RECONSTRUCTING A PROTO-PAHANG RIVER LEXICON: NOW AND IN THE FUTURE

https://doi.org/10.31577/aassav.2026.35.1.01
CC-BY

Muhammad Norsyafiq ZAIDI
Faculty of Humanities and Social Science,
Prince of Songkla University, Thailand
ORCID: https://orcid.org/0000-0003-4281-534X
muhammad.n@psu.ac.th

Shahidi A. HAMID
Faculty of Social Sciences and Humanities,
The National University of Malaysia
ORCID: https://orcid.org/0000-0003-4520-7585
zedic@ukm.edu.my

Nurul Shuhada SHAFIE
Faculty of Social Sciences and Humanities,
The National University of Malaysia
ORCID: https://orcid.org/0009-0004-5894-196X
nshuhadashafie@gmail.com

Proto Pahang Malay (PPM) faces significant challenges in the preservation and dissemination of its reconstructed lexical data. Much of the existing data remains locked within physical archives or behind paywalls, limiting accessibility for researchers and leading to fragmented, unsystematic scholarly engagement with the dialect. This study addresses these challenges by reconstructing core lexicons across eleven semantic domains, based on data collected from nine dialectal locations along the Pahang River. Applying the comparative reconstruction framework, the study identifies both native Austronesian lexicons and a limited set of loanwords, particularly from Tai-Kadai, Indo-European and other language families. Lexical provenance indicators are proposed to classify these loan elements in a future digital repository, enabling future studies to exclude them from analyses of Austronesian lexical reconstruction. The current phase of the study marks a foundational step toward the creation of a centralised, open-access proto-lexical inventory modelled after resources such as the Austronesian Comparative Dictionary. Beyond academic publication, the study envisions a digital platform with features such as audio playback, cross-referencing tools, and public

access. Collaboration with computational, linguistic, and archival experts is identified as essential for realising this goal. Future research is expected to expand lexical coverage, enhance interactivity, and support language documentation efforts for the wider Malayic linguistic heritage.

Key words: Austronesian; comparative reconstruction; digital repository; preservation; Proto-Malay dialect of the Pahang River; reconstruction; Tai-Kadai; linguistic database

Introduction

Pahang is one of the largest states in Peninsular Malaysia, renowned for its lush greenery, hilly terrain, and extensive jungles. It is home to sundry ethnic groups, including Malay, Orang Asli (Aboriginal) and other communities, each contributing to its diverse linguistic composition. Historically, Pahang played a significant role as a trading hub due to its strategic location along the East Coast. Its extensive network of rivers, most notably the Pahang River, further enhanced its importance, forging the Pahang we know today.

The Pahang River not only serves as a lifeline for the state's agriculture and trade but has also acted as a corridor for the transmission of culture and language over centuries. Historical settlements along the river, as documented by Linehan (1928b, 1928a, 1930, 1936, 2020), indicate that human presence in the area dates back more than 10,000 years. The river's influence has fostered the development of distinct local dialects, reflecting the historical interactions between various ethnic groups and languages in the region.

Pahang Isolect

Although the Pahang River has a long history of settlement, its population is not as great as one might imagine. This is because the population in this area is concentrated in only a few remote locations due to difficult accessibility and the deeply inland nature of the region (Cant, 1972). The difficulty of communication between communities has led to the development of linguistic variations, making Pahang one of the states with a high dialectal diversity (Hussein, 1973).

Hussein (1973) generally notes that there is a pattern of variation in the pronunciation of certain words and even the names of things or matters, changing systematically from the downstream to the upstream areas. This pattern was found in studies by Idris (1989) Hasrah et al. (2014) and Zaidi et al. (2021), where both the central and upstream areas of the Pahang River show significant differences in pronunciation.

Omar (2015) identified the downstream area of the Pahang River as the main hub of possible earliest Malay settlement, evidenced by the presence of a royal

governance area. Pekan, which was established as a hub of Malay civilisation in Pahang, likely expanded gradually to the central and upstream river basin areas. Hypothetically, the Malay community once moved from the downstream areas of the Pahang River towards the upstream areas, such as the Lipis region.

Location and Significance of the Pahang River

The selection of the Malay dialect along the Pahang River as the focus of this study is significant due to its unique linguistic characteristics and historical context. Hussein (1973) observed that phonological variations in the Pahang dialect exhibit systematic changes along the river's course, with greater variations observed in upstream areas. This systematic change provides a rich field for studying the processes of linguistic evolution and dialectal variation within a relatively contained geographical space.

Additionally, the Pahang River region is potentially one of the earliest Malay settlement sites that still maintains its distinct dialect (Omar, 2015). Unlike other areas in Pahang, such as Nenasi and the Tembeling River, which have shifted towards the Terengganu dialect (Collins, 1989; Hasrah et al., 2013; Karim & Ibrahim, 1977), or Rompin, where the Johor dialect predominates, the Pahang River retains its original linguistic variation (Omar, 2015). This persistence of the local dialect, largely unaltered by external influences, makes the Pahang River a crucial site for understanding the continuity and preservation of linguistic identity in Malay communities.

The Need for Centralisation of a Proto-Lexical Inventory

In examining the phonological distribution within this topographical region, it becomes necessary to return to the fundamental aim of the present study: to establish a reference corpus for dialectological research. At present, there exists no integrated or centralised proto-lexical inventory specifically dedicated to the Proto-Malay dialects spoken along the Pahang River.

In contrast, proto-language studies at higher taxonomic levels have made considerable progress. For instance, the lexicons of Proto-Austronesian (PAN) and Proto-Malayo-Polynesian (PMP) have been successfully digitised and made accessible via web-based platforms (Blust & Trussel, 2020). Likewise, research on Proto-Malayic researched by Adelaar (1992) has been partially digitised by Bischoffberger (2013). However, there remains a marked absence of comparable efforts at lower levels of linguistic reconstruction, particularly for dialect clusters such as the Proto-Malay of the Pahang River.

While early contributions such as those by Hasrah et al. (2014) and Zaidi et al. (2021) have laid important groundwork, their outputs are primarily limited to physical or static digital publications. Drawing on the model established by Blust and Trussel (2020) who digitised earlier works including Blust (1970) and Dempwolff (1938), it is evident that similar initiatives are needed for the Proto-Pahang dialect to ensure accessibility, sustainability, and interactivity in linguistic research.

Prior to the development of such a digital platform, the essential first step is to reconstruct a substantial number of proto-lexical items. These reconstructions must be rigorously justified and published in peer-reviewed outlets in order to establish both methodological validity and scholarly credibility. Accordingly, the focus of this study is clearly defined. The specific objectives can be found in the following section.

Objectives of the Study

This study aims to achieve three primary objectives. Firstly, it seeks to reconstruct the proto-lexical forms of the Malay dialects spoken along the Pahang River by applying the comparative reconstruction framework outlined by Crowley and Bowern (2010). Secondly, it aims to identify the origins of the reconstructed lexicons, determining whether the forms are inherited from Austronesian roots or are borrowings from neighbouring language families, such as Tai-Kadai and Mon-Khmer. Thirdly, the study aspires to classify the reconstructed lexicons into eleven semantic domains and to propose a framework for developing an open-access digital lexical inventory of the Pahang River Proto-Malay dialect, facilitating broader scholarly access and long-term linguistic preservation.

Literature Review

Before undertaking the systematic reconstruction of the proto-lexicon of the Pahang River Malay dialect, it is essential to first review the existing body of research to identify both foundational contributions and current limitations. To date, only three significant studies have focused on the reconstruction of proto-lexicons in Malay dialects, with two of them specifically examining dialects spoken along the Pahang River.

The studies conducted by Hasrah et al. (2014) and Zaidi et al. (2021) employed a micro-level approach, concentrating on selected areas along the river. Hasrah et al. (2014), for instance, reconstructed a proto-lexicon based on 200 lexical items collected from upstream Malay-speaking communities. However, the recon-

structed data were not classified into semantic domains, and the analysis focused primarily on phonological correspondences with Proto Malayik (PM). A similar methodology was employed in Zaidi et al. (2021), which also lacked etymological depth and domain-specific classification. While both studies succeeded in identifying numerous proto-forms, they offered limited discussion on the potential origins or diffusion of these forms.

This limitation becomes particularly pertinent when considering the historical and sociolinguistic context of the Pahang River region. Although the region is situated inland and somewhat removed from coastal trading centres, it has long been in contact with indigenous Orang Asli communities and immigrants from south Thailand. These communities include speakers of languages from the Tai-Kadai and Mon-Khmer families, whose influence may be reflected in the lexicon of local Malay dialects. Historical records suggest that the presence of Siamese (Ligor) military forces in the East Coast corridor during the 15th century facilitated cultural and linguistic exchange, particularly through intermarriage between soldiers and local communities (Adil 1972; Anderson 1824; Omar 2015; Sakolnakorn et al. 2020). Consequently, lexical borrowing likely occurred, yet this dimension has been largely overlooked in earlier reconstruction studies.

A more structured and comprehensive approach is evident in the work of Che Rosdi et al. (2023), who reconstructed the proto-lexicon of the Upper Perak dialect. This study employed thematic classification, organising lexical items into nine distinct semantic domains: time, numbers, natural environment, colours, household tools, kinship terms, pronouns, body parts, and emotional or behavioural actions. Although this study did not focus on the Pahang River region, it demonstrated a replicable methodology for domain-based and etymologically informed reconstruction.

In summary, while previous studies have laid the groundwork for lexical reconstruction within Malay dialects, none have developed a comprehensive, domain-organised, and etymologically annotated proto-lexical inventory for the dialects spoken along the Pahang River. Although Hasrah et al. (2014), Zaidi et al. (2021), and Che Rosdi et al. (2023) each address the reconstruction of proto-forms within their respective dialect groups, their findings remain largely fragmented and methodologically incompatible, making it difficult to consolidate these efforts into a unified proto-lexical framework. Furthermore, all three studies were published in physical format only, limiting their accessibility and interoperability for subsequent linguistic research. This lack of digital availability hampers efforts to trace, cross-reference, or expand upon previously reconstructed proto-forms in a systematic manner.

