ISSN 1798-4769 (print); ISSN 2053-0684 (online)
Journal of Language Teaching and Research, Vol. 17, No. 2, pp. 463-476, March 2026
DOI: https://doi.org/10.17507/jltr.1702.09

A Computational Dialectology Approach to Mapping Bidayuhic Varieties in Tayan Hulu Using Gabmap

Dedy Ari Asfar
Teachers' Training and Education Faculty, Tanjungpura University, Pontianak, Indonesia

Syarifah Lubna
Language, Literature, and Community Research Center, National Research and Innovation Agency (BRIN), Jakarta, Indonesia

Irmayani Abdulmalik
Language, Literature, and Community Research Center, National Research and Innovation Agency (BRIN), Jakarta, Indonesia

Wiwin Erni Siti Nurlina
Language, Literature, and Community Research Center, National Research and Innovation Agency (BRIN), Jakarta, Indonesia

Edi Setiyanto
Language, Literature, and Community Research Center, National Research and Innovation Agency (BRIN), Jakarta, Indonesia

Sutarsih
Language, Literature, and Community Research Center, National Research and Innovation Agency (BRIN), Jakarta, Indonesia

Yusup Irawan
Language, Literature, and Community Research Center, National Research and Innovation Agency (BRIN), Jakarta, Indonesia

Binar Kurniasari Febrianti
Language, Literature, and Community Research Center, National Research and Innovation Agency (BRIN), Jakarta, Indonesia

Yeni Yulianti
Language, Literature, and Community Research Center, National Research and Innovation Agency (BRIN), Jakarta, Indonesia

Sarwo Ferdi Wibowo
Manuscript, Literature, and Oral Tradition Research Center, National Research and Innovation Agency (BRIN), Jakarta, Indonesia

Febyasti Davela Ramadini
Language, Literature, and Community Research Center, National Research and Innovation Agency (BRIN), Jakarta, Indonesia

Ajeng Rahayu Tjaraka
Language, Literature, and Community Research Center, National Research and Innovation Agency (BRIN), Jakarta, Indonesia

Prima Duantika
Balai Bahasa Provinsi Kalimantan Barat, Indonesia

Abstract—**This study examines the linguistic variation of the Bidayuhic language in Tayan Hulu, West Kalimantan, Indonesia, through a computational dialectological approach using Gabmap. This study applies Levenshtein Distance to measure lexical and phonological differences in six observation sites, analyzing 491 lexical items. The findings show that the Bidayuhic language forms a linguistic continuum, where dialectal variation is not entirely aligned with geographical boundaries. Instead, lexical and phonological differences are influenced by language contact, social mobility, and cultural interaction. This study identifies the merger of the Proto-Malayo-Polynesian (PMP) phonemes R and l into /r/, /ɣ/, and /h/, reflecting phonological innovations in the Bidayuhic language. Furthermore, ablaut in verb morphology is observed, distinguishing between transitive and intransitive verb forms. Cluster analysis via Multidimensional Scaling (MDS) and probabilistic clustering revealed two main groups, confirming that variation is gradual rather than regionally segmented. Despite adding 0.8 probabilistic disturbances, the clustering remained stable, validating the effectiveness of Gabmap in dialect classification. These results emphasize that Bidayuhic variation is shaped more by sociolinguistic interactions than geographical factors. This study highlights the role of Gabmap in linguistic mapping, offering a methodological model for mapping local languages in Indonesia.**

Index Terms—**Bidayuhic, Austronesian language, dialectometry, Gabmap, computational dialectology**

I. Introduction

Local language documentation plays an important role in preserving language diversity by safeguarding cultural heritage, informing language policy, and utilizing technological advances. Local language documentation not only ensures the continuity of language structure but also protects the cultural narratives embedded within it, which are important for maintaining the identities of diverse communities (Francois et al., 2023; Wei & Schnell, 2025). In addition, the dominance of national and global languages continues to threaten language diversity, making documentation efforts increasingly urgent (Lendik & Yuit, 2021). Moreover, from an educational and policy perspective, research on local languages makes a significant contribution to bilingual education and informs language planning by integrating local languages into the curriculum, thus promoting language inclusivity (Mwelwa & Spencer, 2013; Yumnam & Singh, 2024).

Along with the increasing urgency of local language documentation, technological developments have presented various innovative solutions to preserve it. The use of digital devices not only increases efficiency in language documentation but also expands accessibility to linguistic data, especially in remote areas. One prominent initiative is LangDoc, which has proven capable of increasing documentation accuracy in areas with limited access to technology (Spencer, 2024). However, fieldwork in marginalized communities still faces various practical challenges, especially in the process of data collection and analysis, so a more adaptive and innovative approach is needed (Nath, 2008). In addition to technical aspects, the development of standard orthography is also an important factor in supporting literacy and the preparation of educational materials for less documented languages, which ultimately contributes to the survival of the language (Chebanne, 2016). The success of these various documentation efforts can be seen from global and regional initiatives, as evidenced in case studies from the Pacific, Amazonia, and South Africa, which demonstrate diverse and effective strategies in revitalizing local languages and ensuring their continued use in various communities (Beier & Epps, 2020; François, 2020).

In the Indonesian context, one of the regional languages facing challenges in documentation and preservation is Bidayuhic, a language spoken in West Kalimantan, Indonesia. Bidayuhic is a member of the Austronesian language family in the Dayak subgroup of Land Dayak and is known to have a complex variety of dialects (Asfar, 2014; Chong & Gedat, 2012). These dialects are generally classified into three main groups—east, south, and north—each of which exhibits unique linguistic characteristics (Chong & Gedat, 2012). One of the remarkable characteristics of Bidayuhic is the retention of Proto Malay-Polynesian (PMP) forms, a rare feature among Austronesian languages in western Borneo (Asfar, 2015; Chong & Gedat, 2012). In addition, this language shows ablaut, a process of vowel alternation that is not common in other Austronesian languages (Collins, 2021). Phonologically, certain dialects, such as Hliboi, show geminate consonants at the beginning of words, a feature that significantly shapes their phonetic structure (Smith, 2021). Furthermore, changes in historical pronunciation, including syllable complexity and vowel breaking, have influenced the evolution of the language (Smith, 2021). Despite its strong presence in the Bidayuh region of Sarawak, Malaysia, Bidayuhic is facing a gradual shift among younger speakers, which signals a potential challenge to its long-term vitality (Coluzzi et al., 2013). Given its linguistic uniqueness and signs of language shift, further research and documentation efforts are crucial to preserve Bidayuhic as part of the region's rich linguistic heritage.

Faced with the complexity of dialect variation in the Bidayuh language, an accurate analytical approach is a primary necessity in understanding its linguistic patterns. In this case, traditional descriptive dialectology methods often face limitations in systematically capturing variation patterns. Therefore, computation-based approaches such as Gabmap offer more objective quantitative solutions in mapping dialectal differences. Compared to traditional dialectology methods that tend to be descriptive and manual labor-based, Gabmap offers a more quantitative and automated application-based approach to language variation analysis.

One of Gabmap's main advantages is its ability to provide advanced visualizations, such as distribution maps and colored multidimensional diagrams, which enable more intuitive and systematic data representation (Nerbonne, 2011). With this feature, dialectological research can not only identify phonetic and lexical differences between language

varieties but also display patterns of language variation in a form that is easier for researchers and the public to interpret (Wieling et al., 2016). In addition, Gabmap supports statistical-based quantitative analysis that enables more objective data exploration, including the use of histograms, alignment of phonetic transcriptions, and cluster analysis that groups dialects based on their linguistic similarities (Leinonen et al., 2016).

