Linguistik Indonesia, Februari 2025, 19-42 *Volume ke-43, No. 1*

ISSN cetak 0215-4846; ISSN online 2580-2429

SUBDIALECT CLASSIFICATION OF MANGGARAI LANGUAGE USING DIALECTOMETRY CALCULATION AND GABMAP

Salahuddin
Universitas Gadjah Mada
salahuddin1998@mail.ugm.ac.id

Abstract

Verheijen (1967) noted that there are 43 sub-dialects of the Manggarai language, which are grouped into five dialect groups: West Manggarai, West-Central Manggarai, Central Manggarai, East Manggarai, and Far East Manggarai. However, the Language Development and Fostering Agency states that the Manggarai language has five dialects, including the Tangge dialect in West Manggarai Regency. Neither study was accompanied by scientific evidence underlying the sub-dialect grouping. Therefore, this research aims to provide quantitative evidence to re-evaluate the variations of the Manggarai language, especially those spoken in West Manggarai Regency. Data was obtained by asking 200 Swadesh vocabularies in ten sample observation areas. The results show that the Manggarai language in West Manggarai Regency is divided into three sub-dialect variations, namely the Kempo sub-dialect (MSdK), the Kolang sub-dialect (MSdS>H), and the Transition area sub-dialect (MSdT). Consistent grouping based on dialectometric calculations and cluster analysis proves that MSdK includes Mbuit, Watu Wangka, Sano Nggoang, Siru, and Benteng Dewa. On the other hand, MSdS>H only consists of Golo Lajang Barat because it shows high linguistic differences from other nearby areas. The MSdT includes Watu Waja, Poco Rutang, Lale, and Tentang. Still, the inclusion of Watu Waja in this group needs to be reconsidered because of the differences in results between dialectometric calculations and cluster analysis.

Keywords: Dialectometry, gabmap, Manggara, subdialect classification

Abstrak

Verheijen (1967) mencatat bahwa ada 43 sub-dialek bahasa Manggarai, yang dikelompokkan menjadi lima kelompok dialek: Manggarai Barat, Manggarai Barat-Tengah, Manggarai Tengah, Manggarai Timur, dan Manggarai Jauh Timur. Namun, Badan Pengembangan dan Pembinaan Bahasa menyatakan bahwa bahasa Manggarai memiliki lima dialek, termasuk dialek Tangge di Kabupaten Manggarai Barat. Tidak ada penelitian yang disertai dengan bukti ilmiah yang mendasari pengelompokan sub-dialek tersebut. Oleh karena itu, penelitian ini bertujuan untuk memberikan bukti kuantitatif untuk mengevaluasi kembali variasi bahasa Manggarai, terutama yang dituturkan di Kabupaten Manggarai Barat. Data diperoleh dengan menanyakan 200 kosakata Swadesh di sepuluh sampel area observasi. Hasilnya menunjukkan bahwa bahasa Manggarai di Kabupaten Manggarai Barat dibagi menjadi tiga variasi sub-dialek, yaitu Kempo (MSdK), Kolang (MSdS>H), dan daerah Transisi (MSdT). Pengelompokan yang konsisten berdasarkan perhitungan dialektometrik dan analisis kluster membuktikan bahwa MSdK mencakup Mbuit, Watu Wangka, Sano Nggoang, Siru, dan Benteng Dewa. Di sisi lain, MSdS>H hanya terdiri dari Golo Lajang Barat karena menunjukkan perbedaan linguistik yang tinggi dari daerah sekitarnya. MSdT mencakup Watu Waja, Poco Rutang, Lale, dan Tentang. Namun, pengecualian Watu Waja dalam kelompok ini perlu dipertimbangkan kembali karena adanya perbedaan hasil antara perhitungan dialektometrik dan analisis kluster.

Kata Kunci: *Dialektometri, gabmap, Manggarai, klasifikasi subdialek*

INTRODUCTION

Languages can take on different variations influenced by various factors such as geographical location, social status, and the profession of the speakers (Parera, 1991: 26). Dialects refer to language variations that arise due to geographical location, which can happen when language speakers reside in an isolated area with no regular interaction with speakers of the same language from other regions. The isolation of a community can be identified by forms of speech that are significantly different from neighboring communities (McDavid, 1946: 170). For instance, the dialect used by people living on the coast will differ from those living in mountainous regions. Meanwhile, the dialect used by homogeneous rural communities will vary from that of divergent urban areas. Sociolects refer to language variations created due to the social strata within language-speaking communities. For example, Balinese society, divided into different strata, tends to use different variations of the Balinese language. McDavid Jr. (1946: 172) explained that there are discernible differences in vocabulary use between British people with social status and those without.

Mapping dialect variations is crucial in determining the diversity of a language. It contributes to various fields of linguistics, such as phonology and syntax. By developing sound maps, researchers can better understand the different dialects present in a language. This massive mapping of variations can help reformulate previously known sound rules, leading to the mutual benefit of dialectology and other branches of linguistics.