As linguistic research increasingly moves towards open-access, searchable, and interactive formats, the need for digitalisation becomes imperative. Some researchers have already begun familiarising themselves with generative technol-

ogies (Pyysalo, 2017)the output of which equals the Indo-European (IE. While this lies beyond the immediate scope of the present study, the presence of such technologies signals a broader shift in the research landscape. Hence, a digital platform would enable the integration, comparison, and ongoing refinement of proto-lexicons, while also supporting features such as etymological tagging, phonological alignment, and audio playback. In the absence of such infrastructure, the field remains vulnerable to scholarly redundancy and the repeated reinvention of previous work. Therefore, this study seeks to address this gap by proposing a domain-based and etymologically annotated proto-lexicon of the Pahang River Malay dialect as a foundational step toward future digital integration and corpus centralisation.

To ensure clarity of focus in this study, it is crucial to articulate the direction and structure of the work undertaken. The next section therefore outlines the methodological framework employed in this research. Emphasis is given to the identification of etymological sources, whether Austronesian or non-Austronesian in origin, as well as to the marking of lexical borrowings. These aspects are systematically addressed with reference to established comparative linguistic frameworks and earlier proto-language reconstructions. In this way, the study ensures that the reconstructed forms are not only phonologically plausible but also etymologically grounded within a broader historical-linguistic context.

Methods

The lengthy section will be separated into six subsections. These subsections will be used to discuss study approaches, locations, informant selection, data collection, the theoretical frameworks used in this study, and data presentation.

Study Approaches

The study is based on a qualitative approach, equipped two primary methods: literature review and fieldwork. The literature review focuses on understanding the current state of lexical reconstruction research and the future direction of this field. It identifies gaps in previous studies and emphasises the need for macroscopic reconstruction efforts and the documentation of the reconstructed lexicon in high-impact electronic journals. This approach aims to ensure that the research contributes to the development of a proto-lexical inventory system, which will be valuable for research and the study of proto-languages.

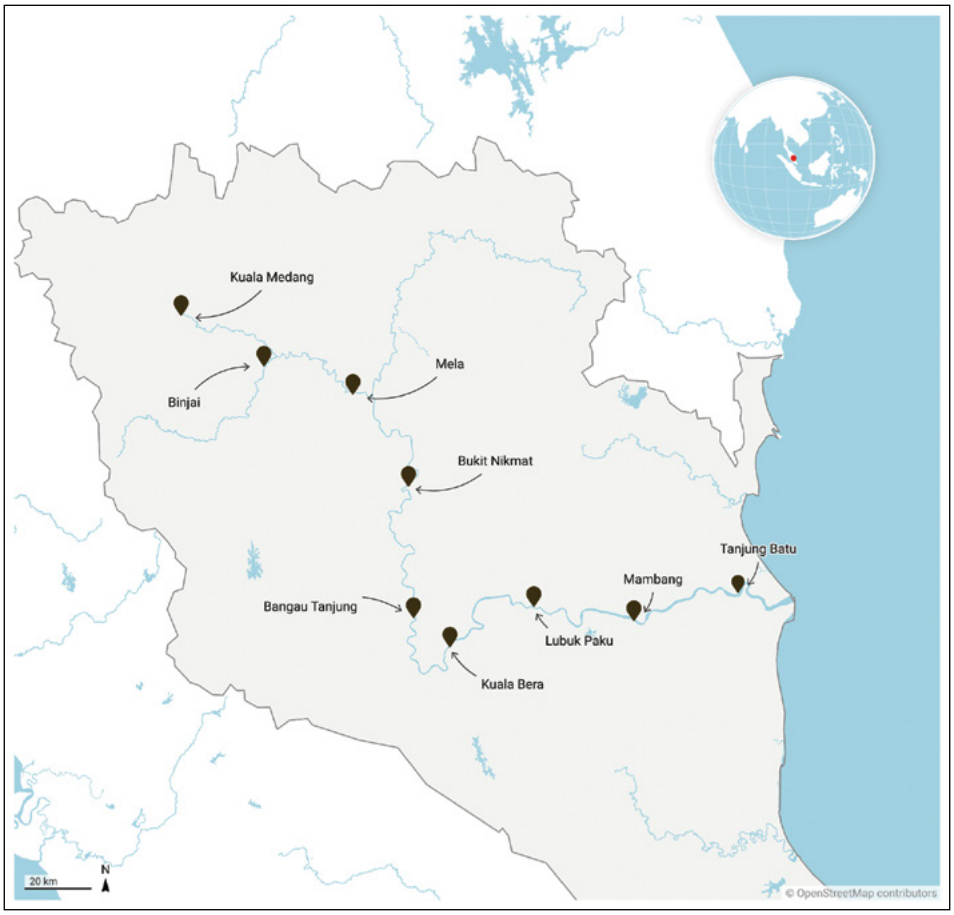

Figure 1. Map of the nine focused locations along the Pahang River

Locations

The study selected the Pahang River as the research site due to the significance of existing data. Previous studies by Hasrah et al. (2014) and Zaidi et al. (2021) have collected a number of proto-lexicons, but their research was limited to only a few areas along the Pahang River. Consequently, this study chose nine locations along the 419-kilometre stretch of the Pahang River, with distances between the study sites ranging from 30 to 50 kilometres.

The selection of study locations was based on two main criteria. Firstly, each location had to have a generational depth of at least three or four layers, ensuring that the chosen sites were not newly established settlements. Secondly, the locations had to be predominantly native Malay communities. This is because the areas around the Pahang River also include settlements of Indigenous Peoples and minority groups from the Tai-Kadai and Mon-Khmer language families who migrated to Pahang.

As a result, the focal areas for the reconstruction of the proto-lexicons include Tanjung Batu, Mambang, Lubuk Paku, Bangau Tanjung, Kuala Bera, Bukit Nikmat, Mela, Binjai, and Kuala Medang. The distribution of these study locations is shown in more detail in Figure 1.

Informant Selection

In the context of dialect studies in the West, Chambers and Peter (2004) suggested several criteria for selecting informants, which they abbreviated as NORM, standing for non-mobile, old, remote, and male. However, Ayatrohaedi (1979) recommended prioritising female informants, especially in Asian societies where men tend to move and travel more frequently than women. As a result, women are more likely to remain within the village setting, thereby playing a crucial role in preserving and transmitting the dialect that has been passed down through the generations (Wu, 2023). Accordingly, this research utilises nine primary informants.

Zaidi et al. (2021) introduced a more justified set of criteria termed NCH: native, complete organ of articulation, and healthy. Male informants tend to leave their homes more frequently and work outside their local areas, leading to increased interaction with outsiders. As a result, the language they use begins to exhibit foreign elements, such as loanwords and pronunciation patterns that increasingly mimic external linguistic features. Consequently, female informants tend to demonstrate greater command of their dialect and are more likely to retain its linguistic features, as their interactions are largely confined to speakers within the same speech community. This set of criteria represents a contemporary approach to informant selection that we should consider adopting.

Data Collection

Several data collection methods were employed, including the use of word lists, as recommended by Omar (2015). She suggested that the words to be tested should be categorised according to their respective domains. Accordingly, in this study,

the data set to be examined is organised into specific domains such as body parts, names of animals, parts of animals, names of plant parts, and other elements closely related to the Malay community of Pahang.

While informants articulate what is considered dialectal pronunciation, researchers simultaneously transliterate these utterances into phonetic symbols. This study utilises the International Phonetic Alphabet (IPA) system, with the symbols being used in the data analysis process to ensure accuracy and consistency in representing the informants' pronunciations.

Theoretical Framework

After collecting the data in the field, the next step was to prepare the data set for reconstruction analysis. As Omar (2019, p. 3) explains, reconstruction is a method used to rebuild a lexical item based on evidence from daughter languages. This study will focus on a phonological perspective, as daughter languages typically evolve from the mother language, adapting over time (Bynon, 1994, p. 43). The study adopted the framework proposed by Crowley and Bowern (2010), which offers a structured approach for reconstructing proto-phonemes. By applying the same methods and principles, this framework can also be used to reconstruct proto-lexicons. To reconstruct both phonemes and proto-lexicons, Crowley and Bowern (2010, pp. 85-92) outlined several fundamental principles:

1. Any reconstruction should involve sound changes that are plausible, unless there is good evidence to the contrary.
2. Any reconstruction should involve as few changes as possible between the proto-language and the daughter languages.
3. Reconstruction should fill gaps in phonological system rather than creating unbalanced systems.
4. A phoneme should not be reconstructed in a proto-language unless it is shown to be necessary from the evidence of the daughter languages.

Example of Reconstruction Framework Application

This section provides a brief explanation of the methods used in reconstructing proto-phonemes and proto-lexicons. Here, we are presented with a set of data related to the *taring* 'canine' gloss. Refer to Table 1 below:

V1 to V9 in Table 1 represent the locations selected for proto-lexical reconstruction. In the first row of the sound correspondence table, the [t] sound consistently appears across all locations. Given its uniform presence, the initial letter of this proto-lexicon is designated as *t. We believe there is no alternation of the

Table 1. Example of phoneme and proto-lexical reconstruction based on the canine gloss (a pointed tooth). The data presented is part of this study.