This approach provides significant advantages over traditional dialectology, which relies more on subjective intuition and manual analysis, which is often time-consuming and prone to inconsistencies. Gabmap is becoming an increasingly important tool in modern dialectological research because it offers advantages in visualization, quantitative analysis, and the ability to facilitate cross-language and cross-modal research. Approaches such as Gabmap are part of a broader development in computational dialectology, a field that integrates computational and statistical techniques in the analysis of language variation.

The computational dialectology approach has brought significant progress in mapping language variation with higher efficiency and accuracy compared to traditional methods. This progress is mainly supported by its ability to handle large-scale data sets. Quantitatively, dialectometry utilizes statistical techniques to measure and represent linguistic features spatially, including multidimensional scaling and fuzzy clustering to compare dialect maps based on syntactic and phonological features (Pröll, 2013; Spruit, 2006). Furthermore, this approach continues to evolve by simultaneously incorporating geographical, social, and linguistic factors, as applied in research on Dutch phonetic variation, which considers community size, speaker age, and word frequency in predicting pronunciation distance (Wieling et al., 2011).

In line with this development, computational dialectology not only focuses on mapping statistical-based linguistic variations but also continues to innovate in its analysis methods. Quantitative studies in computational dialectology increasingly highlight the relationship between language variation and social and geographical factors (Nerbonne & Kretzschmar Jr., 2006; Wieling & Nerbonne, 2015). Innovations in computational dialectology also include the use of Reproducing Kernel Hilbert Space (RKHS), which allows for nonparametric measurement of language variation without reliance on fixed geographical boundaries (Nguyen & Eisenstein, 2017). In addition, the application of Computational Construction Grammar in mapping syntactic features in various languages has provided a broader model in the analysis of regional variations (Dunn, 2019). The integration of computational dialectology with geographic information systems (GIS) further enriches the visualization of language variation by linking linguistic data with relevant social and geographic factors (Kehrein, 2012).

With an increasingly broad scope of analysis, from phonetics to syntax, this approach provides deeper insights into the complexity of language variation, as shown in recent research on tonal and segmental dialectometry in the Yue and Pinghua speech regions (Sung et al., 2024). Therefore, computational dialectology offers a more comprehensive, objective, and efficient approach to understanding the dynamics of language variation in various regions and communities.

Although this development has been widely applied in various language studies around the world, its use in research on local languages in Indonesia is still relatively limited. One example that has received insufficient attention is the dialectal variation of the Bidayuhic language. Although previous research has discussed the dialectal variation of the Bidayuhic language and the application of traditional methods in dialectology, studies that specifically utilize computational dialectology, especially with Gabmap, are still limited in the context of the Bidayuhic language in Indonesia. Most previous studies have focused on descriptive and manual aspects of dialect mapping without utilizing a more objective quantitative-based approach. Therefore, this study offers novelty by applying Gabmap to systematically analyze Bidayuhic dialectal variations, using statistical methods and technology-based visualization to improve the accuracy of linguistic mapping and understand the dynamics of phonetic and lexical differences in the Bidayuhic community.

This research aims to apply Gabmap as a dialectometric analysis tool in mapping the variation of the Bidayuhic language in Tayan Hulu. In addition, this research aims to identify dialectal zones based on lexical and phonological differences using algorithm-based mapping methods. In line with this objective, this study focuses on mapping the dialectal variations of the Bidayuhic language in Tayan Hulu, West Kalimantan, using a computational dialectology approach with Gabmap. With this approach, it is hoped that more systematic and objective dialectal mapping can be obtained, providing a more in-depth understanding of the patterns of Bidayuhic language variation. Academically, this research contributes to the development of computational dialectology methods in regional language studies, as well as enriching Austronesian dialectology studies more broadly. Meanwhile, for the wider community, the results of this research are expected to form the basis for efforts to revitalize and document the Bidayuhic language, as well as provide policy recommendations in language education and linguistic planning to preserve the diversity of regional languages in Indonesia. Thus, this research is expected to provide new insights into the use of technology in the analysis of language variation and open opportunities for computation-based linguistic research in the future.

II. Methodology

This study applies a descriptive-comparative approach with a computational dialectology method, which allows for the analysis of dialectal variation based on quantitative techniques. Gabmap is used as the main tool in mapping dialectal variation, enabling data-based analysis and spatial visualization. The research was conducted in various villages in Tayan Hulu, West Kalimantan, Indonesia, focusing on the Bidayuhic community of speakers. Six observation points were selected for this study. The determination of the six observation points was carried out using the snowball sampling

method (Contandriopoulos et al., 2019; Isaías et al., 2012) because there is an emblematic phenomenon of disambiguation that always exaggerates the differences in group identity based on language in West Kalimantan (Collins, 2018, 2021). To obtain an accurate representation, the research participants consisted of native speakers of Bidayuhic from various generations to capture intergenerational variations in their dialects. In addition, this study applies the NORMs (Non-mobile Older Rural Males) method developed by Chambers (2015) and includes elderly female informants according to the research model by Collins (2021) to get a sample that reflects the conservative dialect that still survives.

The object of this research is the Bidayuhic language as spoken by local communities in six villages along the upper reaches of the Tayan River. The research population includes all Bidayuhic language utterances, including linguistic and non-linguistic aspects. The selected sample includes utterances that have been defined in a word list consisting of 200 Swadesh basic vocabularies, which was then expanded to 491 Bidayuhic language-specific lexemes (Collins, 2021).

In addition, phonetic interviews are applied to capture pronunciation differences and phonological aspects using direct and indirect elicitation techniques. In the direct elicitation technique, informants are asked to explicitly name a specific object, such as when the researcher asks, "What is galangal called in the local language?" On the other hand, in the indirect elicitation technique, the researcher shows or describes an object without mentioning the intended term or lexicon, allowing for a more natural response from the informant, for example, by pointing to the hand or tongue while asking, "What is this called in the local language?" (Asfar, 2016; Chambers, 2015; Effendy et al., 2023; Irawan et al., 2024).

Furthermore, this research applies dialect classification methods through clustering techniques, which are available in Gabmap. Some of the algorithms used include hierarchical clustering, weighted averaging, and Ward's method. Hierarchical clustering allows for tiered grouping based on the similarity between language variants, while weighted averaging is used to consider the weight of linguistic differences in the grouping. Ward's method, on the other hand, helps produce a more balanced classification by minimizing variance within each group (Leinonen et al., 2016). With this combination of techniques, the research can identify the distribution patterns of the Bidayuhic group more comprehensively.

Additionally, another main method in this analysis is dialectometry, which is used to measure the similarities and differences between the dialects being studied. The calculation is performed using the Levenshtein distance, which is a technique that measures the degree of difference between two linguistic forms based on the number of changes (insertions, deletions, or substitutions) required to transform one form into another (Nerbonne, 2011). This approach allows for a more quantitative and objective analysis of dialectal variation.

To present the analysis results more clearly, Gabmap is used in data visualization through various forms of dialectal mapping, such as beam maps, network maps, multidimensional scaling (MDS) plots, and cluster maps. Beam maps and network maps help illustrate the connections between research locations based on linguistic differences, while MDS plots are used to compare classification results to understand how far dialectal variations can be grouped into specific linguistic areas. Cluster maps, on the other hand, allow for the geographical mapping of dialect groupings that have been analyzed using dialectometry techniques (Leinonen et al., 2016; Nerbonne, 2011).