In addition to mapping, dialectological studies rely on several other linguistic concepts, including phonemes and allophones in phonology, morph, allomorph, and allomorphic, and morphophonemic in morphology, and phrases, clauses, and morphosyntax in syntax. Furthermore, Swadesh vocabulary and the field of meaning can provide valuable data for the branch of historical-comparative linguistics to reconstruct proto-languages from languages in the archipelago. Comparative historical linguistic studies can influence the results, mainly reconstructed pre-language etymons. The data obtained in language maps can also be used to develop phonological and morphological theories. Indonesian linguistic experts can rely on this data to formulate linguistic rules while overcoming the problem of structural typology, which still depends on Western ideas.

It can be concluded that mapping the variations in a language is necessary to maintain and develop regional languages in Indonesia. Therefore, this research aims to map the dialect variations of the Manggarai language in West Manggarai Regency, East Nusa Tenggara. The study departs from the basic assumption of dialect geography research, which states that the greater the separation between geographical locations, the greater the difficulty in understanding speech in a particular dialect. West Manggarai Regency has hilly to mountainous topography, which allows for differences in language variations between Manggarai speakers in separate areas due to geographical factors. Therefore, exploring the variations in the Manggarai language in the West Manggarai Regency is essential to determine the extent of linguistic differences between regions and the isolect status.

LITERATURE REVIEW

As initial information, Manggarai is spoken in three West Flores districts: West Manggarai, Manggarai, and East Manggarai. Linguistic experts have identified that Manggarai language has several dialects. Burger (1946) categorized these dialects into three large groups: eastern, central,

and western. Verheijen et al. (1995) conducted research and found that there are 43 subdialects in the Manggarai language, which are grouped into five dialects: West Manggarai, West-Central Manggarai, Central Manggarai, East Manggarai, and Far-East Manggarai. The Far-East Manggarai dialect is located in north-central Flores and is separated from other Manggarai dialects by the Rembong language.

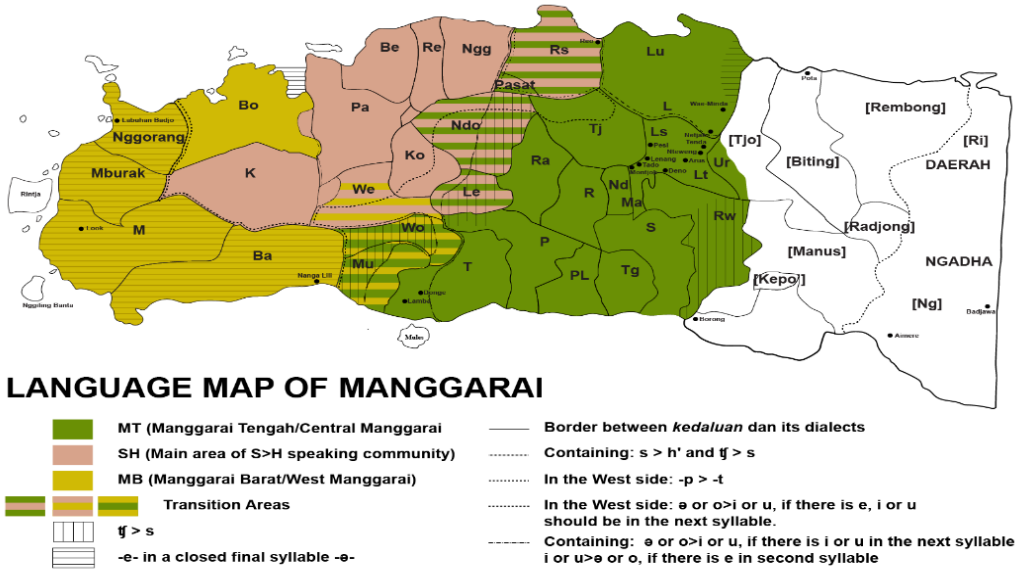

Figure 1. Manggarai Language Map by Verheijen (1967)

On the site page, lexirumah.model-ling.eu details the number of subdialects of the Manggarai language, which are divided based on the location of their speaking by Verheijen (1967). The division of dialects is adjusted to where the Manggarai language is spoken.

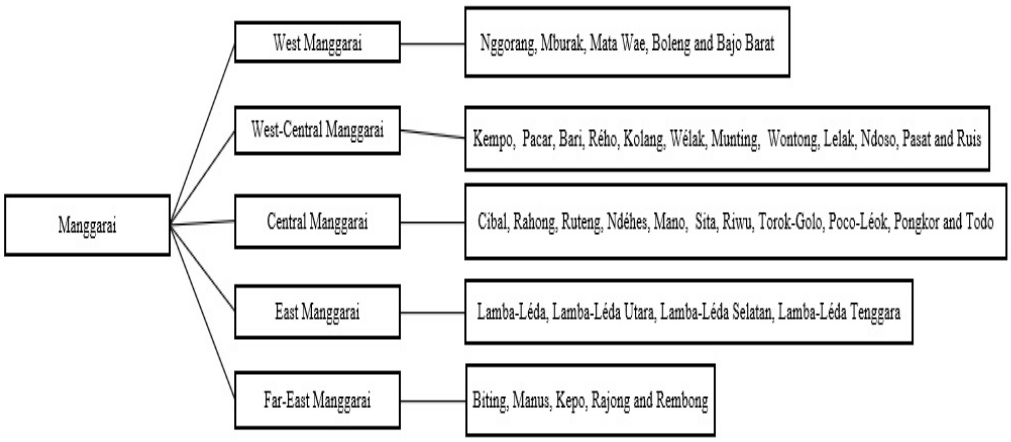

Figure 2. Details of Dialects and Subdialects in the Manggarai Language (Verheijen, 1967)

