairs fall into the language family category, reflecting closer genealogical ties. Qualitatively, these relationships are evident not only from the cognate percentages but also from identical lexical matches, recurring patterns of phonemic correspondence, phonetic similarities, and consistent one-phoneme differences, all of which demonstrate systematic sound change patterns.

The glottochronological calculations show that the divergence times among the languages range from approximately 430 BCE to 3,590 CE when converted. These findings indicate that most of the languages began to separate from their proto-language in very ancient periods, particularly for the Aceh - North Sumatra, North Sumatra - Bangka Belitung, and Aceh - Bangka Belitung pairs. In contrast, a more recent split is observed in the Bengkulu - Bangka Belitung pair. The language grouping based on cognate counts also shows that the Bengkulu - Bangka Belitung pair has the highest degree of kinship at 50.5%, placing it in the language family category. In contrast, the other interprovincial pairs fall within the 17-30.5 percent range and belong to the language stock category, indicating more distant genetic relationships despite originating from the same proto-language. These results demonstrate that the languages share interconnected historical relationships and reflect long-term linguistic evolution across Sumatra. The differences in divergence times illustrate the impact of migration, cultural contact, and community mobility on the region's linguistic history.

Previous research conducted by Meliana et al. (2024) only covered two provinces in Sumatra with three indigenous languages. This study expands the coverage to four provinces by taking each province's two main authentic languages. In the data collection, the vocabulary was expanded to 200 Swadesh words. This study builds on the findings of Zhang & Gong (2016), who underlined that the size of a vocabulary list, such as the Swadesh list, has an important role

and advocated applying more rigorous statistical methods to improve the reliability of language kinship identification by improving data collection. This study makes a significant contribution to historical-comparative linguistics in Indonesia by expanding the geographic scope and enriching the data on genealogical relationships among regional languages in Sumatra. The findings reinforce the central role of language in preserving cultural identity and heritage, while also demonstrating how socio-cultural dynamics shape language evolution. The recommendations include integrating local languages into education, strengthening further research, enhancing interregional collaboration, and developing digital tools to support language preservation.

NOTE

We would like to thank two anonymous reviewers for their valuable comments on the earlier draft. We also extend our gratitude to the informants for their contributions to the data-collection process.

REFERENCES

A'laikum, A., & Ermanto. (2023). Kekerabatan Bahasa Minangkabau di Nagari Mungo Kecamatan Luak Kabupaten Lima Puluh Kota dan Bahasa Melayu Riau di Desa Buantan Besar Kecamatan Siak Sri Indrapura Kabupaten Siak. *PERSONA: Language and Literary Studies*, *2*(2), 166–176.

Anayati, W., Wardana, M. K., Mayasari, M., & Purwarno, P. (2022). Lexicostatistics of Malay and Malagasy Languages: Comparative Historical Linguistic Study. *English Review: Journal of English Education*, *10*(3), 875–882. https://doi.org/10.25134/erjee.v10i3.6690

Arokoyo, B. E., & Lagunju, O. O. (2019). A Lexicostatistics Comparison of Standard Yorùbá, Àkúrẹ́ and Ìkàrẹ́ Àkókó Dialects. *Journal of Universal Language*, *20*(2), 1–27. https://doi.org/10.22425/jul.2019.20.2.1

Bast, F. (2015). Time Calibration of Linguistic Phylograms: A Molecular Clock for Historical Linguistics. *Journal of Phylogenetics & Evolutionary Biology*, *03*(03). https://doi.org/10.4172/2329-9002.1000e115

Bastardas-Boada, A. (2018). The Ecology of Language Contact: Minority and Majority Languages. In *The Routledge Handbook of Ecolinguistics* (pp. 57–76). Routledge: Taylor & Francis.

Creswell, J. W., & Creswell, J. D. (2018). Research Design: Qualitative, Quantitative, and Mixed Methods Approaches. Fifth Edition. *SAGE Publications, Inc.* https://spada.uns.ac.id/pluginfile.php/510378/mod_resource/content/1/creswell.pdf

Dardanila, Widayati, D., & Gustianingsih. (2023). Language Kindship of Jamee, Gayo, and Malay. *Migration Letters*, *21*(2), 901–912. https://migrationletters.com/index.php/ml/article/view/6310

Dauletovna, A. N. (2024). Comparative Linguistics and Translation Studies. *World of Scientific News in Science International Journal, Germany*, *Vol 2*(Issue 2), 812–817. https://worldofresearch.ru/index.php/wsjc/article/view/310