With this approach, this research not only provides a more systematic understanding of the dialectal variations of the Bidayuhic group but also produces more accurate and replicable data for further studies. The results of this research are expected to contribute to language policy planning, especially in efforts to preserve regional dialects that are increasingly threatened by linguistic homogenization due to the dominance of national and global languages.

III. FINDINGS

A. Distribution Maps of Bidayuhic in Tayan Hulu

Figure 1. Research Observation Point Map

The data studied in this research comes from Bidayuhic in the upper reaches of the Tayan River, displaying a base map with six georeferenced points as seen in Figure 2 below. The sixth observation point of these georeferences was precisely

created using the Google Earth Pro application. The six observation points have six variations hypothesized to be in the Bidayuhic language (Figure 1, above), extracted and presented in Figure 2. Administratively, there are five observation points originating from the Tayan Hulu District and one observation point located within the administration of the Tayan Hilir District. As a result, one point is outside the boundaries of the Tayan Hulu District. As for the five observation points in the Tayan Hulu District, they locally identify themselves as Dayak Pruwan, Dayak Banyadu', Dayak Hibun/Ribun, and Dayak Taba/Temiang Taba, while one point in the Tayan Hilir District in Senyabang Village identifies the local ethnic group as Dayak Keneles.

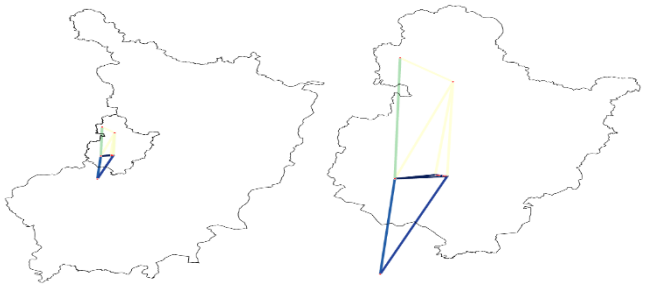

Figure 2. Map of the Selected Georeferenced Points of Six Bidayuhic Tayan Hulu

In this study, the digital data that were compared and classified include 491 lexical items related to the human body, kinship, animals, colors, numbers, directions, and cultural vocabulary. In addition, this data also includes a subset of various nouns and verbs that are part of the corpus collection in Gabmap. This corpus was collected from six different locations, allowing for comparative analysis of lexical variation in the language used. With a wordlist-based approach, this research presents a statistical overview of word distribution in a corpus, which serves as the foundation for linguistic analysis.

Gabmap notes that in the entire Bidayuhic Tayan Hulu text uploaded to the application, there are 16,300 characters with 53 unique characters, reflecting the variation of letters, numbers, and symbols used in the data. Overall, the number of tokens detected reached 16,187, but it only consisted of 74 unique tokens. This indicates a high level of word repetition in the dataset, which has implications for the patterns of use and lexical distribution in the analyzed language. This statistic serves as an indicator of lexical diversity and consistency within the corpus, ultimately providing insights into the structure and patterns in the analyzed language data.

The lexical distribution map shows the distribution of the word Bidayuhic Tayan Hulu *kulit* ("skin") in Figure 4 and the distribution of the Bidayuhic Tayan Hulu word *telur* ("egg") in Figure 3, which can be seen clearly and in detail on Gabmap. This map shows specific variants of pronunciations, indicating where the selected variants can be found. In addition, this lexical distribution shows the number of words used, providing a more complete picture of the distribution of those words and the locations where they are used.

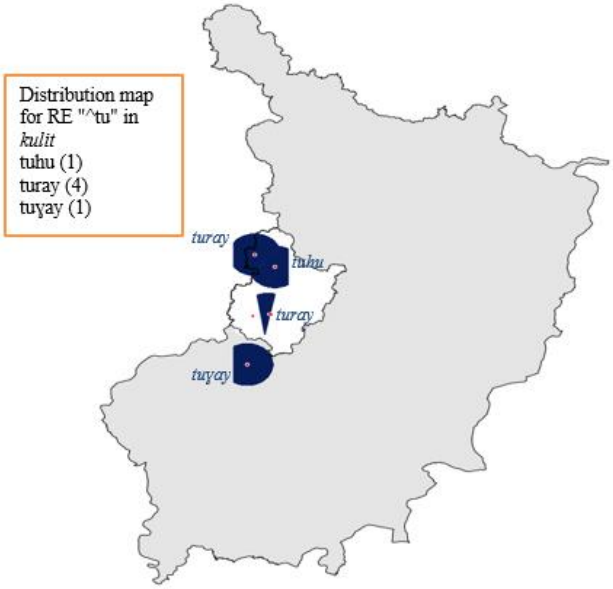

Figure 3. Distribution Map for Pronunciations of the Word *Telur* ("Egg")

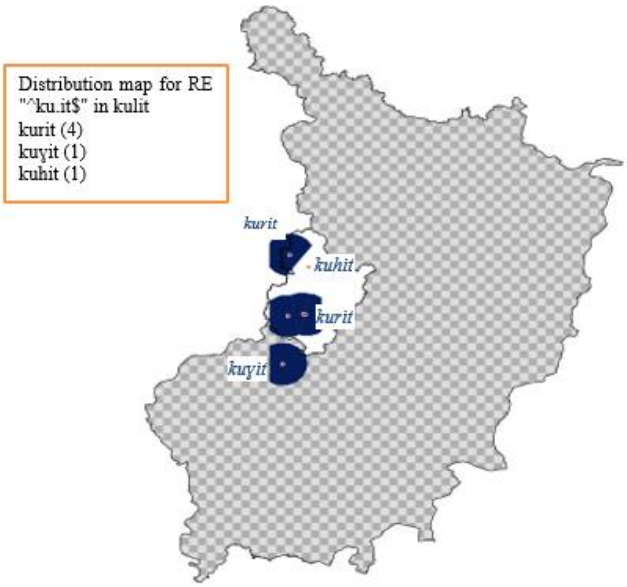

Figure 4. Distribution Map for Pronunciations of the Word *Kulit* ("Skin")

This Bidayuhic Tayan Hulu map shows its affirmation as a Bidayuhic group. This can be seen through one of the Proto-Malay Polynesian (PMP) *R > /r/, /h/ /#/ dan fonem PMP *l > r (Asfar, 2014; Collins, 2021). As a result, it has been seen in the word distribution maps in Figure 3 and Figure 4 that in Bidayuhic Tayan Hulu there is also an innovation in the PMP phoneme *l, namely *l > r. PMP *l appears as /r/ in the Menyabo, Tabat, and Kubing variants, *l appears as /ɣ/ in the Senyabang variant, and *l appears as /h/ in the Riyai variant. Thus, it can be said that *R and *l PMP have undergone a merger in the Bidayuhic variant in Sungai Tayan Hulu diachronically, as they show typical sound correspondences in this group.

The lexical distribution map of *mati* ("die") (Figure 5) and the distribution of the word *membunuh* ("kill") (Figure 6) of the Bidayuhic Tayan Hulu can be seen in Gabmap with clarity and detail. This map shows the specific variants (pronunciations), indicating where the chosen variant can be found. In addition, the lexical distributions in these two maps when compared show the ablaut phenomenon, which in the Bidayuhic group is considered a retention of Proto-Malay Polynesian (PMP) that has been lost in some variants (Collins, 2021).