According to recent Language Development and Fostering Agency data, the Manggarai language has five dialects spread across three districts. These dialects include the Tangge dialect spoken in Tangge Village within the Lembor District of West Manggarai Regency, the Manus dialect spoken in Golo Meni Village and Mukun Village (Pong Bali) within the Kota Komba District of East Manggarai Regency, the Rajong (Kesar) dialect spoken in Mbengan Village

within the Kota Komba District of East Manggarai Regency, as well as in Nanga Meje and Langga Sai Villages within the South Elar District of East Manggarai Regency, the Kepo dialect spoken in Mbengan, Gising (South Elar), Golo Linus, and Sangan Kalo Villages within the South Elar District of East Manggarai Regency, and finally, the Rembong dialect spoken in Sangan Kalo Village, also within the South Elar District of East Manggarai Regency.

Based on the linguistic situation information above, two conclusions can be drawn. First, the Manggarai dialect spoken in the West Manggarai Regency is called the 'West Manggarai Dialect' as stated by several linguistic experts (Burger, 1946; Verheijen, 1967; Grimes, 1997). Second, the West Manggarai dialect has several subdialects based on the location where they are spoken, as seen on the website page lexirumah.model-ling.eu and the Language Development and Fostering Agency. However, the lack of uniformity regarding reports on the types and number of subdialects in West Manggarai Regency is of concern in this research. For example, the Language Development and Fostering Agency reports that there is only one dialect in West Manggarai Regency, namely the Tangge subdialect. On the other hand, this report contradicts the subdialect report by Verheijen (on the site https://lexirumah.model-ling.eu/), which does not include the Tangge subdialect as one of the subdialects in West Manggarai Regency.

To accurately map the distribution of subdialects, dialectology researchers in Indonesia need suitable and easily accessible software. This study employs dialectometric calculations and gabmap software to map subdialects of the Manggarai language based on differences in linguistic features found in various survey sites. Gabmap is a platform that facilitates dialectometric and cartographic calculations. It enables researchers to create maps and perform statistical analysis of dialect data by gathering variants of the pronunciation of many words in different locations within an area. Researchers can compare the collected terms and generate a dialect map using gabmap.

No other research has addressed the variations of the Manggarai dialect since Verheijen's study in 1967. However, it is important to highlight two thesis studies that explore this dialect's variation in West Manggarai Regency: Baru's (2022) thesis titled "Variation of the Manggarai Dialect of East Nusa Tenggara" and Ridwan's (2019) thesis called "Variation of the Manggarai Dialect: A Diachronic Dialectology Study." Both studies focus on describing the phonological and lexical differences among the dialects across the three regencies, but they do not offer comprehensive information on dialectometric calculations or the development of dialect distribution maps.

METHOD

Data Collection

The survey sites are determined by considering the differences in the use of isolect (according to administrative criteria). Determining the survey sites at an early stage is essential because it directly impacts the selection of informants and wordlists. The survey sites were decided qualitatively by considering several things: the survey site is a rural area/far from a big city/isolated area with low mobility (Ayatrohaedi, 1978) with a maximum population is 6000 people, and should have at least 30 years old rural areas. According to the law of spatial autocorrelation, closer locations are expected to show similar variations (Jeszenszky et al., 2021: 3), or in other words, dialects/subdialects that are geographically far apart are less identical than adjacent dialects (Chambers et al., 2004). The selection of the survey sites must be based on findings that indicate differences in linguistic variation in the area by involving data samplings.

At this stage, the term 'planning the grid' is known, which aims to identify the geographical area to be investigated and decide where to collect information (Boberg et al., 2018: 241). This study aimed to identify and gather evidence of speakers of traditional dialects who were less influenced by educationally promoted standard forms of language so that the target area was smaller rural communities.

The sampling method is required to find out whether an area can be representative of other regions to investigate the indication of different dialect variations. The research is conducted in ten representative survey sites in the West Manggarai district, such as Sano Nggoang in Sano Nggoang District, Watu Wangka in Mbeliling District, Mbuit in Boleng District, Poco Rutang and Siru in Lembor District, Benteng Dewa and Watu Waja in South Lembor, Golo Lajang Barat in Pacar, Lale in Welak District, and Tentang Village in Ndoso. These ten villages meet the criteria for survey sites suggested by the experts above, especially for village age and geographical location.