Sound Correspondence																		
	V1		V2		V3		V4		V5		V6		V7		V8		V9		*
/	t	:	t	:	t	:	t	:	t	:	t	:	t	:	t	:	t	/	t
/	ʁ	:	ʁ	:	ʁ	:	ʁ	:	ʁ	:	ʁ	:	ʁ	:	ʁ	:	ʁ	/	ʁ
/	ɣ	:	ɣ	:	ɣ	:	ɣ	:	ɣ	:	ɣ	:	ɣ	:	ʁ	:	ʁ	/	ɣ
/	ɛ	:	e[a]	:	ɛ	:	ɛ	:	e	:	ɛ	:	e	:	ɛ	:	e	/	e
Rebuilt Proto Lexical																		*tʁɣe	

[t] sound, so Principle 1 is applied. The reconstruction process continues to assess the second row, where the sound [ʁ] appears consistently across all locations, which is similar to [t]. Therefore, there is no issue in reconstructing *ʁ for this proto-lexicon associated with the *taring* 'canine'.

In the third row, a minor enigma arises with the presence of two sounds, [ɣ] and [ʁ]. Based on the majority rule concept outlined by Campbell (2020) and Principle 2 mentioned earlier, the sound [ɣ] is more suitable for reconstruction. Choosing [ʁ] would result in an "unfair" reconstruction, as [ɣ] is present in more than half of the daughter languages. Thus, the study successfully reconstructs three proto-phonemes forming the *taring* 'canine' gloss of the proto-Pahang dialect. However, attention must be given to the fourth row, where three variations of the final vowel in the open syllable appear: [ɛ], [ea], and [e].

Here, we encounter the issue of phonemic and allophonic determination. It is crucial to first determine the proto-phoneme to ensure that the reconstructed proto-lexicon does not contain allophonic elements. Hasrah et al. (2014) and Zaidi et al. (2021) refute the inclusion of [ɛ] as a proto-phoneme, attributing its presence to the expansion of the phonemic vowel /e/. Consequently, in such situations, allophonic elements like [ɛ] are automatically excluded. Therefore, the canine gloss is reconstructed as *tʁɣe in the proto-Malay dialect of Pahang.[1]

[1] The following is a reconstruction method that is simplified for the understanding of readers and reviewers. If readers and reviewers wish to learn more about this reconstruction method, we recommend referring to Crowley and Bowern (2010, pp. 78–93) for further details.

Data Presentation

We have chosen to reconstruct 11 domains, each containing 10-20 lexical items. Accordingly, each domain will have its own table, resulting in at least eleven tables. For data presentation, we will display the reconstructed proto-lexical correspondences along with relevant information, such as the distribution of pronunciations across related languages and etymological indicators. An example of such a table can be found in Table 2 below.

As illustrated in Table 2, we have included the etymological details of the reconstructed terms. Various proto-lexical databases were consulted to trace the connections between the reconstructed lexical items in this study and those found in existing proto-lexical databases. For instance, if a lexical item is cognate with a form from the Proto-Malayo-Polynesian language, it is marked with the symbol PPM, referring to Proto-Pahang Malay. Additionally, if it corresponds with a language from the Mon-Khmer language family, it is indicated by the abbreviation PMK (Proto-Mon-Khmer).

Regarding databases, this research utilises lexical data and historical information mostly from several databases listed below:

1. A Malay-English Dictionary: Romanised (Wilkinson, 1932)
2. Proto-Austronesian lexical list: (Blust & Trussel, 2020)
3. Proto-Malayo-Polynesian (Blust & Trussel, 2020)
4. Proto-Malayic (Adelaar, 1992; Bischoffberger, 2013)
5. Mon-Khmer dictionary (Shorto, 2006)
6. Tai-Kadai database (Starostin, n.d.)
7. Dravidian database (Starostin, n.d.)

Table 2. Example of data presentation related to the reconstructed proto-lexicon, along with its corresponding variations and etymological indicators.

Gloss	Proto lexicon	Correspondences variation	Etymology
canine	*tɐɣe	V1, V3, V4, V6; [tɐɣɛ] V5, V7; [tɐɣe] V2; [tɐɣeª] V8; [tɐʁɛ] V9; [tɐʁe]	PPM
...	…	…	…

Research Findings

This section will be divided into eleven subsections, where the presentation of data is based on the thematic classification of proto-lexicons, covering topics such as body parts, animal names, household tools, characteristics, nature, animal

body parts, plants, sea and river animals, insects, pronouns, family members, and general verbs and nouns. This section will present the outcomes of the reconstruction processes. A detailed discussion of these outcomes will be provided in the relevant section.

Body Parts

There are nineteen lexical items that share the same semantic field: body parts. All reconstructed lexical items are presented in Table 3 below, accompanied by relevant explanatory information:

Table 3. Conspectus of the reconstructed lexical domain for human body parts

Gloss	Proto lexicon	Correspondences variation	Ety.
head	*ǩəpɐlə	TB, LP, BT, KMD; [pɐlə] MBA, KB, BNT, MLA, BNJ; [ǩəpɐlə]	SKT
face	*mukɐ	All dialects; [mukə]	SKT
nose	*ido	TB, LP, BT, KB, BNT, MLA, BNJ, KMD; [idə] MBA; [ido]	PPM
mouth	*mulut	TB, LP, KB, MLA, KMD; [muləʔ] BT; [mulə$^{j\gamma}$] BNT, BNJ; [muləʔ]	PPM
eyes	*mɐtɐ	All dialects; [mɐtə]	PPM
ears	*ťəliŋɐ	All dialects; [ťəliŋɐ]	PPM
chin	*dɐgu	TB, MBA, LP, BT, BNT, MLA, BNJ, KMD; [dɐgu] KB; [dɐguʔ]	PPM
forehead	*dɐi	TB, MBA, LP, BT, KB, BNT, MLA, BNJ; [dɐi] KMD; [dɐj]	PPM
cheek	*pipi	All dialects; [pipi]	PPM
tongue	*lidɐh	TB, MBA, LP, BT, KB, BNT, BNJ; [lidəh] MLA; [lidɐh] KMD; [lidə]	PPM
teeth	*gigi	TB, LP, BT; [gigij] MBA, KB, BNT, MLA, BNJ, KMD; [gigi]	PPM
hand	*tɐŋe	TB, MBA, KMD; [tɐŋɛ] LP, BT; [tɐŋe] BNT, BNJ; [tɐŋɐ] KB, MLA; [tɐŋɐˀ]	PPM
elbow	*siku	TB, MBA, LP, BT, KMD; [sikuw] KB, BNT, MLA, BNJ; [siku]	PPM
arms	*ləŋɐ	TB, MBA, LP; [ləŋɛ] BT; [ləŋe] BNT, BNJ, KMD; [ləŋɐ] KB, MLA; [ləŋɐˀ]	PPM
stomach	*pəɣut	TB; [pɣɔʔ] LP; [pɣuʔ] BT; [pəɣu$^{j\gamma}$] MBA, MLA; [pəɣɔʔ] BNT, BNJ, KMD; [pəɣəʔ] KB ø	PPM

Gloss	Proto lexicon	Correspondences variation	Ety.
buttock	*puŋgo	TB; [buntɔʔ] MBA; [puŋgoᵛ] LP; [buntojʔ] BT; [dʒubɔ] KB, BNT, MLA, BNJ, KMD; [puŋgɔ]	PPM
calf	*bətis	TB, MBA, LP, KB, BNT, KMD; [bəteh] BT; [bətɛʲ] MLA, BNJ; [bətɛh]	PPM
feet	*kɐki	TB, MBA, LP, KMD; [kɐkij] BT, KB, BNT, MLA, BNJ; [kɐki]	PPM
finger	*dʒɐɣi	TB, MBA, LP, BT, KB, BNT, MLA; [dʒɐɣi] BNJ, KMD; [dʒɐʁi]	PPM

Based on the preceding table, one instance of borrowing from a foreign language family can be identified. The abbreviation SKT in the table refers to Sanskrit. This borrowing is the result of prolonged linguistic contact that has occurred over thousands of years. For comparative purposes, the etymon for the gloss *head* is reconstructed as *hulu(ʔ) at a higher proto-level (Adelaar, 1992).

Animal Names

There are twenty-one lexical items that share the same semantic field, namely animal names, which currently show a noticeable increase in the number of lexical borrowings from foreign language families. Refer to Table 4 below for further details:

Table 4. Conspectus of the reconstructed lexical domain for animal names

Gloss	Proto lexicon	Correspondences variation	Ety.
chicken	*ɐjɐ	All dialects; [ɐjɐ]	PPM
duck	*itik	TB, MBA, LP, BT, KB, BNT, MLA, BNJ; [itɛʔ] KMD; [itɛʔ]	PPM
cattle	*ləmbu	TB, BNT, BNJ; [ləmbu] MBA; [ləːm̄ũ] LP; [ləmũ] BT; [ləbuw] MLA; [ləmːu] KMD; [ləmbuw] KB; ø	SKT
dog	*ɐndʒe	TB, LP; [andʒeᵉ] MBA, BT, KB, BNT, MLA, BNJ, KMD; [andʒɛ] LP; [andʒeᵃ]	PPM
wild chicken	*ɐjɐ.ute	TB; [ɐjɐ.utae] MBA, BT, BNT, BNJ, KMD; [ɐjɐ.utɛ] LP, MLA; [ɐjɐ.utɐ] KB; [ɐjɐ.utɐm]	PPM
cat	*kutʃe	TB, LP, KB, BNT, MLA, BNJ, KMD; [kutʃɛ] MBA, BT; [kutʃ̃ɛᵃ]	PPM