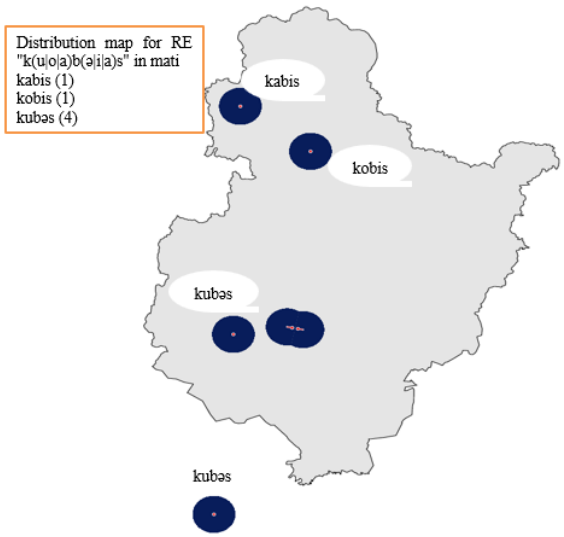

Figure 5. Distribution Map for Pronunciations of the Word *Mati* ("Die")

As shown in Figure 5, the distribution of the word *mati* ("die") in several Bidayuhic variants in Tayan Hulu shows variation in the use of words to refer to the concept of death. In the Dayak Taba/Temiang Taba, Dayak Pruwan (Tabat and Kubing), and Dayak Senyabang Keneles variants, the word used is *kubəs*, which shows similarity in lexical form among these variants. Meanwhile, in the Berakak Dayak Banyadu variant, the word used is *kabis*, which is slightly different in pronunciation but still refers to the concept of death. Lastly, in the Dayak Hibun/Ribun (Riyai) variant, the word used is *kobis*, which undergoes a vowel change compared to the other variants. The difference in lexical forms indicates the

presence of phonological variation in the use of words to express death, which is related to geographical factors or dialects within the communities that use these variants.

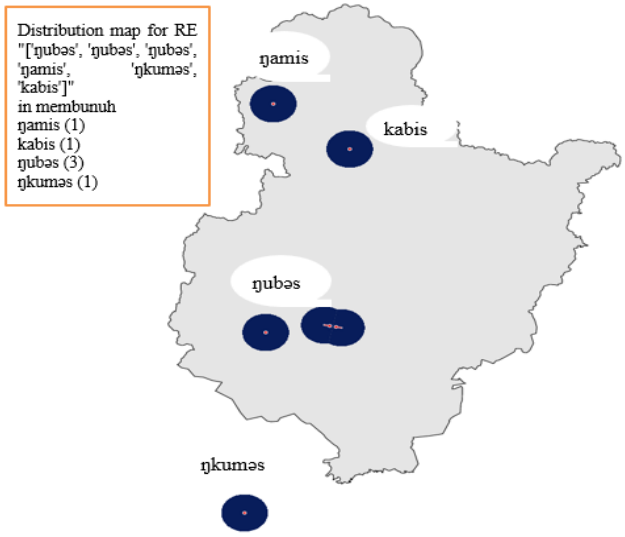

Figure 6. Distribution Map for Pronunciations of the Word *Membunuh* ("Kill")

Distribution of the word *membunuh* ("kill") in several Bidayuhic variants in Tayan Hulu shows different pronunciations, as seen in Figure 6, to express the act of killing. In the Dayak Taba/Temiang Taba, Dayak Pruwan (Tabat and Kubing), and Dayak Senyabang Keneles variants, the word used is *ŋubəs*, which has a consistent pronunciation across these three variants. Meanwhile, in the Dayak Banyadu variant, the word used is *ŋamis*, which is phonologically different from the other variants. In the Dayak Hibun/Ribun (Riyai) variant, the word used is *kabis*, which also shows a vowel change compared to other variants. The difference in the pronunciation of the word *membunuh* reflects the presence of phonological variations that may be influenced by dialectal and geographical factors of each community in the Tayan Hulu region.

B. Classification of Bidayuhic Tayan Hulu Using Gabmap

The classification of the Bidayuhic group in Tayan Hulu uses the dialectometric analysis available in Gabmap. This analysis is usually based on the linguistic distance between pairs of words in the data. The distance measure used for string data is the string edit distance (or Levenshtein Distance). The edit distance of a string calculates the minimum number of insertions, deletions, and substitutions required to transform one character string into another (Leinonen et al., 2016). The results of the pronunciation distance measurements for each pair of words in the sample are 491 lexical items. The results of these measurements can be seen through multidimensional scaling (MDS) in Gabmap, which proposes a set of dimensions and coordinates for each input point.

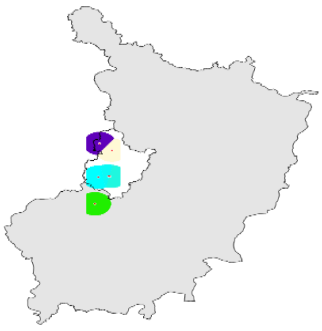

Figure 7. MDS Map of Lexicon

The first classification that considers this general similarity is projected onto the MDS map (with r = 0.97) in Figure 7. This MDS map combines the measured lexical differences among the six Bidayuhic variants in Tayan Hulu. This MDS classification not only confirms that the Bidayuhic variants of Tayan Hulu form a language continuum, meaning that one language variation gradually merges into another when they are close together, but also that this continuum contains several main subdivisions, corresponding to lexical differences (purple, cream, turquoise, and green aggregates). This can be seen in Bidayuhic Berakak Banyadu' (purple aggregate), Bidayuhic Riyai Hibun/Ribun (cream aggregate), Bidayuhic Pruwan and Temiang Taba (turquoise aggregate), and Bidayuhic Keneles (green aggregate).

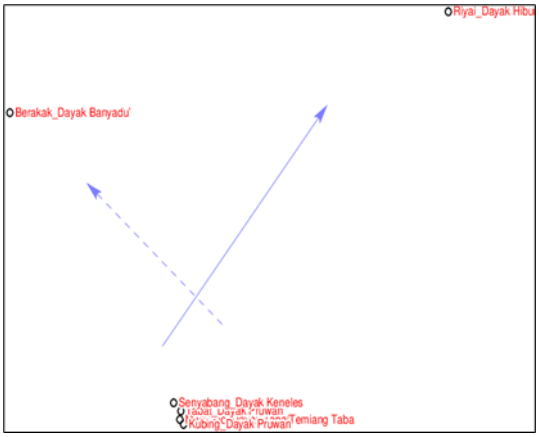

Figure 8. MDS Plot Map of Lexicon

Figure 8 shows the validated aggregates through the appropriate MDS scatter plot, a stable technique offered by Gabmap for this purpose and which displays 100% variation in the data. The distance measured by the plot shows a high correlation with the distance provided in the linguistic distance table, with a value of r=0.98. Another technique that confirms that the obtained aggregates are very stable and thus explains the adequate appearance of the data aggregates is probabilistic clustering techniques. This probabilistic clustering essentially consists of continuously adding a quantity of noise during the clustering process and maintaining the cophenetic distance from the compared sites (Lafkioui, 2018; Nerbonne et al., 2008). Even after 0.8 noise is added, while the default additional noise is 0.2, the aggregate remains stable.