The required informants or native speakers must meet the NORMs criteria (nonmobile, older, rural male) proposed by Chambers et al. (2004: 29). The term 'native speaker' implies that speakers are a valuable source of information and hold the key to the target language's structure in their linguistic competence (Chelliah et al., 2010). Native speaker can take a different role in linguistic fieldwork project. However, considering their task as someone who provides information about the language spoken in the area being observed, in this study, I refer to them as consultant or informant interchangeably. Informant is defined as an articulated individual in the language community who provides crucial information about the language or culture being studied and can build a long-term relationship with the investigator. At the same time, the term ‘informant’ itself characterized native speakers as a machine that works to provide linguistics data only. Thus, Other fieldworkers prefer using the term ‘consultant’ to refer the native speaker as having active roles in the fieldwork situation (Chelliah et al., 2010). We chose one informant for each survey site according to the above criteria. The informant criteria are at least the age of a middle-aged man or woman, being native to the community (Boberg et al., 2018: 242), farmers/laboring, being physically and mentally healthy (Mahsun, 2004: 106), and have not yet entered a senile age (Lauder, 1993). These informants included MK (53) in Watu Wangka, TM (52) in Lale, LM (50) in Watu Waja, DU (54) in Mbuit, SH (55) in Tentang, RA (48) in Siru, GH (59) in West Golo Lajang, YB (54) in Sano Nggoang, SN (51) in Poco Rutang, AK (54) in Benteng Dewa.

Determining the criteria for survey sites and informants also directly impacted the type of wordlists used when surveying dialects. Additional data collection is needed to complete various linguistic information regarding the dialects of the languages studied (Fernandez, 1993: 24). Therefore, questions related to culture are also used to see the differences in Manggarai language variation between survey sites. Researchers should find in-depth information about the language and history of the research area as a reference in determining what kind of questions to prepare (Mahsun, 2004: 107). Information about the language and history of the research area can be obtained through previous research, embodied in a dictionary or history book (Fernandez, 1993: 25; Mahsun, 2004: 107). In this study, the information contained in the Manggarai-Indonesian Dictionary compiled by Verheijen (1967) helps to determine the type of wordlists used in this study. Thus, the words investigated were based on the Swadesh list, which consists of 200 words with slight adjustments according to the criteria that have been mentioned.

Data Analysis

This study employs two quantitative methods to determine the degree of differences in isolects between observation areas. It concludes with a geographical classification of variations in the Manggarai language. In dialectology, dialectometry is the commonly used quantitative method for measuring variations in a particular language (Nadra et al., 2009). Séguy (1973) introduced the term "dialectometry" in the *Atlas Linguistique de la Gascogne*. It is derived from the word 'dialectometry' which means 'the measure of dialect' or dialect measurement (Heeringa et al., 2001: 4). Dialectometry is a method of calculating dialect differences, which are linguistic differences determined mainly by geographic location (Nerbonne et al., 2023: 145). The formula for the dialectometric method is as follows.

$$\frac{\boldsymbol{s.100}}{\boldsymbol{n}} = \boldsymbol{d}\%$$

The amount of data that shows the differences between the areas being compared is represented by the letter **s**. The letter **n** refers to the total amount of data with phonologically or lexical differences, and the letter **d** denotes the percentage of linguistic differences between the areas. Based on these values, if the difference is more significant than 81%, the two regions are considered to have different languages. They are deemed to have dialect differences if the difference is between 51% and 80%. If the difference is between 31% and 51%, they are considered to have subdialect differences. The difference between 21% and 30% is considered speech difference, while less than 20% is regarded as no difference.

Séguy (1973) looked for an objective way that would make it possible to analyze maps without having to resort to traditional methods. Then, Séguy introduced a new concept that could track location points that showed different dialect variations and record them in large quantities. The contrast between the two locations is described in terms of percentages, which indicate linguistic evidence between the two locations. Nerbonne (in Nerbonne, 2010) explains that dialectometry distils aggregate relationships from a set of linguistic correspondences by systematically comparing large sets of corresponding linguistic data and measuring the differences. Different (not identical) forms contribute to the linguistic distance between areas of observation.

Dialectometry adopts a spatially oriented approach similar to wave theory introduced by Wellentheorie and gravity theory in studying language change and enriches it with a quantitative perspective (Koile et al., 2022). Compared to geolinguists and dialectometricians try to infer patterns of spatial variation from large data sets based on aggregate differences rather than from analysis of individual features. Dialectometry works at a high level of spatial detail and deals with closely related linguistic varieties.

Dialectometric calculations rely on phonological and lexical differences. Phonological differences include variations in sound and phonemes, while lexical differences involve variations in vocabulary. Research in syntax and semantics has shown minimal dialectal or sub-dialectal variation, and therefore, these two fields are often not considered in dialectological studies.

Studies in dialectology have shown that language variation is complex both linguistically and geographically and cannot be simplified (Nerbonne et al., 2023: 245). Computational linguistics offers techniques that can handle large amounts of data and be used to analyze them. One of the ways this can be done is by using the Gabmap application, which applies Levenshtein calculations to calculate the linguistic distance between observation points. The introduction of

the edit distance (or Levenshtein distance) in dialectometry was an essential innovation, first used by Kessler in 1995 (in Wieling et al., 2015). Heeringa's (2004) dissertation dealt with edit distances in dialectometry, enabling the analysis of large amounts of data collected while compiling dialect atlases without manually characterizing and classifying them. Therefore, computational methods are valuable in dialectometry since they efficiently analyze large amounts of data.