Gloss	Proto lexicon	Correspondences variation	Ety.
buffalo	*kəɣəbʁ	TB, MBA, LP, BT; [kəɣbʁ] KB, BNT, MLA; [kəɣbɔ] BNJ, KMD; [kəbɔ] LP; [kəːbʁ]	PPM
snake	*ulo	All dialects; [ulɔ]	PPM
cobra	*ulo.tədo	TB, BNT, MLA, BNJ, KMD; [ulɔ.tədɔ] MBA; [ulɔ. tədɔa] BT; [ulɔ tədɔŋ] LP, KB ø	PPM
python	*ulo.sʁwə	All dialects except KB; [ulɔ.sawə] KB ø	PPM
monkey	*kəɣə	TB, LP, BT, KB, BNT, MLA; [kəɣə] MBA; [moɲeʔ] BNJ; [moɲiʔ] KMD ø	PPM
monkey	*bəɣuk	TB, MBA, LP, BT, KB, BNT, MLA, KMD; [bəɣɔʔ] BNJ; [bəʁɔʔ]	PPM
agile gibbon	*uŋkʁ	TB, LP, BT, MLA, BNJ, KMD; [uŋkə] BNT; [tənəkɔh] MBA, KB ø	PPM
elephant	*gʁdʒʁh	TB, MBA, LP, BT, KB, BNT, MLA, BNJ; [gʁdʒɔh] KMD; [gʁdʒɔ]	SKT
tiger	*ɣimʁ	TB, MBA, LP, BNJ; [ɣimʁ] BT, BNT, MLA, KMD; [ɣimɔ] KB ø	PPM
tortoise	*kəkuɣʁ	TB, MBA, LP, BT, BNT; [kːkuɣə] KB, MLA; [kuɣə kuɣə] BNJ; [kuʁə kuʁə] KMD; [kəkuʁə]	PPM
lesser mouse-deer	*kʁntʃe	TB, LP; [plandɔʔ] BT, KB, BNT, MLA, BNJ, KMD; [kantʃɛ] MBA ø	PPM
deer	*ɣusʁ	TB, MBA, LP, BT, BNT, MLA, KMD; [ɣusə] BNJ; [ʁusə] KB ø	SKT[2]
goat	*kʁmbe	TB, BT; [kʁmbɛʁ] LP, KB, BNT, MLA, BNJ, KMD; [kʁmbɛ]	PPM
horse	*kudʁ	All dialects; [kudə]	PPM
belatuk	*bəlʁtuk	TB; [ŋɔ] MBA, BNJ; [blʁtʃɔʔ] LP, BT, KMD; [blʁtɔʔ]	PPM

Based on the preceding table, twenty-one reconstructed lexical items have been successfully established at the macro level. However, during data collection, a consistent linguistic attitude towards borrowing from foreign sources, particularly from the Sanskrit language family, was observed. For instance, the lexical items *cattle* and *deer* provide some insight into the presence of these animals following the migration of Sanskrit-speaking communities and the spread of their religion into the Pahang region.

[2] The proto-Malayo-Polynesian lexical inventory (Blust & Trussel, 2020) included proto--lexical borrowings from Sanskrit, although there is no Borrowing/Loan indicator.

Home Appliances

A total of twenty reconstructed lexical items exhibit a notable increase in the number of borrowings from foreign language families. This pattern of uniformity is presented in Table 5 below:

Table 5. Conspectus of the reconstructed lexical domain for home appliances

Gloss	Proto lexicon	Correspondences variation	Ety.
door	*pintu	TB, BNT, MLA, BNJ; [pintu] MBA, LP, KMD; [pitu] BT; [pintuw] KB ø	PMK
floor	*lɐnte	TB, BT, BNT, BNJ; [lɐntɛ] MBA, LP, MLA; [lɐntɐ] KMD; [lɐtɐ] KB ø	PPM
table	* medʒɐ	TB, MBA, LP, BT, KB, BNT, MLA, BNJ, KMD; [medʒə]	LTN[3]
chairs	*kəɣusi	TB, BT, BNT, KMD; [kɣusi] MBA; [kːusi] LP; [kusi] KB, MLA; [kəɣusi] BNJ; [kəʁusi]	SMR[4]
floor mat	*ɐlɐh.kɐki	TB, LP, KMD; [pəŋəsɛʔ] MBA, BT, KB, BNT, MLA; [ɐlɐh kɐki] BNJ; [lɐpeʔ kɐki]	PPM
flower vases	*pɐsu	TB, MBA, LP, BT, KB, BNT, MLA, BNJ, KMD; [pɐsu]	LTN
urn	*təmpɐje	TB, MBA, LP, BT, BNT, MLA, BNJ; [təmpɐjɛ] KMD; [təpɐjɛ] KB ø	PPM
coconut shell[5]	*tʃebuk	TB, BT, BNT, BNJ, KMD; [tʃebɔʔ] MBA; [tʃɛbɔʔ] LP, KB, MLA; [gɐjə]	PPM
water dipper	*gɐjo	TB, LP, BT, KB, MLA; [gɐjə] MBA, BNT; [tʃɛbɔʔ] BNJ, KMD; [tʃebɔʔ]	PPM
mirror	*tʃəɣəme	TB, MBA, BT, BNT; [tʃəɣəmɛ] LP, KMD; [tʃəɣmɛ] KB; [tʃəmɛ] MLA; [tʃəɣmeŋ] BNJ; [tʃəʁmin]	SKT
hanger	*pəɲɐŋkut	BT; [pəɲɐŋkɔʔ] BNT; [sɐŋkɔʔ] KMD; [pəɲidɛ]	PPM
tiles	*dʒube	LP, KMD; [dʒube] BNT; [dʒubɛ]	U[6]
mat	*tiko	TB, MBA, LP, BT, KB, BNT, MLA, BNJ, KMD; [tikɔ]	PPM
rugs	*kɐpet	TB, LP, BT, KMD; [kɐpɛʔ]	ARM[7]

[3] LTN = Latin.
[4] SMR = Samaria.
[5] The coconut shell is used for scooping water. It usually comes with an urn, which is placed in front of the house. Before entering the house, the feet will be washed with an urn full of water. Aside from that, coconut shells are used for bathing.
[6] Unknown origins.
[7] ARM = Armenia.

Gloss	Proto lexicon	Correspondences variation	Ety.
curtain	*lɐŋse	TB, KMD; [lɐsːɛ] MBA, LP, BT, KB, BNT, MLA, BNJ; [lɐŋsɛ]	PPM
picture	*gɐmbo	TB, MBA, LP, BT, KB, BNT, MLA, BNJ, KMD; [gɐmbɔ]	PPM
flower	*buŋɐ	TB, MBA, LP, BT, KB, BNT, MLA, BNJ, KMD; [buŋə]	PPM
bed	*kɐte	TB; [kɐtːɛ] MBA; [kːɐtɛ] LP, BT, KB, BNT, MLA, BNJ, KMD; [kɐtɛ]	PPM
mattress	*tilɐ	TB, MBA, LP, BNT, MLA, BNJ, KMD; [tilɐ] BT; [tilɐŋ] KB; [tilɐˀ]	PMKTKD[8]
sheets	*tʃɐdɐ	TB, BT, BNT; [tʃɐdɔ] LP, MLA, BNJ, KMD; [tʃɐdɐ] MBA; [gəbɔ] KB; [tʃɐdɛ]	IIR[9]

Table 5 displays a list of reconstructed lexicons, several of which are borrowed from various foreign language families, including Tai-Kadai, Mon-Khmer, Latin, Sumerian, and Armenian. There are also words whose origins remain unidentified. The reconstructed household items reflect the cultural transition resulting from the assimilation of foreign influences into the normalised linguistic and cultural practices of the Pahang community.

Characteristics

At least twenty reconstructed lexical items belong to the semantic field of qualities or attributes. Unlike the previous tables, the reconstructed vocabulary in this set demonstrates a relatively lower degree of lexical borrowing. Refer to Table 6 below for details:

Table 6. Conspectus of the reconstructed lexical domain for characteristics

Gloss	Proto Lexicon	Correspondences variation	Ety.
angry	*mɐɣɐh	TB, MB, LP, BT, KB, BNT; [mɐɣɔh] MLA; [mɐɣɔh] BNJ; [mɐʁɔh] KMD; [mɐɣɔ]	PPM
disquiet	*hibɐ	TB, MB; [hibə]	PPM
sad	*sədih	TB, MB, LP, BT, KB, BNT, MLA, BNJ; [sədeh] KMD; [sədɐ]	PPM

[8] TKD = Tai-Kadai.
[9] IIR = Indo-Iranian.

Gloss	Proto Lexicon	Correspondences variation	Ety.
scare	*tɐkut	TB, LP, BNT, MLA, BNJ, KMD; [tɐkɔʔ] BT; [tɐkuʲʔ] KB; [tɐkuʔ]	PPM
worry	*ɣisɐ	TB, BT, KB, BNT, MLA, KMD; [ɣisɔ] MB, LP; [ɣisɐ] BNJ; [ʁisɐ]	PPM
restless	*gəlisɐh	TB, KB; [glisɔh] BNT, BNJ; [gəlisɔh] BT; [gəlisɔh]	PPM
happy	*gəmbiɣɐ	TB, BNT; [gəmbiɣə] BNJ; [gəmbiʁə]	SKT
like	*sukɐ	TB, MB, LP, BT, KB, BNT, MLA, BNJ, KMD; [sukə]	SKT
hot	*pɐnɐs	TB; [pɐneh] MB, LP, BT, MLA, BNJ; [pɐnɐh] KB, BNT; [pɐnɐx] KMD; [pɐnɛ]	PPM
cold	*sədʒuk	TB, BNT, BT, MLA, BNJ, KMD; [sədʒuʔ] MB, KB; [kətːɔ]	PPM
warm	*suɐ	TB, MB, LP, BT, KB, BNT, MLA, BNJ; [suɐ] KMD; [dədːɔ]	PPM
frozen	*bəku	TB; [bəkɔ] MB, LP, BT, KB, BNT, MLA, BNJ; [bəku] KMD; [kəʁɛh]	PPM
exist	*ɐdɐ	TB, BNT, MLA, BNJ, KMD; [ɐdə] MB, LP; [ɐdə] BT, KB; [ɐndə]	PPM
cease	*tɐk.dɐk	TB, MB, MLA, BNJ, KMD; [tɐʔdɔʔ] LP, BNT; [tɐʔdə] KB; [tɐʔdəʔ]	PPM
weak	*ləmɐh	TB, LP, BT, KB, BNT, MLA, BNJ; [ləmɔh] KMD; [lemɔ]	PPM
strong	*ɐnde	TB, MB; [kuɐˀʔ]LP, KB, BNJ; [kueʔ] BT; [kuɐʲʔ] BNT, MLA; [kuɐʔ]	PPM
healthy	*sihɐt	TB, MB, LP, BT, KB, BNT, MLA, BNJ; [sihɐʔ] KMD; [sihɐʔ]	AA[10]
sick	*sɐkit	TB, MB, LP, BT, KB, BNT, MLA, BNJ, KMD; [sɐkɐʔ]	PPM
noise	*bise	TB, MB; [bisɛɐ] LP, BT, BNT, MLA, BNJ; [bisɛ] KB; [bisim]	PPM
silent	*səɲɐp	TB, MB, LP, BT, KB, BNT, MLA, BNJ, KMD; [səɲɐʔ]	PPM

Table 6 contains three borrowed lexical items, consisting of entries from Sanskrit and Afro-Asiatic origins. The Sanskrit borrowings are reflected in the glosses *happy* and *like*, while the Afro-Asiatic borrowing appears in the gloss *healthy*. This semantic field shows a notably lower degree of borrowing, possibly due to the natural tendency of communities to express emotions and experiences through their own linguistic resources.