The dendrogram shown in Figure 9 illustrates clustering with an additional noise level of 0.2. The stability of the main aggregate of the Bidayuhic Tayan Hulu lexicon is also verified by other algorithmic classification techniques, as shown in Figures 10, 11, and 12, which present the clustering classification results based on the weighted average algorithm. Thus, the Bidayuhic Tayan Hulu data reinforces that this algorithm has the advantage of producing consistent and representative clusters because it assigns the same weight to the joining clusters, even though the number of sites forming each cluster is not the same. Note that this cluster is also validated through the Gabmap cluster validation technique, which utilizes MDS and its two-dimensional plot.

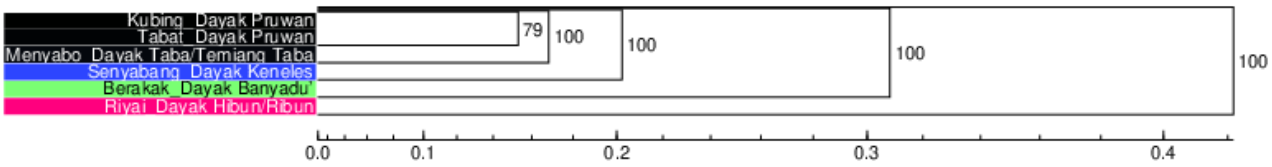

Figure 9. Probabilistic Dendrogram of the Bidayuhic Lexicon of Tayan Hulu

In the Gabmap application, numbers like 79, 100, 100, 100, 100 in the probabilistic dendrogram indicate the level of certainty or probability that these language groups form a cluster at a certain distance. Here, 100% indicates that the languages in that group are certainly in one cluster without doubt, while 79% indicates that there is a 21% chance that the languages in that group could be classified into another group at a certain distance. This means that at a distance of 0.2 (20%), the language group Keneles, Pruwan, and Taba forms a single cluster with 100% certainty using the weighted average method, while the two Pruwan variants appear to have a closer relationship but only with 79% certainty, which means there is still a chance they could join another group if the clustering method or distance changes. Thus, the numbers in this dendrogram reflect the strength of clustering, indicating that the higher the number, the stronger the inter-language relationships within that group.

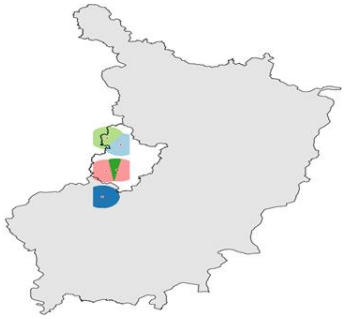

Figure 10. Map of the Bidayuhic Tayan Hulu Cluster Analysis

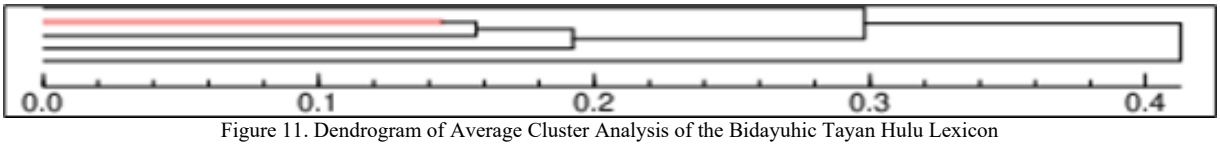

Figure 11. Dendrogram of Average Cluster Analysis of the Bidayuhic Tayan Hulu Lexicon

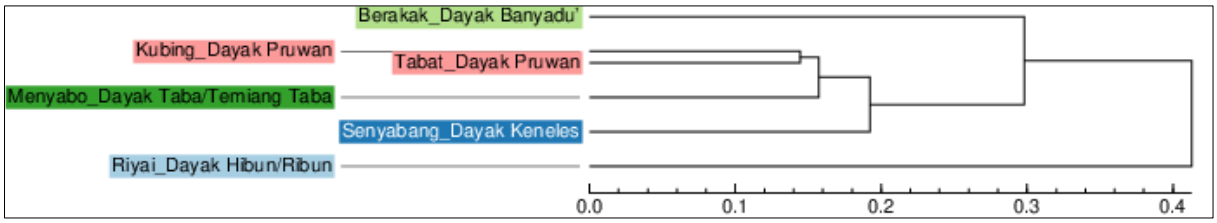

Figure 12. Dendrogram Cluster Analysis Weighted Average Bidayuhic Tayan Hulu

The dendrogram in Figure 11 and Figure 12 also briefly shows that the main subdivisions of the Bidayuhic Tayan Hulu Continuum are divided into two main language clusters. First, the Riyai Dayak Hibun/Ribun language (light sky blue aggregate) as a distinct language continuum. On the other hand, the second most important division is made into a separate group, which is also a combination of the Berakak Dayak Banyadu' variety (light green aggregate) and the Senyabang Dayak Keneles variety (dark blue aggregate), the Menyabo Dayak Taba/Temiang Taba variety (dark blue aggregate), and the Tabat Dayak Pruwan and Kubing Dayak Pruwan varieties (peach aggregate).

The results of the clustering based on string edit distance (Levenshtein distance) as a parameter to observe the difference statistics in Gabmap (Kessler, 1995; Leinonen et al., 2016; Nerbonne et al., 2011) can be seen in Figure 13. Therefore, based on the Gabmap analysis, it is suspected that there are two main linguistic classifications within the Bidayuhic group in Tayan Hulu. As a result, all the quantitative evidence from this Gabmap application supports the findings of several linguists (Asfar, 2014, 2015; Chong & Gedat, 2012; Collins, 2021).

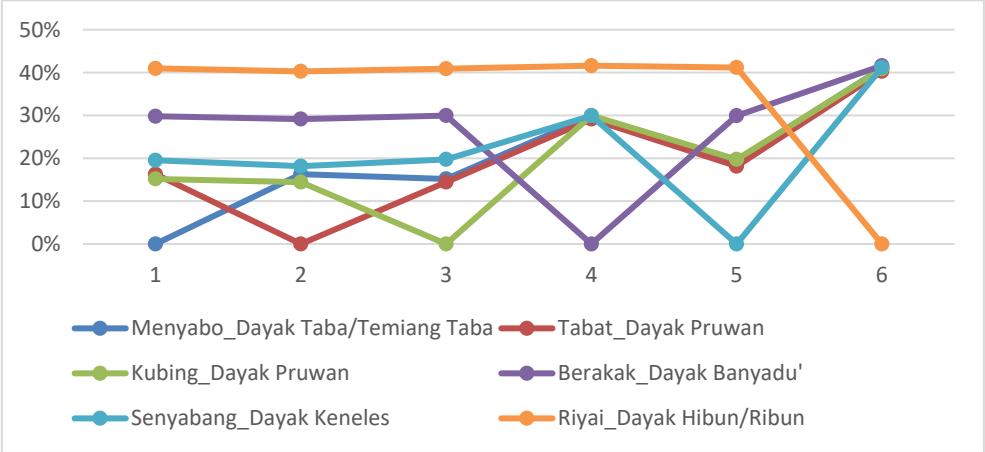

Figure 13. Language Classification in the Bidayuhic Group of Tayan Hulu

Lastly, this algorithmic study also systematically examines which lexical items are responsible for the main geolinguistic differences evidenced in the Tayan Hulu region. In other words, this study objectively identifies which lexical features determine the formation of different aggregates. To do this, both representative quantitative measurements and distinctiveness, provided by Gabmap, were used. The main result of this data mining task points to a set of lexemes that serve as cluster differentiators, which fall into various semantic fields, although slightly higher scores were found in fields related to the human body and expressions of time and space.