The Levenshtein distance is a numerical value that measures the number of insertions, deletions, or substitutions required to transform one string into another. This technique involves comparing rows of fonts, with all calculations having the same value, usually 1. For example, when two fonts are the same but have different diacritics, they are considered different and given a value of 1. Similarly, [a] and [p] are different and also receive a value of 1. Kessler (1995) uses this method to calculate the Levenshtein distance for phonetic variants of words and lexical differences. The Levenshtein distance is a highly objective measure, and its results can be proven, making it suitable for computational methods. However, it is essential to note that the data used for applying the Levenshtein distance must consist of samples from examples of variation.

From the results of these calculations, gabmap automatically displays a dialect map that illustrates the linguistic distance between dialects. Gabmap can generate maps that show the distribution of specific types, such as particular words or fonts. To locate a specific variant, you can select a variable (word) and then a particular variant (pronunciation). It is important to note that white areas on the map indicate no examples were found in that area. At the same time, darker colors represent a higher relative frequency of that character in the data at that location (Leinonen et al., 2016).

Maps play a vital role in studying geographic dialectology, and their position is determined by the research carried out, according to Ayatrohaedi (1979: 30). In this era of computational advancements, dialectology research can be made easier through the use of specialized software that supports automatic analysis and presentation of data. The automated analysis increases the efficiency of researchers and the replicability of analyses. It allows them to focus on more abstract aspects of research and conclusions drawn from studies (Leinonen et al., 2016).

This study utilizes gabmap software for dialectometric calculations and isolect mapping in West Manggarai Regency. Gabmap is a dialect classification and visualization software developed by linguists at the University of Groningen (Leinonen et al., 2016). The software provides various statistical methods to group similar languages (Feleke et al., 2020). Gabmap can summarize data, calculate linguistic distances, perform statistical analysis and create maps. Researchers can get a data summary by uploading data on the site, which includes the number of places, linguistic variables, and transcription symbols used. The software creates an index map of the data locations, and researchers can also create a distribution map of specific gloss variants. Some analyses in gabmap software are based on the linguistic distance between language variants, dialects or subdialects. The edit distance series measures linguistic distance, displayed on some maps. All maps and images in gabmap can be downloaded as image files. With statistical tools in gabmap, researchers can explore the structure of dialect data. Multidimensional scaling can examine the extent to which dialects form a continuum, whereas cluster analysis classifies dialects and detects dialect areas. The results of statistical analysis are displayed in maps and figures.

Cluster analysis is a technique to group dialects based on their similarities and identify dialect regions. Clustering involves dividing a set of objects into different groups or clusters. In

the case of Gabmap, the things are geographic places grouped based on the similarities in the language spoken there. The statistical analysis and maps presented on Gabmap go through several processing stages. The dialect data is collected from several observation areas through phonetic transcription, and the geographic data is obtained from Google Earth and uploaded to the Gabmap website (https://gabmap.nl/doc/manual/). Once the data is uploaded, Gabmap summarizes the inspection, including an index map, data overview and distribution map of each variation for a particular gloss.

The processing type chosen for the dialect data is string data, as it is in phonetic transcription. String edit distance, also known as Levenshtein distance, is used when the dialect data consists of transcriptions of lexical items.

$$\mathrm{I}_l \quad = \quad \frac{1}{n}\sum_{i=1}^{n}\frac{D_i^L - D_i^G}{D_i^G}$$

Where *n* is the number of observation area locations, D_i^L is linguistic distance, and D_i^G is geographic distance. These two distances can be obtained in detail using the formula:

$$D_i^L \quad = \quad \sum_{j=1}^{k} d_{i,j}^L \cdot 2^{-0.5j}$$

$$D_i^G \quad = \quad \sum_{j-1}^{k} d_{i,j}^G \cdot 2^{-0.5j}$$

This formula generates maps and stats based on input distance.

FINDINGS AND DISCUSSION

To create maps of observation areas, we used Google Earth and uploaded research data files encoded in Unicode text on the gabmap page https://gabmap.let.rug.nl/. All the data uploaded on the page will be automatically generated, summarizing linguistic distances between observation areas. For dialectometry calculations using the triangle permutation method, the names of the observation areas will be encoded into a numbered index map. The observation area code follows a number coding pattern generated by gabmap, which is used to call the name of the following observation area.

Table 1. Generalized observation region automation sequence by Gabmap

Indexed Map Number	Observation Area
1	Mbuit
2	Watu Waja
3	Benteng Dewa
4	Lale
5	Poco Rutang
6	Siru
7	Golo Lajang Barat
8	Sano Nggoang
9	Watu Wangka
10	Tentang

Dialectometric Calculations using the Triangle Permutation Method

Calculations involving permutations between observation areas are done by computing the linguistic distance from one observation area to another (Lauder, 1993: 142). The main objective of permutation analysis is to determine the level of mutual understanding of observations between different regions. As noted by Voegelin and Haris (in Mahsun, 2004: 147), the degree of understanding is directly proportional to the spatial distance between observation areas.