[10] AA = Afro-Asiatic.

Environmental

The lexical items related to the natural environment were collected during fieldwork; however, only a limited number were obtained. A total of eleven words were successfully reconstructed. Refer to Table 7 below:

Table 7. Conspectus of the reconstructed lexical domain for environmental items

Gloss	Proto lexicon	Correspondences variation	Ety.
tree	*pokok	TB, MB, LP, BT, KB, BNT, MLA, BNJ; [pokɔʔ]	PPM
sky	*lɐŋit	TB, MB; [lɐŋɛʔ] LP, BT, KB, BNT; [lɐŋeʔ] MLA, BNJ; [lɐŋiʔ]	PPM
cave	*guwɐ	TB, MB, LP; [guwə]	SKT
river	*suŋe	TB, LP, BT, KB, BNT, MLA, BNJ; [suŋɛ] MB; [suŋɐ]	PPM
beach	*pɐnte	TB, BT, BNT, MLA, BNJ; [pɐntɛ] MB, LP, KB; [pɐntɐ]	PPM
trail	*dəne	TB, BNT, MLA, BNJ; [dənɛ] MB, KB; [dənɐ] LP, BT; [loɣɔ]	PPM
green pigeon[11]	*pune	TB, BT, BNT, MLA, BNJ; [punɛ] MB, LP; [punɐ]	PPM
estuary	*kuwɐlɐ	TB, MB, LP, BT, KB, BNT, MLA; [kwɐlə] BNJ; [kɔlə]	PPM
cloud	*ɐwɐ	TB, MB, BT, KB, BNT, MLA; [ɐwɐˀ] LP; [ɐwɛ] BNJ; [ɐwɐ]	PPM
road	*dʒɐlɐ	TB, MB, BT, KB, BNT, MLA; [dʒɐlɐˀ] LP; [dʒɐlɛ] BNJ; [dʒɐlɐ]	PPM
shore	*lɐut	TB, MB, BT, KB, MLA; [lɐɔʔ] LP; [lɐoʔ] BNT, BNJ; [lɐəʔ]	PPM

Table 7 presents the reconstructed lexical items, most of which are of Austronesian origin, with one identified as a borrowing from Sanskrit. Interestingly, the word *gua* (cave) was borrowed from Sanskrit, which may be hypothesised as a reflection of the religious background of ancient communities, many of whom practised Hinduism. Hindu religious practices often involved the use of large caves, such as those found in Gua Charas, Pahang and Batu Caves in Selangor, serving as an example of how religious activities could influence the incorporation of a foreign lexicon into the local linguistic repertoire.

[11] A bird species: green pigeon.

Animal Parts

At least ten reconstructed lexical items are identified in this section, encompassing both terrestrial and aquatic animals. Refer to Table 8 for details of this phenomenon:

Table 8. Conspectus of the reconstructed lexical domain for animal parts

Gloss	Proto lexicon	Correspondences variation	Ety.
comb[12]	*bɐlo	TB, MBA, LP, BT, KB, BNT, MLA, BNJ, KMD; [bɐlɔ]	PPM
tail	*eko	TB, LP, BT, KMD; [ekɔ] MBA, KB, BNT, MLA, BNJ; [ɛkɔ]	PPM
fish scales	*sisik	TB, BT; [sisɛʔ] MBA, LP, KB, BNT, MLA, BNJ, KMD; [siseʔ]	PPM
fin	*siɣip	TB, BNJ; [siʁɛʔ] MBA, LP, BT, KB, BNT, MLA, KMD; [siɣeʔ]	PPM
gill	*insɐ	TB, LP, BT, KB, BNT, MLA, BNJ, KMD; [insɐ] MBA; [ensɐŋ]	PPM
fur	*bulu	TB, LP, BT; [buluw] MBA, KB, BNT, MLA, BNJ, KMD; [bulu]	PPM
canine	*tɐɣe	TB, LP, BT, BNT; [tɐɣɛ] KB, MLA; [tɐɣe] MBA; [tɐɣeª] BNJ, KMD; [tɐʁe]	PPM
whisker	*mise	TB, MBA, BNJ, KMD; [misɛ] LP, MLA; [misɐ] KB, BNT; [misɐª]	PPM
spur	*tɐdʒi	TB, MBA, BT, KB, BNT; [tɐdʒij] LP, MLA, KMD; [tɐdʒi] LP; [tadʒi/susɔʔ]	PPM
pouch	*kɐnto	TB, LP, KMD; [kɐntɔ] BT; [pokeʔ]	U

Table 8 presents the reconstructed proto-lexicon, comprising nine Austronesian-origin lexical items that have been retained within the Proto-Pahang Malay inventory, alongside one foreign lexicon whose origin remains unidentified. The predominance of Austronesian lexical items within the Proto-Pahang Malay inventory reflects the community's close interaction with the animal world, whether through hunting activities or animal domestication.

[12] The comb referred to the fleshy part on top of the head of a chicken, usually coloured red and only found in males.

Parts of a Tree

The following section presents the reconstruction of lexical items within the domain of plant parts. In this case, fourteen lexical items were successfully reconstructed, as listed in Table 9 below:

Table 9. Conspectus of the reconstructed lexical domain for part of tree

Gloss	Proto lexicon	Correspondences variation	Ety.
leaf	*dɐo	TB, MBA, LP, KB, MLA, KMD; [dɐɔ] BT; [dɐɔŋ] BNT, BNJ; [dɐᵊ]	PPM
root	*ɐko	TB, MBA, LP, BT, KB, BNT, MLA, BNJ, KMD; [ɐkɔ]	PPM
branch	*dɐhe	TB, LP, KB, BNJ, KMD; [dɐhɛ] MBA, BT, BNT, MLA; [dɐhɐ] MLA; [dɐhɐᵊ]	PPM
shoots	*putʃuk	TB, MBA, LP, BT, BNT, MLA, KMD; [putʃɔʔ] KB, BNJ; [putʃuʔ]	PPM
flower	*buŋɐ	TB, MBA, LP, BT, KB, BNT, MLA, BNJ, KMD; [buŋə]	PPM
thorn	*duɣi	TB; [duɣi] BNJ, KMD; [duʁi]	PPM
seed	*bidʒik	TB; [bidʒi] MBA, LP, BT, KB, BNT, MLA, BNJ, KMD; [bidʒiʔ]	PPM
rubber	*gətɐh	TB, MBA, LP, BT, BNT, MLA, BNJ, KMD; [gətəh] KB; [gətɔx]	PPM
weed	*lɐlɐ	TB, MBA, LP, BT, KB, BNT, MLA, BNJ, KMD; [lɐlɐ]	PPM
lotus	*təɣɐte	TB, BT, BNT, KMD; [təɣɐtɛ] LP, MLA; [təɣɐtɐ] BNJ; [təʁɐtɛ]	PPM
logs	*bɐlɐk	TB, LP, BT, KB, BNT, MLA, BNJ, KMD; [bɐlɔʔ] MBA; [bɐlɐʔ]	PPM
wood	*kɐju	TB, MBA, LP, BT, KB, BNT, MLA, BNJ, KMD; [kɐju]	PPM
ginger	*ɐlijɐ	TB, MBA, LP, BT, KB, BNT, MLA, BNJ; [hɐliə] KMD; [ɣoliə]	PPM
lemon-grass	*səɣe	TB, KB, BNT; [səɣɛ] MBA, LP; [səɣɐ] BNJ, KMD; [səʁɛ]	PPM

This study successfully reconstructed fourteen Austronesian-origin lexical items that were seeded by lower proto variants such as Proto-Pahang Malay. Table 9 demonstrates that there is no evidence of borrowing from foreign languages. This finding provides insight into the community's profound familiarity and intimate relationship with the natural environment.