IV. Discussion

This research shows that the Bidayuhic language in Tayan Hulu forms a linguistic continuum, meaning that dialectal variations do not have strict boundaries but rather undergo gradual transitions between variations. These findings contribute to modern dialectology, which emphasizes data-driven quantitative analysis in mapping linguistic differences across regions (Markus, 2022; Nerbonne & Kretzschmar, 2013). In the traditional dialectology approach, language variations are often categorized discretely based on specific geographical regions (Kristophson, 2013; Lindström & Pilvik, 2018). However, the results of this study indicate that the approach is not entirely applicable in the case of Bidayuhic.

Lexical and phonological variations in the Bidayuhic language in Tayan Hulu show that dialectal boundaries do not always align with geographical boundaries but instead form a linguistic continuum influenced by social factors and interactions between communities. Evidence from this study shows that several word forms, such as *tuhu*, *turay*, and *tuyay* for "egg" and *kubəs*, *kabis*, and *kobis* for "dead," are spread across various locations without a clear geographical

pattern, thus refuting the assumption that each region has a fixed dialect form. Furthermore, MDS analysis and probabilistic dendrograms show that geographically close locations do not always have high linguistic similarity, while some more distant locations are actually in the same dialectal cluster. Therefore, these results affirm that the differences in Bidayuhic language are more influenced by language contact, social mobility, and cultural interaction rather than by geographical boundaries alone (Chambers, 2015; Mikuleniene, 2013; Nerbonne & Kretzschmar Jr., 2006; Nevaci, 2016).

Furthermore, dialectal variation in Bidayuhic also shows a correlation with geographical factors, as seen in the distribution map of observation points in Figure 1 and Figure 2. The distribution of the Dayak Pruwan, Banyadu', Hibun/Ribun, Taba/Temiang Taba, and Keneles-speaking communities along the Tayan Hulu River shows that intergroup interactions play an important role in the spread of language variation. In modern dialectology studies, methods based on Geographic Information Systems (GIS) and geolinguistic mapping have shown that geographic isolation and social interaction can influence dialect distribution (Chambers, 2015; Dezsö, 2016). The results of this study support these findings by showing how phonological and lexical variations in Bidayuhic can be mapped based on the spatial distribution of its speakers.

In addition to phonological changes, this study also found ablaut phenomena in the Bidayuhic morphological system, which distinguishes between transitive and intransitive verb forms. Figure 5 and Figure 6 show how the word "die" (kubəs, kabis, kobis) undergoes vowel changes in the form "kill" (*ŋubəs*, *ŋamis, kabis*, *ŋkuməs*). This pattern of change is unique because it differs from the morphological strategies in other Austronesian languages, which predominantly use affixation to mark transitivity (Chong & Gedat, 2012; Collins, 2021). By applying the Levenshtein Distance in Gabmap, this research is able to measure these vowel changes objectively, proving that these differences are not merely random variations but part of a stable linguistic system in Bidayuhic.

Further cluster analysis using Gabmap reveals the presence of two main clusters within the Bidayuhic group, as seen in Figure 7 to Figure 12. This clustering shows that Bidayuhic Riyai Hibun/Ribun is a distinct group. Meanwhile, the variants Berakak Banyadu', Keneles, Pruwan, and Taba form another cluster. These findings align with the dialectometry classification model based on Principal Component Analysis (PCA) and Linear Mixed-Effects Modelling, which are used in big data-based dialect mapping (Huisman et al., 2021). Thus, Gabmap not only produces more accurate dialectal mapping but also enables data-driven validation of linguistic hypotheses regarding language variation.

The importance of validation in dialectometry studies is also supported by probabilistic clustering, which in this research was used with an added noise of 0.8 to test the stability of clustering. Figure 9 and Figure 12 show that the clustering results remain stable even when noise is added, which confirms that the dialectal distribution in Bidayuhic is not affected by random factors. This technique has been proven in the studies by Lafkioui (2018) and Nerbonne et al. (2008) as an effective method in addressing linguistic data variability and ensuring the accuracy of dialectal mapping.

The Bidayuhic group would find Gabmap, a type of computational dialectology, quite helpful as they work to plan and bring their language back to life. It could help people learn where different dialects are spoken and make learning resources in their own language. Gabmap is a wonderful way to map out regional dialects in Indonesia, just like Colombian Spanish is. It also recommends that you should use Gabmap with other dialectometry tools, such as ALT-Web, to get a better and more complete view. The results show that Gabmap is a better and fairer way to look at how language changes over time. This helps Indonesia keep track of languages, learn about dialects in different sections of the country, and develop rules (Nevaci, 2016; Bonilla, 2023).

V. Conclusion

The use of Gabmap in dialectometry research has proven effective in visualizing dialectal variations based on lexical and phonological data. This software enables a more objective quantitative analysis by generating visual maps that depict the linguistic distance between language variations. The research results show that this method can identify dialect zones and isogloss distribution patterns more systematically compared to traditional approaches. Thus, Gabmap makes a significant contribution to dialectology studies by offering an efficient and accurate computational-based approach. Although it has advantages in quantitative analysis, Gabmap also has limitations, especially in capturing more complex sociolinguistic aspects, such as the social and cultural influences on language variation. Moreover, the effectiveness of the analysis heavily depends on the quality and completeness of the data used. Therefore, although Gabmap is an innovative tool in linguistic research, this approach is ideally combined with qualitative methods to provide a more holistic understanding of language variation in a region.

References

[1] Asfar, D. A. (2014). Klasifikasi bahasa Dayak Pruwan sebagai bahasa Bidayuhik [Classification of Pruwan Dayak as a Bidayuhik language]. *Kandai*, *10*(2), 138–152. https://doi.org/10.26499/jk.v10i2.318

[2] Asfar, D. A. (2015). *Bahasa Ribun: Refleks fonem Proto-Melayu Polinesia dalam bahasa Ribun* [Ribun language: Polynesian Proto-Malay phoneme reflexes in Ribun languages]. Top Indonesia.

[3] Asfar, D. A. (2016). Kearifan lokal dan ciri kebahasaan teks naratif masyarakat Iban [Local wisdom and linguistic features of Iban narrative texts]. *Litera*, *15*(2), 366-378. https://doi.org/10.21831/ltr.v15i2.11835

[4] Beier, C., & Epps, P. (2020). Reflections on fieldwork: A view from Amazonia. *Language Documentation and Conservation*, *Special Issue*, *(15)*, 321–329.