Calculations with triangles between observation areas are performed under the following conditions: first, only observation areas that are located in direct communication with each other are compared; second, each observation area that communicates directly is connected by a line, resulting in triangles with various shapes (i.e., isogloss lines); and third, the lines in the dialectometric triangle must not intersect. Only one possibility should be chosen, and selecting those closer to each other is better.

Following certain principles when using dialectometry with triangles between observation areas is essential. Firstly, if a meaning has multiple variations and one is known in another area being compared, then the difference is not considered to exist. Secondly, if one of the observation areas being compared does not have a particular form of meaning realization, it is considered different. Thirdly, if the observation areas being compared do not share a specific form of meaning realization, it is believed that there is no difference. Fourthly, phonological and morphological differences are set aside when calculating dialectometric differences at the lexicon level. Lastly, calculation results are mapped using the "polygons de Thiessen" construction system on a triangular dialectometric map, where all points are triangulated into an irregular network.

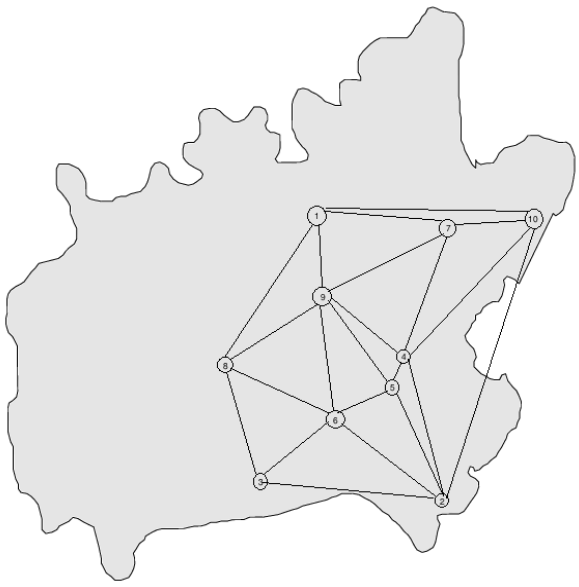

Figure 3. Permutation of Triangles Between Observation Areas

Automated calculation of permutation values between observation areas involves using Python scripts. The first requirement for this process is a database that contains all observation areas and their respective variations of 200 words each (see Appendix 1). This database is used to compare the vocabulary between observation areas. Sites with the same variation are assigned a value of 0, while those with different variations receive a value of 1. However, by the

requirements of dialectometric triangular permutation calculations, if a location has more than one variation, one of the variations is assumed to be non-existent if it appears in another observation area. This is achieved through coding logic similar to that used in Python automation (see Appendix 2). All comparison results are then summarized in a separate table, which shows the comparison of all words between two observation areas in the permutation table (see Appendix 3). We can determine the phonological and lexical differences in each directly adjacent observation area using these calculations.

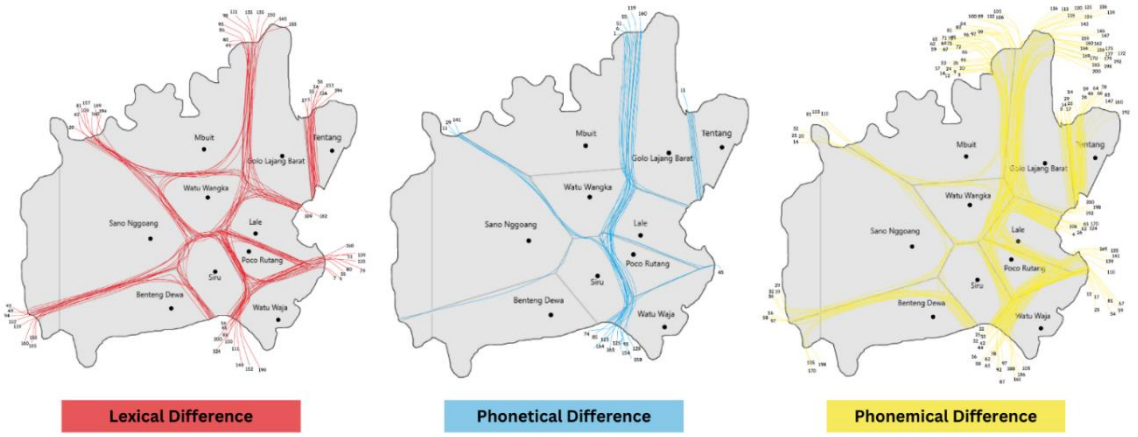

Figure 4. Compilation of Isogloss Lines for Lexical and Phonological Differences

Compiling isogloss line maps reveals that lexical differences are evenly distributed throughout the observation area. However, phonemic contrasts appear to be more frequent than phonetic differences when it comes to phonological differences. Nonetheless, phonological differences are more dominant on a larger scale than lexical differences. This conclusion is supported by the isogloss map displayed below.