Gapur, A., Siregar, D. S. P., & Pujiono, M. (2018). Language Kinship Between Mandarin, Hokkien Chinese and Japanese (Lexicostatistics Review). *Aksara*, *30*(2), 301. https://doi.org/10.29255/aksara.v30i2.267.301-318

Ghanbar, H., Cinaglia, C., Randez, R. A., & De Costa, P. I. (2024). A methodological synthesis of narrative inquiry research in applied linguistics: What's the story? *International Journal of Applied Linguistics*, ijal.12591. https://doi.org/10.1111/ijal.12591

Gonzales, W. D. W. (2024). The Holistic Advantage: Unified Quantitative Modeling for Less-Biased, In-Depth Insights into (Socio)Linguistic Variation. *Languages*, *9*(5), 182. https://doi.org/10.3390/languages9050182

Granić, J. (2025). *The (Non-)Acceptance of Otherness in Multilingual Settings* (pp. 247–261). https://doi.org/10.18485/dpls_plucast.2025.6.ch15

Grant, A. P. (2010). On using qualitative lexicostatistics to illuminate language history: Some techniques and case studies. *Diachronica*, *27*(2), 277–300. https://doi.org/10.1075/dia.27.2.06gra

Harianto, Zulfitri, & Amin, T. S. (2021). Lexicostatistics Study of Mandailing and Angkola Languages. *Jurnal Educatio*, *7*(1), PP. 265-275. https://doi.org/DOI: https://doi.org/10.31949/educatio.v7i1.850

Hendrokumoro, Darman, F., Nuraeni, N., & Ma'shumah, N. K. (2024). The genetic relationship between Alune, Lisabata, Luhu, and Wemale (Western Seram, Indonesia): A historical-comparative linguistics approach. *Cogent Arts & Humanities*, *11*(1), 2306718. https://doi.org/10.1080/23311983.2024.2306718

Istanti, W., Seinsiani, I. G., Visser, J. G., & Lazuardi, A. I. D. (2020). Comparative Analysis of Verbal Communication Vocabulary between Indonesian-Afrikaans for Foreign Language Teaching. *International Journal of Language Education*, *4*(3), 389–397. https://doi.org/10.26858/ijole.v4i3.15106

Kazakova, I., & Shakhnazaryan, V. (2020). Amazing Integration of Autochthonous Languages Into Allochthonous on The Example of Maorisms and Maisms. *ICERI2020 Proceedings*, 6711–6719. https://doi.org/10.21125/iceri.2020.1427

Kumala, S. A., & Lauder, M. R. (2021). Makna Toponim di Tangerang sebagai Representasi Keberadaan Etnis Cina Benteng: Sebuah Kajian Linguistik Historis Komparatif. *Ranah: Jurnal Kajian Bahasa*, *10*(2), 304. https://doi.org/10.26499/rnh.v10i2.4048

Lim, W. M. (2024). What Is Qualitative Research? An Overview and Guidelines. *ANZMAC: Australian and New Zealand Marketing Academy*, *1*(31). https://doi.org/DOI: 10.1177/14413582241264619

Liu, L. (2016). Using Generic Inductive Approach in Qualitative Educational Research: A Case Study Analysis. *Journal of Education and Learning*, *5*(2), 129. https://doi.org/10.5539/jel.v5n2p129

Liu, Y., Luo, W., & Wang, X. (2023). Exploring the relationship between students' note-taking and interpreting quality: A case study in the Chinese context. *Frontiers in Education*, *8*, 1157509. https://doi.org/10.3389/feduc.2023.1157509

Mahriyuni, Pramuniati, I. & Maftuhah, R.A. (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. (2017). *Metode Penelitian Bahasa: Tahapan, Strategi, Metode, dan Tekniknya*. RAJAWALI PERS.

Mahsun, Fernandez, I. Y., Laksono, K., Lauder, M. R., & Nadra. (2017). *Bahasa dan Peta Bahasa di Indonesia*. Badan Pengembangan dan Pembinaan Bahasa.

Mailani, O., Nuraeni, I., Syakila, S. A., & Lazuardi, J. (2022). Bahasa Sebagai Alat Komunikasi Dalam Kehidupan Manusia. *Kampret Journal*, *1*(2), 1–10. https://doi.org/10.35335/kampret.v1i1.8

Makkawaru, & Hendrokumoro. (2022). The Genetic Relationship between Bugis and Kaili. *Journal Educational Verkenning*, *Volume 3*(Issue 1), Pages 017-027. https://hdpublication.com/index.php/jev

McLeod, W. (2009). *A new multilingual United Kingdom? The impact of the European Charter for Regional or Minority Languages*. Palgrave Macmillan.