Sea and River Creatures

This section discusses the reconstruction of lexical items related to marine and riverine life. At least twenty proto-forms were successfully reconstructed and are summarised in Table 10 below:

Table 10. Conspectus of the reconstructed lexical domain for sea and river creatures

Gloss	Proto lexicon	Correspondences variation	Ety.
crab	*kətɐ	All dialects [kətɐ]	PMK
horse crab	*bəlɐŋkɐs	TB; [bəlɐŋkeh] MBA, BT, KB; [blɐŋkeh] LP; [bəlɐŋkɐʲʰ] BNT; [blɐŋkɐx] MLA; [blɐŋkex] BNJ; [blɐŋkɐx] KMD; [blɐŋkɛh]	PPM
tilapia	*tɐlɐpiɐ	TB, MBA, LP, BT, BNT, MLA, BNJ, KMD; [tɐlɐpiɐ] KB; [tɐlɐpiɔ]	BNT[13]
jellyfish	*obo.obo	TB, LP, BT; [obo.obo] BNT, MLA, BNJ; [obɔ.obɔ] KMD; [əbɔ.əbɔ]	PPM
catfish	*kəli	TB, MBA, BT; [kəlij] LP, KB, BNT, MLA; [kəli] MLA, BNJ, KMD; [smilɐ]	U
snake-head fish	*we	TB; [wɛ] MBA; [ɣuɐn] LP; [bəɣətʃɐʔ] BT; [ɣuwe] KB, BNT; [ɣuɛ] MLA, BNJ; [tomɛ] KMD; [ɣuwɛ]	PPM
barb fish	*lɐmpɐ	TB, MBA, LP, BNT, MLA, BNJ, KMD; [lɐmpɐ] BT; [lɐmpɐŋ] KB; [lɐmpɛ]	PPM
scads fish	*təmule	MBA; [təmoleh] LP, MLA; [təməlijɛ] BT; [təmolɛh] BNT; [tɐmɐlijɛ] BNJ; [təmolɛ] KMD; [təmule]	U
bagrid fish	*bɐwo	TB, MBA, BT, KB, BNT, MLA, BNJ, KMD; [bɐɔ] LP; [bɐɔŋ]	PPM
silver catfish	*pɐte	TB, LP, KB, BNT, MLA, BNJ, KMD; [pɐtɛ] MBA, BT; [pɐtɛᵉ]	PPM
eel	*bəlut	TB, MBA, LP; [bəluʔ] BT, KB; [bələʔ] BNT, BNJ, KMD; [bələʔ] MLA; [bəloʔ]	PPM
shark	*ju	All dialects [ju]	PPM
clam	*kəpɐh	All dialects [kəpɔh]	PPM
fish	*ike	TB, MBA; [ikɐᵊ] LP; [ikɐʲ] BT, KB, BNT; [ikɛɐ] MLA; [ikɛ] BNJ; [ikɐ] KMD; [ikɛ]	PPM

[13] BNT = Bantu.

Gloss	Proto lexicon	Correspondences variation	Ety.
shellfish	*kəɣʁ	TB, MBA, LP, BT, BNT, MLA; [kəɣʁ] BNJ, KMD; [kəʁʁ]	PPM
squid	*soto	TB, KMD; [soto] MBA, LP, BT, KB, BNT, MLA, BNJ; [sotɔ]	PPM
softshell turtle	*l:abi	TB, LP, BT, KMD; [l:ʁbi] MBA; [lʁlʁbi] KB, BNT, MLA, BNJ; [lʁbi.lʁbi]	PPM
crocodile	*bojʁ	TB, LP, BT, KMD; [bojə] MBA, KB; [bwʁjə] BNT, MLA; [bwɔjə] BNJ; [bɔjə]	PPM
prawns	*udʁ	All dialects [udʁ]	PPM

Table 10 presents the reconstructed lexical items, most of which exhibit genuine Austronesian features, except for a few whose origins remain uncertain due to the absence of corresponding forms in proto-language corpora. Examples include the glosses *catfish* and *scads fish*. The term *catfish* refers to a non-native species that does not originate from the freshwater ecosystems of Pahang and its surrounding areas, and its etymology remains unknown similarly for *scads fish*. In the case of *tilapia*, while its biological origin is known, the lexical form itself was borrowed from the general genus term that refers to *fish* in Bantu languages.

In addition, the term *crab*, believed to be borrowed from the Mon-Khmer language family (Uthai, 1993), was likely assimilated through linguistic contact with indigenous communities residing near the Malay-speaking population of Pahang. However, the precise mechanism through which the language infiltrated remains uncertain. Further research is required to determine whether the Mon-Khmer speakers had direct contact with the Pahang Malay community, or whether the Pahang Malay lexicon absorbed the term indirectly through intermediary linguistic communities, a question that remains open for scholarly discussion.

Insects

Table 11 below presents the reconstructed lexical items primarily of Austronesian origin, except for the words specified otherwise. To further clarify this, refer to the discussion below.

Table 11 illustrates the phonological uniformity observed among most Proto-Pahang Malay correspondences. A total of ten lexical items were successfully reconstructed at the proto level, among which only one, *crickets*, reconstructed as *tʃəŋkəɣik, is of non-Austronesian origin. This term was borrowed from the Mon-Khmer language family through a process of linguistic contact. However, the direction of borrowing remains uncertain, as it is unclear whether the influ-

ence occurred directly or indirectly, given that the same lexical item is also attested in several other Malayic languages.

Table 11. Conspectus of the reconstructed lexical domain for insects

Gloss	Proto lexicon	Correspondences variation	Ety.
ant	*səmut	TB, MBA, BT; [səməʔ] LP, KB, MLA; səmuʔ BNT, BNJ, KMD;[səməʔ]	PPM
grass-shopper	*bːelɐlɐ	TB, MBA, LP, KB, MLA BNT, BNJ, KMD; [bːelɐlɐ] BT; [bːelɐlɐŋ]	PPM
fly	*lɐlɐt	TB, MBA, LP, KB, MLA BNT, BNJ, BT; [lɐlɐʔ] KMD; [lɐləʔ]	PPM
green fly	*lɐŋa	TB, LP, BT; [lɐŋɐʔ] MBA, KB; [lɐŋɐ] BNT, MLA, BNJ, KMD; [lɐŋɔ]	PPM
crickets	*tʃəŋkəɣik	TB, LP, BT, KB, MLA; [tʃəŋkəɣiʔ] MBA; [tʃəkəɣeʔ] BNT; [kəɣideʔ] BNJ; [tʃəŋkəʁeʔ] KMD; [tʃəŋkəʁiʔ]	PMK
cock-roaches	*lipɐs	TB; [lipeh] LP, BT; [lipɐh] MBA, BNT, MLA; [lipɐx] KB, KMD; [lipex] BNJ; [lipɛh]	PPM
bee	*ləbɐh	TB; [ləbɔ] MBA, LP, BT, KB, MLA, BNJ, KMD; [ləbɔh] BNT; [ləbɔx]	PPM
butterflies	*ɣːəɣɐmə	TB; [ɣːɐmə] LP; [ɣɐmə] BT, BNT; [ɣɐmə.ɣɐmə] BNJ; [ʁəʁɐmə] KMD; [ʁːɐmə] MBA, KB, MLA ø	PPM
millipede	*ulɐt. goŋgok	TB, BT, KB; [goŋgɔʔ] BNT, KMD; [ulɐʔ goŋgɔʔ] MBA; [goŋɔʔ] LP, BNJ; [gəluntɔ] MLA; ø	PPM[14]
mosquito	*ɲɐmuk	TB, MBA, LP, BT, KB, BNT, BNJ, KMD; [ɲɐməʔ] MLA; [ɲɐmuʔ]	PPM

Pronouns

Pronouns in the Malay language exhibit considerable complexity, largely due to the influence of non-Austronesian elements on pre-modern Malay, as well as the adaptation and normalisation of usage that incorporated syntactic and lexical features from Arabic culture. In this study, however, only ten pronouns were elicited, which are believed to have largely eliminated potential Arab influences. Refer to Table 12 for further observations:

[14] Hamilton (1925, p. 32) provides the etymology for the word Gonggok (millipede) in the Johor dialect and Sepah Bulan in the Kedah dialect.

Table 12. Conspectus of the reconstructed lexical domain for pronouns

Gloss	Proto lexicon	Correspondences variation	Ety.
I	*koi/keh	TB; [kɐwe] LP, KB; [koi] BT; [kɔi] BNT, BNJ, KMD; [kɛh] MLA; [kɐh] MBA ø	PMK
I	*keh	TB; [kɐwe] MBA, KB; [ɐku(w)] BT; [kɔi] BNT, BNJ, KMD; [kɛh]	PMK
you	*ĕok	TB, KB, BNJ, KMD; [ĕɔʔ] MBA, LP, BNT, MLA; [ɐwɔʔ] BT; [ɔ̆ʔ]	PPM
you15	*ĕok	TB, KB, KMD; [ɐɔʔ] LP, BNT; [ɐwɔʔ] BT; [ɔʔ] MBA; [kɐmu] BNJ; [demə] MLA ø	PPM
you16	*ĕok	TB, KMD; [ɐɔʔ] BT; [ɔʔ] LP; [ɐwɔʔ] MBA, KB, BNT, MLA, BNJ ø	PPM
you17	*ŋ:kɐ	TB, LP, BT; [ŋ:kɐ] KMD; [ɐɔʔ] MBA, KB, BNT, MLA, BNJ; ø	PPM
we	*komə	TB; [kɐmij] MBA, LP; [kɐmi] BT; [ɔʔkomə] KB, BNJ, KMD; [komə] BNT, MLA; [kɔmə]	RAO[18]
us	*kitɐ	TB, MBA, LP, BT, KB, MLA, BNJ; [kitə] BNT; [demə] KMD; [kɔmə]	PPM
they	*demə	TB, BT; [ɐɔʔ] MBA, LP, BNT; [ɐwɔʔ] KB; [mekə] MLA, BNJ, KMD; [demə]	RAO
he/she	*dijɐ	TB, KB, MLA, BNJ; [dijə] BT; [demə] MBA, LP, BNT; [ɐwɔʔ] KMD; [ɐɔʔ]	PPM

Although the study strategically avoided Arabic-derived lexicon during data elicitation, some foreign words of non-Austronesian origin were nevertheless recorded. Their clear articulation highlights the extensive cross-cultural interaction that took place. The source language families include Mon-Khmer, as well as Malayic varieties brought originally from present-day Indonesia by people who later settled in regions of Peninsular Malaysia. While previous research indicates that pronouns are generally resistant to borrowing (Alikhon, 2022), in the case of Proto-Pahang Malay, the level of borrowing, though relatively low, is no exception.