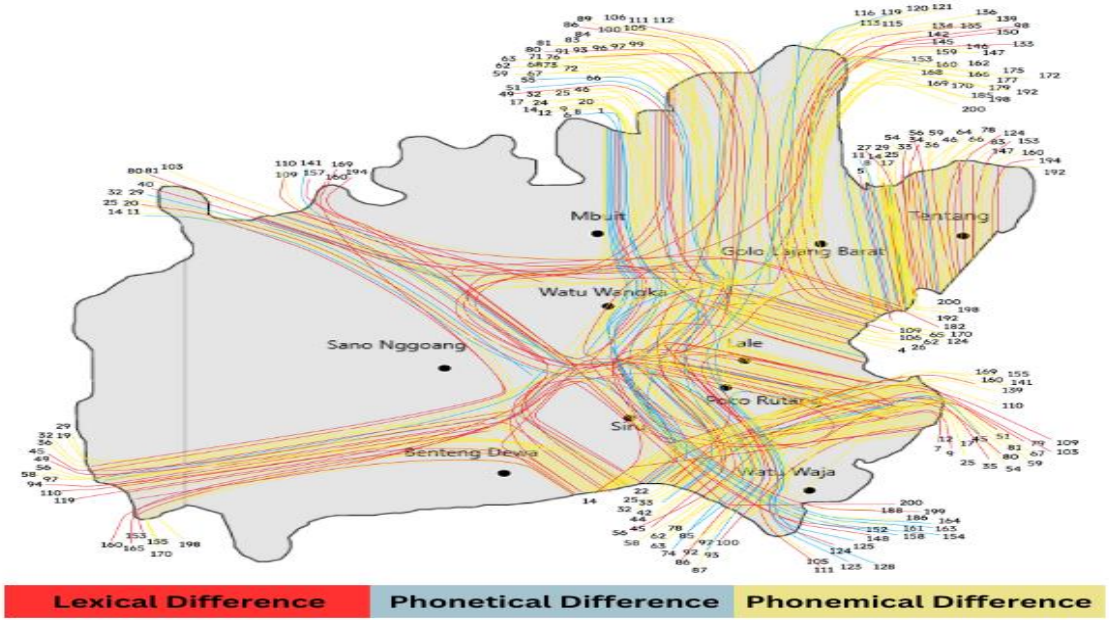

Figure 5. Isogloss File Compilation

The first step is to perform a dialectometric calculation using the formula proposed by Séguy (1973) and redefined by Goebl (1984).

$$\frac{s.100\%}{n}$$

The method involves comparing multiple maps and counting the number of differences (s) between observation areas. Then, the percentage vocabulary distance (d) is calculated. Based on these results, the linguistic differences between observation areas are determined using the following criteria: a percentage distance above 81% is considered a language difference, 51-80% is regarded as a dialect difference, 31-50% is considered a subdialect difference, 21-30% is regarded as a speech difference, and below 20% is considered as no difference.

Table 2. Permutations Between Nearby Observation Areas

Compared OA	Number of Phonological Differences	Number of Lexical Differences	Total
1-7	79	24	103
1-8	22	15	37
1-9	11	12	23
1-10	43	20	63
2-3	76	16	92
2-4	36	17	53
2-5	49	21	70
2-6	76	26	102
2-10	60	25	85
3-6	13	18	31
3-8	21	16	37
4-5	32	18	50
4-7	52	18	70
4-9	66	14	80
4-10	46	15	61
5-6	43	29	72
5-9	45	26	71
6-8	9	16	25
6-9	16	14	30
7-9	83	21	104
7-10	73	25	98
8-9	18	13	31

The table indicates that areas 7-9 exhibit the greatest phonological and lexical differences, followed by locations 7-1, 2-6, and 7-10. On the other hand, areas 1-9 display a relatively low level of lexical and phonological differences, suggesting that these two areas are part of the same group.

Dialectometric Calculations of Phonological Differences

The next step is to measure the level of difference by applying the dialectometric formula proposed by Séguy (1973). All findings indicating lexical differences must be ignored when calculating phonological differences, or vice versa.

Table 3. Interregional Relationships Based on the Phonological Differences Found

Compared OA	Number of Phonological differences (*s*)	100/*n*	Total (%)	Classification
1-7	79	100/135	58,5	Dialect
1-8	22		16,3	No difference
1-9	11		8,1	No difference
1-10	43		31,9	Subdialect
2-3	76		56,3	Dialect
2-4	36		26,7	Speech difference
2-5	49		36,3	Subdialect
2-6	73		54,1	Dialect
2-10	60		44,4	Subdialect
3-6	13		9,6	No difference
3-8	21		15,6	No difference
4-5	32		23,7	Speech difference
4-7	52		38,5	Subdialect
4-9	66		48,9	Subdialect
4-10	46		34,1	Subdialect
5-6	43		31,9	Subdialect
5-9	45		33,3	Subdialect
6-8	9		6,7	No difference
6-9	16		11,9	No difference
7-9	73		54,1	Dialect
7-10	73		54,1	Dialect
8-9	18		13,3	No difference

The table above displays dialectometric calculations for phonological differences between the observation areas. The following results can be summarized:

1. The observation areas with different dialects are 1-7, 2-3, 2-6, 7-9, and 7-10.
2. The observation areas with two different subdialects are 1-10, 2-5, 2-10, 4-7, 4-9, 4-10, 5-6, and 5-9.
3. Speech differences are observed in areas 2-4 and 4-5.
4. The areas that do not show differences are 1-8, 1-9, 3-6, 3-8, 6-8, 6-9, and 8-9.