McMahon, A., & McMahon, R. (2012). Lexicostatistics and Glottochronology. In C. A. Chapelle (Ed.), *The Encyclopedia of Applied Linguistics* (1st ed.). Wiley. https://doi.org/10.1002/9781405198431.wbeal0701

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

Muñoz, J. (2018). De la glotocronología a la filogenética: Estado de la cuestión y los nuevos desarrollos en la metodología de clasificación lingüística. *Revista de Investigación Lingüística*, *21*, 170–184. https://doi.org/10.6018/ril.21.367611

Nefaa, A. (2023a). Genetic relatedness of Tunisian Sign Language and French Sign Language. *Frontiers in Communication*, *8*, 1201148. https://doi.org/10.3389/fcomm.2023.1201148

Nefaa, A. (2023b). Genetic relatedness of Tunisian Sign Language and French Sign Language. *Frontiers in Communication*, *8*, 1201148. https://doi.org/10.3389/fcomm.2023.1201148

Nixon, C. (2022). *Constructing Language and Comparative Linguistics*. Bibliotex Digital Library.

Ntelu, A., & Djou, D. N. (2017). The Language Family Relation of Local Languages in Gorontalo Province (A Lexicostatistic Study). *Journal of Arts and Humanities*, *6*(11), 48. https://doi.org/10.18533/journal.v6i11.1285

Oksanen, J. (2024). Designer-aligned Automated Interview Note-taking. *Aalto University School of Science: Master's Programme in International Design Business Management (MSc)*. https://aaltodoc.aalto.fi/items/eba3a690-c00f-4140-a026-b5f38e729bec

Onuoha, C. E., & Esther, C. (2020). Lexicostatistics Comparison of Standard Igbo and Achi Dialect. *Journal of Chinese & African Studies (JOCAS)*, *3*(1), 51–62. https://nigerianjournalsonline.com/index.php/JOCAS/article/view/4653/4517

Parmini, N. P. (2024). The Genetic Relationship Between Balinese and Madurese. *International Journal of Education, Vocational and Social Science*, *03*(01).

Parmini, N. P., Mawa, I. W., Soper, I. W., Suparta, I. M., Sueni, N. M., & Temaja, I. G. B. W. B. (2023). The Genetic Relationship Between Balinese and Madurese. *International Journal of Education, Vocational and Social Science*, *2*(1), 283–295. https://doi.org/10.99075/ijevss.v2i01.169

Petroni, F., & Serva, M. (2011). Automated Word Stability and Language Phylogeny*. *Journal of Quantitative Linguistics*, *18*(1), 53–62. https://doi.org/10.1080/09296174.2011.533589

Rahmawati, R. (2022). Proto Language Relationship with Mandailing Language. *Randwick International of Education and Linguistics Science Journal*, *3*(2), 362–367. https://doi.org/10.47175/rielsj.v3i2.482

Rakgogo, T. J., & Mandende, I. P. (2023). Lexical similarities between Khelobedu dialect and Tshivenḓa and Sepedi languages. *Literator*, *44*(1). https://doi.org/10.4102/lit.v44i1.1910

Rama, T., & Wichmann, S. (2020). A test of Generalized Bayesian dating: A new linguistic dating method. *PLOS ONE*, *15*(8), e0236522. https://doi.org/10.1371/journal.pone.0236522

Ratcliffe, R. R. (2020). *The Glottometrics of Arabic: Quantifying Linguistic Diversity and Correlating It With Diachronic Change*. *11*(1), 1–29. https://doi.org/10.1163/22105832-01001100

Reagan, T. (2021a). Historical Linguistics and the Case for Sign Language Families. *Sign Language Studies*, *21*(4), 427–454. https://doi.org/10.1353/sls.2021.0006

Reagan, T. (2021b). Historical Linguistics and the Case for Sign Language Families. *Sign Language Studies*, *21*(4), 427–454. https://doi.org/10.1353/sls.2021.0006

Rozov, N. (2022). Towards the Multistage Ecosocial Theory of Glottogenesis: Modern Evolutionary Concepts, Principles, and Extension of the Nomological Approach. *Open Journal for Studies in Philosophy*, *6*(2). https://centerprode.com/ojsp/ojsp0602/coas.ojsp.0602.02049r.html