[5] Bonilla, J. E. (2023). Superdialects, Dialects, and Subdialects of Colombian Spanish. *Lexis (Peru)*, *47*(2), 536–564. https://doi.org/10.18800/lexis.202302.002
[6] Chambers, J. K. (2015). Dialectology. In *International Encyclopedia of the Social & Behavioral Sciences: Second Edition*. https://doi.org/10.1016/B978-0-08-097086-8.52005-4
[7] Chebanne, A. (2016). Writing Khoisan: Harmonized orthographies for development of under-researched and marginalized languages: The case of Cua, Kua, and Tsua dialect continuum of Botswana. *Language Policy*, *15*(3), 277–297. https://doi.org/10.1007/s10993-015-9371-1
[8] Chong, S., & Gedat, R. A. (2012). An introduction to the Austronesian languages in western Borneo. *Language and Linguistics*, *13*(2), 321-349.
[9] Collins, J. T. (2018). The Sekujam language of West Kalimantan (Indonesia). *Wacana*, *19*(2), 425–458. https://doi.org/10.17510/wacana.v19i2.702
[10] Collins, J. T. (2021). *Keberagaman Bahasa dan Etnisitas di Kalimantan Barat* [Language Diversity and Ethnicity in West Kalimantan]. Pontianak: Indonesia Melestarikan Bahasa Ibu.
[11] Coluzzi, P., Riget, P. N., & Wang, X. (2013). Language vitality among the Bidayuh of Sarawak (East Malaysia). *Oceanic Linguistics*, *52*(2), 375–395. https://doi.org/10.1353/ol.2013.0019
[12] Contandriopoulos, D., Sapeha, H., & Larouche, C. (2019). Some insights related to social network analysis data collection challenges–a research note. *International Journal of Social Research Methodology*, *22*(5), 463–468. https://doi.org/10.1080/13645579.2019.1574957
[13] Dezsö, J. (2016). A magyar történeti dialektológia korszakai [Periods of Hungarian historical dialectology]. *Magyar Nyelv*, *112*(1), 17–31. https://doi.org/10.18349/MagyarNyelv.2016.1.17
[14] Dunn, J. (2019). Global Syntactic Variation in Seven Languages: Toward a Computational Dialectology. *Frontiers in Artificial Intelligence*, *2*, 1-22. https://doi.org/10.3389/frai.2019.00015
[15] Effendy, C., Sulissusiawan, A., Syahrani, A., Jupitasari, M., Asfar, D. A., & Lubna, S. (2023). Marine fauna lexicon of Malay community in West Kalimantan. *AIP Conf. Proc. 2913*, 060017. https://doi.org/10.1063/5.0175681
[16] François, A. (2020). In search of island treasures: Language documentation in the Pacific. *Language Documentation and Conservation*, *15*(Special Issue), 276–294.
[17] Francois, S., Wu, K., Doe, E., Tucker, A., & Theall, K. (2023). The influence of racial violence in neighborhoods and schools on the psycho-behavioral outcomes in adolescence. *Research in Human Development*, *20*(1–2), 48-64. https://doi.org/10.1080/15427609.2023.2171694
[18] Huisman, J. L. A., Franco, K., & van Hout, R. (2021). Linking linguistic and geographic distance in four semantic domains: Computational geo-analyses of internal and external factors in a dialect continuum. *Frontiers in Artificial Intelligence*, *4*, 1-19. https://doi.org/10.3389/frai.2021.668035
[19] Irawan, Y., Setiawan, F. A., Asfar, D. A., Irmayani, Herpanus, & Pramulya, M. (2024). Lexical and post-lexical prosodic documentation of Embaloh language. *ILS*, *13*(1), 22–40. https://doi.org/10.33736/ils.6025.2024
[20] Isaías, P., Pífano, S., & Miranda, P. (2012). Subject recommended samples: Snowball sampling. In *Information Systems Research and Exploring Social Artifacts: Approaches and Methodologies* (pp.43-57). https://doi.org/10.4018/978-1-4666-2491-7.ch003
[21] Kehrein, R. (2012). Linguistic Atlases: Empirical Evidence for Dialect Change in the History of Languages. In *The Handbook of Historical Sociolinguistics*. https://doi.org/10.1002/9781118257227.ch26
[22] Kessler, B. (1995). Computational dialectology in Irish Gaelic. In *Proceedings of the Seventh Conference on European Chapter of the Association for Computational Linguistics* (pp. 60-66). Morgan Kaufmann Publishers Inc. https://doi.org/10.3115/976973.976983
[23] Kristophson, J. (2013). Theory of dialect (descriptive). In *Die slavischen Sprachen / The Slavic Languages. Halbband 2* (pp. 2061–2067). Retrieved February 11, 2025, from https://www.scopus.com/inward/record.uri?eid=2-s2.0-85119566486&partnerID=40&md5=8d9cc21b95de7d12077103b5bd33de09
[24] Lafkioui, M. B. (2018). The Rif Berber language continuum: An algorithmic geolinguistic study. In *HuldeAlbum Voor Jacques Van Keymeulen*, *hal-01914354*. Retrieved February 11, 2025, from https://hal.science/hal-01914354/document
[25] Leinonen, T., Çöltekin, Ç., & Nerbonne, J. (2016). Using Gabmap. *Lingua*, *178*, 71–83. https://doi.org/10.1016/j.lingua.2015.02.004
[26] Lendik, L. S., & Yuit, C. M. (2021). A preliminary study on the use of epithets in Kenyah Long Wat. *Journal on Asian Linguistic Anthropology*, *3*(1), 56–75. https://doi.org/10.47298/jala.v3-i1-a3
[27] Lindström, L., & Pilvik, M.-L. (2018). Korpuspõhine kvantitatiivne dialektoloogia [Corpus-based quantitative dialectology]. *Keel ja Kirjandus*, *61*(8–9), 643–662.
[28] Markus, M. (2022). A critical assessment of English dialect feature catalogues: Towards a dialectometrical evaluation of the English Dialect Dictionary Online. *Lingua*, *279*. https://doi.org/10.1016/j.lingua.2022.103428
[29] Mikuleniene, D. (2013). Contemporary linguistic situation in Lithuania: Geolinguistic aspects and new descriptive possibilities. *Acta Baltico-Slavica*, *37*, 459–471. https://doi.org/10.11649/abs.2013.031
[30] Mwelwa, J., & Spencer, B. (2013). A bilingual (Bemba/English) teaching resource: Realising agency from below through teaching materials designed to challenge the hegemony of English. *Language Matters*, *44*(3), 51–68. https://doi.org/10.1080/10228195.2013.840011
[31] Nath, P. K. (2008). *Doing fieldwork on the Singpho language of North Eastern India.* Cambridge University Press. https://doi.org/10.1017/UPO9788175968431.016
[32] Nerbonne, J., Colen, R., Gooskens, C., Kleiweg, P., & Leinonen, T. (2011). Gabmap—A web application for dialectology. *Dialectologia*, *II*(SPEC. ISSUE 2), 65–89. https://raco.cat/index.php/Dialectologia/article/view/245345
[33] Nerbonne, J., Kleiweg, P., Heeringa, W., & Manni, F. (2008). Projecting Dialect Distances to Geography: Bootstrap Clustering vs. Noisy Clustering. In C. Preisach, H. Burkhardt, L. Schmidt-Thieme, & R. Decker (Eds.), *Data Analysis, Machine Learning and Applications* (pp. 647–654). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-540-78246-9_76
[34] Nerbonne, J., & Kretzschmar Jr., W. (2006). Progress in dialectometry: Toward explanation. *Literary and Linguistic Computing*,