[15] The word *you* is plural in this context. Some Malay dialects use repetition; [awok.awok] refers to a group of individuals who are listening to the speaker. In Pahang dialect, they only use one term to address a large number of people.

[16] The glossary used during data collection is in Malay. The original word for this is *anda*, which is rarely used except in writing.

[17] Another *you* we have here has a strong connotation. This is typically used when a speaker is irritated or angry with someone.

[18] RAO = Rawa.

This suggests a significant degree of interaction between the Pahang Malay community and neighbouring linguistic groups.

Family Names

Finally, this study successfully reconstructed proto-lexical items within the semantic domain of kinship terms. For this section, eighteen lexical items of Austronesian origin were extracted. Refer to Table 13 below:

Table 13. Conspectus of the reconstructed lexical domain for family names

Gloss	Proto lexicon	Correspondences variation	Ety.
mother	*mok	TB, MBA, LP, KB, BNT, MLA, BNJ, KMD; [mɔʔ] BT ø	PPM
father	*ɐjoh	TB, LP; [ɐbɔh] MBA, KB, BNT, MLA, BNJ, KMD; [ɐjɔh] BT ø	PPM
brother	*ɐbɐ	TB, MBA, LP, BT, KB, BNT, MLA, BNJ, KMD; [ɐbɐ]	PPM
sister	*kɐkɐk	TB, MBA, LP, BT, KB; [kɐkɔʔ] BNT; [ɐkɔʔ] MLA, BNJ; [ɐkɐʔ] KMD; [ɐbɐ]	PPM
grand-father	*ɐki	TB, MBA, LP; [tɔʔˑkij] BT, BNT, MLA, BNJ, KMD; [ɐkij] KB; [ɐki]	PPM
grand-mother	*we	TB, MBA, LP; [tɔʔ] BT; [tɔʔːwuɛ] KB; [we] BNT; [wiʔ] MLA; [wɐˀ] BNJ, KMD; [wɛ]	PPM
younger sibling	*ɐdik	TB, MBA, LP, BT, KB, BNT, MLA, BNJ, KMD; [ɐdeʔ]	PPM
great-grand-parent	*mojɐ	TB, MBA, LP, BT, BNT, MLA, BNJ, KMD; [mojɐ] KB; [jɐ]	PPM
cousin	*šəpupu	TB, LP; [špupuw] MBA, BT, KB, BNT, KMD; [špupu] MLA, BNJ; [ø]	PPM
uncle	*pok. šədɐɣɐ	TB, MBA, LP, BT; [pɔʔˑsdɐɣə] KB, BNT; [pɔʔ] BNJ, KMD; [pɔʔˑšdɐʁə] MLA ø	PPM
aunt	*mok. šədɐɣɐ	TB, MBA, LP, BT; [mɔʔˑsdɐɣə] KMD; [mɔʔˑsdɐʁə] KB, BNT, MLA, BNJ ø	PPM
co-sibling-in-law	*biɣɐs	TB, BT; [biɣɛh] MBA, LP, KB; [biɣɐh] BNT, MLA; [biɣɐx] BNJ; [biʁɛh] KMD; [biʁeh]	PPM
co-parents-in-law	*bese	TB, LP, BNJ; [besɛɐ] MBA, MLA; [besɐˀ] BT, BNT; [besɐ] KB; [bese] KMD; [besɛ]	PPM
grandchild	*tʃutʃu	TB, LP, BT, BNT, MLA, KMD; [tʃutʃu] MBA, KB, BNJ; [ø]	PPM

Gloss	Proto lexicon	Correspondences variation	Ety.
father/ mother in-law	*mətuwʁ	TB, MBA, LP; [m̆ətuwʁ] BT, KB, BNT, BNJ, KMD; [mətuwə] MLA; [pɔʔˑmətuwə]	PPM
son/daughter in law	*m̆ənʁntu	TB, LP; [m̆ːnʁntuw] MBA, BT, BNT, MLA, KMD; [m̆nʁntu] BNJ; [mənʁntu] KB; [ø]	PPM
brother/ sister-in-law	*ipo	TB, MBA, LP, BT, KB, BNT, MLA, BNJ, KMD; [ipɔ]	PPM
step-relative	*tiɣi	TB, LP, BT, KB, BNT, MLA; [tiɣi] BNJ, KMD; [tiʁi] MBA ø	PPM

The uniformity in the use of an Austronesian-origin lexicon reflects a strong sense of familiarity and cohesion among family members. Table 13 also includes terms for more distant or infrequent kinship relations, such as step-relatives, which, while exhibiting a relatively lower degree of uniformity, still trace their origins to Austronesian roots. This indicates that the Malay Pahang community has largely maintained traditional kinship terminology and forms of address. Interestingly, it may also suggest limited intermarriage with outsiders, particularly those from foreign-language families. Nonetheless, this observation should not be regarded as conclusive, as individuals from non-Austronesian-speaking households may also incorporate Austronesian-origin terms within their families. Consequently, this finding does not constitute a definitive or theoretical conclusion.

Discussion

Listing reconstructed proto-forms alongside their corresponding phonological variants across geographic locations is a foundational method for developing a systematic proto-lexical inventory. In the context of the Pahang Proto-Malay dialect, this study has successfully reconstructed proto-lexicons across eleven lexical domains, using comparative phonological evidence from nine selected locations along the Pahang River.

The reconstruction process was guided by the methodological framework outlined by Crowley and Bowern (2010), resulting in a set of preliminary data that forms the basis for a potential digital inventory system. These eleven domains represent a thematically structured approach, similar to that practised by Che Rosdi et al. (2023), who extended their semantic domains from nine to eleven to improve the comprehensiveness of their classification.

However, not all reconstructed forms are of Austronesian or core Malay origin. Some proto-forms exhibit phonological or lexical traits traceable to foreign linguistic sources, including the Tai-Kadai and Mon-Khmer families. This suggests active linguistic contact and convergence in the region's historical landscape, where lexical borrowing may have occurred. Bischoffberger (2013), in her digital database of proto-forms, similarly categorises words believed to be borrowed from external language families, indicating this phenomenon is not unique to Pahang.

In light of this, this study proposes the integration of etymological indicators such as the label "LOAN" to mark reconstructed proto-forms that show strong evidence of borrowing. The inclusion of such indicators provides researchers with a nuanced understanding of each item's origin and enables more targeted filtering and analysis. This practice is inspired by the digital Austronesian Comparative Dictionary developed by Blust and Trussel (2020), which features proto-form metadata, including etymological status.

The selection of eleven domains reflects fundamental aspects of human experience and offers a holistic linguistic representation of the Proto-Pahang Malay dialect. These domains support in-depth intra-domain analysis, cross-domain comparison, and comparative classification with higher-order proto-languages, such as Proto-Malayic (PM) which was reconstructed by Adelaar (1992).

Nonetheless, this reconstruction represents only an initial phase. Further research is necessary to expand the lexical inventory beyond the current domains. Applying the same methodology, researchers can identify and reconstruct additional lexicons, thereby increasing the scope and depth of the dictionary. Digitising this dataset will not only facilitate regular updates and wider accessibility but will also allow for the integration of interactive functions such as audio playback, cross-referencing tools, and advanced search functions.

To ensure that these findings are sustainable and accessible to the wider scholarly community, the following sections will detail the strategies for digital development and the collaborative framework required for this long-term vision.

Current Situation and Future Development

This study concludes its first phase of reconstruction by establishing a foundational lexical dataset of eleven semantic domains across nine representative locations along the Pahang River. These reconstructed forms reflect both the prototypical phonology of the dialect and its geolinguistic distribution across regional subdialects.

A portion of the identified lexicon has been categorised as having foreign origin, particularly from Tai-Kadai, Sino-Tibetan, and Indo-European (specifically Germanic) families. To address this, the study proposes incorporating etymolog-

ical tags (e.g., "LOAN") within the digital display, thereby offering users insight into historical linguistic contact and potential borrowing.

In order to realise the long-term goal of establishing a comprehensive and openly accessible lexical database for the Pahang Proto-Malay dialect, this study proposes six strategic measures:

1. Development of an open-access digital repository that houses reconstructed proto-forms alongside supporting data such as semantic domains, pronunciation variants, etymological classifications, and audio files.
2. Implementation of open licensing frameworks, such as Creative Commons Attribution Share-Alike (CC-BY-SA), to facilitate legal reuse and redistribution of data.
3. Design of a taxonomic structure and user-friendly interface (UI) that enables advanced filtering and cross-referencing by domain, location, or etymological origin.
4. Phased system development, beginning with a static, read-only display, followed by the integration of audio features and cross-linking functions, and eventually allowing controlled expert contributions through collaborative editing features.
5. Alignment with established repositories such as the Austronesian Comparative Dictionary (Blust & Trussel, 2020), SEAlang, Starling, PARADISEC, and ELAR to ensure broader dissemination and long-term preservation.
6. Gradual publication of results in indexed, high-impact academic journals, ensuring both scholarly recognition and sustainable momentum for the project's ongoing development.
7. Together, these strategies are expected to form the technical and academic foundation of a long-lasting, dynamic, and internationally accessible proto-lexical database that contributes to the field of historical Malay linguistics and beyond.

Collaboration

The creation of a comprehensive digital proto-lexicon cannot be achieved by a single researcher working in isolation. Its development necessitates interdisciplinary collaboration involving linguistics, computer science, audio archiving, and digital data management. From the earliest stages, this study has engaged with individuals with programming and UI/UX design expertise to build an interactive system capable of displaying various data formats, including mapped variants, metadata, and audio playback.

Equally important is collaboration with phoneticians and phonologists to ensure accurate transcription and audio representation of proto-sounds. Using tech-

niques such as elicited phonetic approximation, these audio files serve to preserve aural aspects of the reconstructed dialect, thereby enhancing the documentation and educational value of the resource.