Schoonenboom, J. (2023). The Fundamental Difference Between Qualitative and Quantitative Data in Mixed Methods Research. *Forum : Qualitative Social Research (Sozial Forschung)*, *Volume 24, No. 1, Art.* http://www.qualitative-research.net/

Setiawan, L. G. I. P. S. (2020). Hubungan Kekerabatan Bahasa Bali dan Sasak dalam Ekoleksikon Kenyiuran: Analisis Linguistik Historis Komparatif. *Jurnal Inovasi Penelitian*, *1*(1), 27–30. https://doi.org/10.47492/jip.v1i1.44

Sovacool, B. K., Axsen, J., & Sorrell, S. (2018). Promoting novelty, rigor, and style in energy social science: Towards codes of practice for appropriate methods and research design. *Energy Research & Social Science*, *45*, 12–42. https://doi.org/10.1016/j.erss.2018.07.007

Starostin, S. (1999). Comparative-historical linguistics and Lexicostatistics. *Starlingdb.Org*. https://starlingdb.org/Texts/Starostin_Glotto.pdf

Starostin, S. A. (2000). *Comparative-historical linguistics and lexicostatistics," in Time Depth in Historical Linguistics* (C. Renfrew, A. McMahon, and L. Trask, Vol. 1). Cambridge: McDonald Institute for Archaeological Research.

Swadesh, M. (1952). Lexicostatistic dating of prehistoric ethnic contacts. *Proc. Am. Philos. Soc*, 452–463.

Swadesh, M. (1954). Perspectives and Problems of Amerindian Comparative Linguistics. WORD, *10*(2–3), 306–332. https://doi.org/10.1080/00437956.1954.11659530

Swadesh, M. (1955). Towards greater accuracy in lexicostatistic dating. *Int. J. Am. Linguist*, *21*, 121–137. https://doi.org/10.1086/464321

Tantri, A., Saddhono, K., & Mulyono, S. (2024). Kekerabatan Bahasa Jawa dan Bahasa Madura dalam Kajian Linguistik Historis Komparatif. *DIALEKTIKA: Jurnal Pendidikan Bahasa Indonesia*, *3*(2), 75─84. https://journal.peradaban.ac.id/index.php/jdpbsi/article/view/1847/1160

Tao, Y., Wei, Y., Ge, J., Pan, Y., Wang, W., Bi, Q., Sheng, P., Fu, C., Pan, W., Jin, L., Zheng, H.-X., & Zhang, M. (2023). Phylogenetic evidence reveals early Kra-Dai divergence and dispersal in the late Holocene. *Nature Communications*, *14*(1), 6924. https://doi.org/10.1038/s41467-023-42761-x

Troike, R. C. (1969). The Glottochronology of Six Turkic Languages. *International Journal of American Linguistics*, *35*(2), 183–191. https://doi.org/10.1086/465053

Virella, P., & Woulfin, S. (2024). Tell me about your trauma: An empathetic approach-based protocol for interviewing school leaders who have experienced a crisis. *Qualitative Research Journal*. https://doi.org/10.1108/QRJ-09-2022-0121

Williams, S., & McWilliams, K. (2024). "Just to Jog My Memory": An Examination of Forensic Interviewers' Note-taking Behaviors and Perceptions of Notes With Child Witnesses. *Journal of Interpersonal Violence*, 08862605241243346. https://doi.org/10.1177/08862605241243346

Zafar, D. (2023). *ISSN:XXXX-XXXX Analysis the Studies into Comparative Linguistics*. *1*(3).

Zafar, D. (2024). Analysis the Studies into Comparative Linguistics. *European Journal of Artificial Intelligence and Digital Economy*, *1*(2). https://journal.silkroad-science.com/index.php/JAIDE

Zhang, M., & Gong, T. (2016a). How Many Is Enough?—Statistical Principles for Lexicostatistics. *Frontiers in Psychology*, *7*. https://doi.org/10.3389/fpsyg.2016.01916

Zhang, M., & Gong, T. (2016b). How Many Is Enough?—Statistical Principles for Lexicostatistics. *Frontiers in Psychology*, *7*. https://doi.org/10.3389/fpsyg.2016.01916

Zulham, Rahim, Abd. R., & Agus, M. (2022). Kekerabatan Bahasa Makassar dan Bahasa Selayar: Analisis Leksikostatistik dan Glotokronologi. *Gema Wiralodra*, *13*(1), 215–232. https://doi.org/10.31943/gw.v13i1.215