21(4), 387–397. https://doi.org/10.1093/llc/fql034
[35] Nerbonne, J., & Kretzschmar, W. A. (2013). Dialectometry. *Literary and Linguistic Computing*, *28*(1), 2–12. https://doi.org/10.1093/llc/fqs062
[36] Nevaci, M. (2016). O cercetare sociolingvistica asupra dialectului aromân [A Sociolinguistic Research on the Aromanian dialect]. *Fonetica si Dialectologie*, *35*, 145–154.
[37] Nguyen, D., & Eisenstein, J. (2017). A Kernel independence test for geographical language variation. *Computational Linguistics*, *43*(3), 567–592. https://doi.org/10.1162/COLI_a_00293
[38] Pröll, S. (2013). Detecting structures in linguistic maps-fuzzy clustering for pattern recognition in geostatistical dialectometry. *Literary and Linguistic Computing*, *28*(1), 108–118. https://doi.org/10.1093/llc/fqs059
[39] Smith, A. D. (2021). The historical phonology of Hliboi, a bidayuh language of Borneo. *Oceanic Linguistics*, *60*(1), 133–159. https://doi.org/10.1353/ol.2021.0004
[40] Spencer, P. T. (2024). Documenting Endangered Languages with LangDoc: A Wordlist-Based System and A Case Study on Moklen. *FieldMatters 2024—3rd Workshop on NLP Applications to Field Linguistics—Proceedings of the Workshop* (pp. 28–36). Retrieved February 11, 2025, from https://www.scopus.com/inward/record.uri?eid=2-s2.0-85204305321&partnerID=40&md5=8b870e068336e500a469c5dc988d2787
[41] Spruit, M. R. (2006). Measuring syntactic variation in Dutch dialects. *Literary and Linguistic Computing*, *21*(4), 493–505. https://doi.org/10.1093/llc/fql043
[42] Sung, H. W. M., Prokić, J., & Chen, Y. (2024). A New Dataset for Tonal and Segmental Dialectometry from the Yue- and Pinghua-Speaking Area. In *SIGTYP 2024—6th Workshop on Research in Computational Linguistic Typology and Multilingual NLP*, *Proceedings of the Workshop* (pp. 25–36). https://www.scopus.com/inward/record.uri?eid=2-s2.0-85189635282&partnerID=40&md5=b0c08abb36e6fba15b92311f5e3b2f82
[43] Wei, W., & Schnell, J. (2025). *The Routledge Handbook of Endangered and Minority Languages*. Routledge. https://doi.org/10.4324/9781003439493
[44] Wieling, M., & Nerbonne, J. (2015). Advances in Dialectometry. *Annual Review of Linguistics*, *1*(1), 243–264. https://doi.org/10.1146/annurev-linguist-030514-124930
[45] Wieling, M., Nerbonne, J., & Baayen, R. H. (2011). Quantitative social dialectology: Explaining linguistic variation geographically and socially. *PLoS ONE*, *6*(9), 1-14. https://doi.org/10.1371/journal.pone.0023613
[46] Wieling, M., Sassolini, E., Cucurullo, S., & Montemagni, S. (2016). ALT explored: Integrating an online dialectometric tool and an online dialect atlas. In *Proceedings of the 10th International Conference on Language Resources and Evaluation, LREC 2016* (pp. 3265–3272). Retrieved February 11, 2025, from https://www.scopus.com/inward/record.uri?eid=2-s2.0-85037117736&partnerID=40&md5=02b3dea69e1081b3cdaeadb1cc95667f
[47] Yumnam, G., & Singh, C. I. (2024). A Bibliometric Perspective of Regional Languages on Select Scholarly Articles. *DESIDOC Journal of Library and Information Technology*, *44*(1), 37–44. https://doi.org/10.14429/djlit.44.1.18938

Dedy Ari Asfar earned a Master of Letters from the National University of Malaysia and also a Master of Indonesian Education at Tanjungpura University, Pontianak, Indonesia. He is currently a lecturer at the Teachers' Training Education Faculty, Tanjungpura University, Pontianak, Indonesia. He has experience and is actively involved in multiple projects in applied linguistics and other linguistics fields as a young expert researcher with interdisciplinary linguistics expertise. He is also the national editor of the Fifth Edition of the Big Indonesian Dictionary (KBBI), published by the Agency for Language Development and Cultivation, Ministry of Education and Culture, Indonesia.

Syarifah Lubna earned both her bachelor's and master's degrees at Tanjungpura University, majoring in language education (English and Indonesian), in 2005 and 2019. She worked as a language researcher at the Ministry of Education and Culture from 2006 until 2022, when she was assigned to the National Research and Innovation Agency.

Irmayani Abdulmalik completed her undergraduate program in 2001 at Tanjungpura University, Pontianak, majoring in Indonesian Education. She earned a master's degree from Gadjah Mada University in 2012. Irma started her career as a researcher at the Language Agency, Ministry of Education and Culture, in 2001, then joined the National Research and Innovation Agency in 2022.

Wiwin Erni Siti Nurlina is a researcher at the Research Center for Language, Literature, and Community, the National Research and Innovation Agency, Indonesia. She completed her studies in Nusantara Literature at Universitas Gadjah Mada in 1987 and her postgraduate program in linguistics at Universitas Gadjah Mada in 1999. She worked as a researcher at the Language Development and Fostering Agency from 1994 to 2021, and now she is a senior researcher at the National Research and Innovation Agency. Her research interests include linguistics, literature, and culture.

Edi Setiyanto completed his bachelor's degree at Gadjah Mada University, Yogyakarta, Faculty of Literature in 1987 and his master's degree from the Master of Linguistics Program at Gadjah Mada University in 2000. He worked as a language researcher at the Ministry of Education and Culture from 1994 to 2021. In 2022, he served as a researcher at the Research Agency for National Innovation.

Sutarsih is a postgraduate in the Language Education Science doctoral program at Semarang State University. She is a senior researcher at the National Research and Innovation Agency, Indonesia. Her research area is language, literature, oral tradition, folklore, and cultural studies.

Yusup Irawan received his master's degree in linguistics at the University of Indonesia in 2022. He has been involved in many collaborative research projects in his institution, the National Research and Innovation Agency, Indonesia. He is the author of three books: (1) Fonetik Akustik: Telaah Wujud Akustik Bahasa, (2) Fonetik & Fonologi Melodi Bahasa: Prosody, and (3) Belajar Membaca Metode Segitiga AIU. His main research interests include theoretical linguistics, especially phonetics and phonology. He is also interested in other research topics, such as culture and communication style.

Binar Kurniasari Febrianti completed her undergraduate education majoring in English Language Education at Semarang State University and her master's at Tanjungpura University, Pontianak. She worked as a civil servant at the West Kalimantan Province Language Center from June 2005 to December 2021. Since January 2022, she has been a researcher at the National Research and Innovation Agency. Additionally, she has published several scientific papers in journals, proceedings, and book anthologies.

Yeni Yulianti completed her undergraduate program in 2005 at Sanata Dharma University, Yogyakarta, majoring in literature. She earned a master's degree from Gadjah Mada University in 2015. Yeni started her career as a researcher at the Language Agency, Ministry of Education and Culture, in 2006, then joined the National Research and Innovation Agency in 2022.

Sarwo Ferdi Wibowo graduated with a master's degree from Gadjahmada University, majoring in literature study, in 2019. He worked as a junior researcher at the National Research and Innovation Agency.

Febyasti Davela Ramadini is a researcher at the National Research and Innovation Agency with expertise in interdisciplinary literature. She is a bachelor's graduate of French Language and Literature from Brawijaya University and earned her Master of Communication Science degree from the University of Indonesia. She has written national and international scientific articles as well as several books related to literature, language, and communication.

Ajeng Rahayu Tjaraka is a researcher at the National Research and Innovation Agency. She obtained her bachelor's degree in German Studies from Universitas Padjadjaran and her master's degree in Literature from Universitas Indonesia. Her research interests are Indonesian literature, German literature, comparative literature, gender studies, language, and cultural studies.

Prima Duantika is a Widyabasa (authorized employee to develop, foster, and protect language and literature) at the West Kalimantan Provincial Language Center. She completed her bachelor's program in Indonesian Language and Literature Education at Tanjungpura University in Pontianak, Indonesia. Her areas of study include modern literature, literacy, and corpus linguistics.