In the medium- and long-term, further partnerships with research institutions, digital archives, and non-profit organisations are essential for ensuring sustainability, system upkeep, and data updates. Platforms such as ELAR (Endangered Languages Archive), PARADISEC, and the SEAlang Library have been identified as potential collaborators for repository integration and data maintenance.

Additionally, community-level collaboration with local speakers, educators, and cultural practitioners along the Pahang River is envisioned to broaden the utility of this digital resource. Such involvement may support dialect revitalisation, oral history preservation, and community-led language documentation.

To Sum Up

Since this study has successfully reconstructed eleven semantic domains of the proto-lexicon of the Pahang Malay dialect, several subsequent steps remain for researchers to pursue. To summarise the broader workflow involved in the development of a proto-lexicon dictionary for the Pahang River Malay dialect, Figure 2 below outlines the key phases of the project:

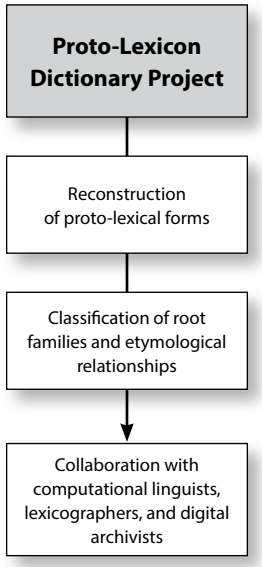

Figure 2. Basic structure of the main phases in the Proto-Lexicon Dictionary project

Following the steps illustrated in the Figure 2 above, the next logical phase involves the classification of reconstructed forms into root families and the identification of their etymological origins whether Austronesian or otherwise. This classification is essential for establishing a historically grounded and semantically coherent lexical system. Ultimately, collaboration with computational linguists, lexicographers, and digital archivists will be critical to transform this lexical data into an accessible, open-source, and dynamically searchable digital platform. Such a platform would not only facilitate cross-linguistic comparison and phonological alignment but also support long-term preservation and scholarly accessibility of the proto-language data.

Conclusion

The digitalisation of the proto-lexical inventory of the Pahang River Malay dialect necessarily begins with systematic reconstruction. This study has undertaken such a task by applying established reconstruction theory to field-collected data, successfully identifying proto-forms across eleven semantic domains. The findings reveal not only core Austronesian lexicon but also a degree of lexical borrowing from neighbouring language families, reflecting a long history of linguistic contact and sociocultural integration.

This research contributes foundational data for a comprehensive digital proto-dictionary and outlines strategic steps for the development of an open-access, interactive, and sustainable platform. By incorporating etymological indicators and audio features, this initiative aligns with emerging best practices in digital language documentation.

The study also highlights the necessity of interdisciplinary collaboration in ensuring the scalability, accessibility, and long-term maintenance of such a linguistic resource. It recognises the importance of engaging experts in phonetics, computational linguistics, and digital archiving, as well as local community stakeholders.

It is hoped that future research will continue to expand the lexical inventory beyond the initial eleven domains, and that this digital resource will not only serve academic inquiry but also promote awareness, preservation, and revitalisation of the Pahang River Malay dialect within the broader Austronesian heritage.

Declaration of AI Use

The author(s) disclose the use of a generative AI tool (ChatGPT version GPT-4o) solely for language editing and/or technical assistance. The author(s) confirm that they are fully responsible for the content, originality, interpretation, and conclusions of this article.

REFERENCES

Adelaar, K. A. (1992). *Proto-Malayic: The reconstruction of its phonology and parts of its lexicon and morphology*. Pacific Linguistics.

Adil, B. (1972). *Sejarah Pahang [History of Pahang]*. Dewan Bahasa dan Pustaka.

Alikhon, L. (2022). Pronouns are the most original words in the language. *JETA (Journal of English Teaching and Applied Linguistic)*, *3*(2), 91-97. https://doi.org/10.52217/jeta.v3i2.1033

Ayatrohaedi. (1979). *Dialektologi: Sebuah pengantar* [Dialectology: An introduction]. Pusat Pembinaan dan Pengembangan Bahasa Departemen Pendidikan dan Kebudayaan.

Bischoffberger, J. (2013). *Wordlist Proto Malayic*. ASJP Database. https://asjp.clld.org/languages/PROTO_MALAYIC

Blust, R. A. (1970). Proto-Austronesian addenda. *Oceanic Linguistics*, *9*(2), 104-162. https://doi.org/10.2307/3622723

Blust, R., & Trussel, S. (2020). Austronesian Comparative Dictionary. https://www.trussel2.com/ACD/

Bynon, T. (1994). *Linguistik sejarawi* [Historical linguistics] (N. E. Mohd Noor, Ed. & Trans.). Dewan Bahasa dan Pustaka.

Campbell, L. (2020). *Historical linguistics: An introduction* (4th ed.). The MIT Press.

Cant, G. R. (1972). *An historical geography of Pahang*. Malaysian Branch of the Royal Asiatic Society.

Chambers, J. K., & Peter, T. (2004). *Dialectology* (2nd ed.). Cambridge University Press.

Che Rosdi, N. H., Aman, R., & A. Hamid, S. (2023). *Dialek Hulu Perak Utara: Pendekatan linguistik perbandingan* [The dialects of North Hulu Perak: A comparative linguistic approach]. Penerbit Universiti Kebangsaan Malaysia.

Collins, J. T. (1989). Malay dialect research in Malaysia: The issue of perspective. *Bijdragen tot de Taal-, Land- en Volkenkunde*, *145*(2/3), 235-264.

Crowley, T., & Bowern, C. (2010). *An introduction to historical linguistics* (4th ed.). Oxford University Press.

Dempwolff, O. (1938). *Austronesisches Wörterverzeichnis (Vergleichende Lautlehre des austronesischen Wortschatzes, III)*. Reimer.

Hamilton, A. W. (1925). Some Malay words. *Journal of the Malayan Branch of the Royal Asiatic Society*, *3*(95), 32-55.

Hasrah, M. T., Aman, R., & A. Hamid, S. (2013). Inovasi dan retensi dalam dialek Hulu Tembeling [Innovation and retention in the Hulu Tembeling dialect]. *GEMA Online® Journal of Language Studies*, *13*(3), 211-222. https://journalarticle.ukm.my/6603/1/4230-9741-1-SM.pdf

Hasrah, M. T., Aman, R., & A. Hamid, S. (2014). *Fosil dialek Melayu Hulu Pahang* [Fossils of the Hulu Pahang Malay dialect]. Penerbit Universiti Kebangsaan Malaysia.

Hussein, I. (1973). Malay dialect in the Malay Peninsula. *Jurnal Nusantara*, *3*, 69-79.

Idris, A. (1989). Fonologi dialek Pahang: Satu kajian awal di kampung Sentang [Phonology of Pahang dialect: A preliminary study in Sentang village]. *Jurnal Dewan Bahasa*, *33*(6), 415-431.

Karim, N. S., & Ibrahim, T. (1977). *Penyelidikan bahasa di Ulu Tembeling* [Language research in Ulu Tembeling]. Jabatan Pengajian Melayu.

Linehan, W. (1928a). Notes on the remains of some ancient brick structures in Pekan district. *Journal of the Malayan Branch of the Royal Asiatic Society*, *6*(4), 78-81. https://www.jstor.org/stable/41559723

Linehan, W. (1928b). Some discoveries on the Tembeling. *Journal of the Malayan Branch of the Royal Asiatic Society*, *6*(4), 66-77. https://www.jstor.org/stable/41559722

Linehan, W. (1930). Notes on some further archaeological discoveries in Pahang. *Journal of the Malayan Branch of the Royal Asiatic Society*, *8*(2), 314-317. https://www.jstor.org/stable/41559754

Linehan, W. (1936). A history of Pahang. *Journal of the Malayan Branch of the Royal Asiatic Society*, *14*(2), 1-257. http://www.jstor.org/stable/41585562

Linehan, W. (2020). *A history of Pahang* (R. Krishnan, Ed.). Silverfish.

Omar, A. (2015). *Susur galur bahasa Melayu* [Malay language lineage] (2nd ed.). Dewan Bahasa dan Pustaka.

Omar, A. (2019). *Rekonstruksi kata dalam bahasa Melayu induk* [Word reconstruction in native Malay]. Dewan Bahasa dan Pustaka.

Pyysalo, J. (2017). Proto-Indo-European lexicon: The generative etymological dictionary of Indo-European languages. In J. Tiedemann (Ed.), *NoDaLiDa 2017 – 21st Nordic Conference of Computational Linguistics, Proceedings of the Conference* (pp. 259-262). Linköping University Electronic Press. https://aclanthology.org/W17-0234.pdf

Shorto, H. L. (2006). *A Mon-Khmer comparative dictionary* (P. Sidwell, D. Cooper, & C. Bauer, Eds.). Pacific Linguistics, Research School of Pacific and Asian Studies, The Australian National University. https://doi.org/10.15144/PL-579

Starostin, S. (n.d.). *Starling database server*. https://starlingdb.org/cgi-bin/query.cgi?root=config&morpho=0&basename=%5Cdata%5Cthai%5Ckdaiet

Uthai, R. (1993). *A comparison of word formation in Standard Malay and Pattani Malay* [Unpublished master's thesis]. Chulalongkorn University.

Wilkinson, R. J. (1932). *A Malay-English dictionary: Romanised*. Salavopoulos and Kinderlis.

Wu, J. (2023). *Malayic varieties of Kelantan and Terengganu: Description and linguistic history*. LOT. https://doi.org/10.48273/LOT0651

Zaidi, M. N., Aman, R., & A. Hamid, S. (2021). *Genetik purba dialek Melayu: Di lembangan tengah sungai Pahang* [Ancient genetics of Malay dialects: In the central Pahang river basin]. Penerbit Universiti Kebangsaan Malaysia.