These conclusions can be visualized in the map below.

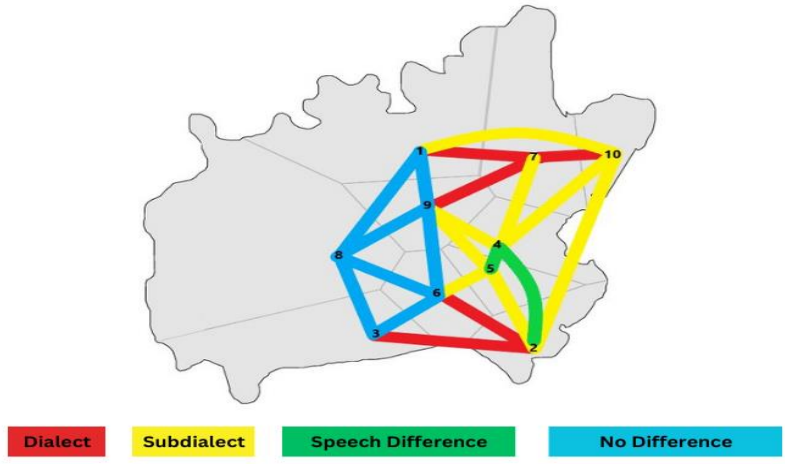

Figure 6. Isogloss Lines for Calculation of Phonological Distinctions

Based on the map, it can be inferred that the people in area 7 (Golo Lajang Barat) speak a different dialect compared to those in areas 1 (Mbuit), 10 (Tentang), and 9 (Watu Wangka). Similarly,

the people in area 2 (Watu Waja) use a different dialect than those in areas 3 (Fortress of God) and 6 (Siru). On the other hand, areas 2 (Watu Waja), 10 (Tentang), and 5 (Poco Rutang) only exhibit subdialect differences. The same goes for area 5 (Poco Rutang) with areas 6 (Siru) and 9 (Watu Wangka), areas 9 (Watu Wangka) with area 4 (Lale), area 4 (Lale) with areas 7 (Golo Lajang Barat) and 10 (Tentang), and area 10 (Tentang) with area 1 (Mbuit). However, it's important to note that area 2 (Watu Waja) only shows differences in speech with area 4 (Lale) and area 4 (Lale) with area 5 (Poco Rutang). Meanwhile, areas 1 (Mbuit), 9 (Watu Wangka), 6 (Siru), 8 (Sano Nggoang), and 3 (Fortress of God) did not display any differences at all.

Dialectometric Calculations of Lexical Differences

It is also essential to consider lexical differences to determine whether they affect the level of understanding between observation areas. Table 17 shows the lexical differences between observation areas, which were determined using triangular permutations based on phonological differences. A total of sixty glosses indicated the lexical variations of the 200 words tested. To calculate the lexical differences in each permutation area, the number of differences was multiplied by 100% and divided by the 48 glosses with that particular lexical variation.

The results show that the permutation areas with the most lexical differences are 5-6 (Siru-Poco Rutang), followed by areas 2-6 (Watu Waja-Poco Rutang) and 5-9 (Siru-Watu Wangka). On the other hand, areas 1-9 (Mbuit-Watu Wangka) and 8-9 (Sano Nggoang-Watu Wangka) have the least number of lexical differences.

Table 4. Interregional Relationships Based on the Lexical Differences Found

Compared OA	Number of Lexical differences (*s*)	100/*n*	Total (%)	Classification
1-7	24	100/135	17,8	No difference
1-8	15		11,1	No difference
1-9	12		8,9	No difference
1-10	20		14,8	No difference
2-3	16		11,9	No difference
2-4	17		12,6	No difference
2-5	21		15,6	No difference
2-6	26		19,3	No difference
2-10	25		18,5	No difference
3-6	18		13,3	No difference
3-8	16		11,9	No difference
4-5	18		13,3	No difference
4-7	18		13,3	No difference
4-9	14		10,4	No difference
4-10	15		11,1	No difference
5-6	29		21,5	Speech difference
5-9	26		19,3	No difference
6-8	16		11,9	No difference
6-9	14		10,4	No difference
7-9	21		15,6	No difference
7-10	25		18,5	No difference
8-9	13		9,6	No difference

Linguistik Indonesia, Februari 2025, 19-42 *Volume ke-43, No. 1*
Copyright©2025, Masyarakat Linguistik Indonesia
ISSN cetak 0215-4846; ISSN online 2580-2429

Based on the results of the dialectometry method, only two areas, namely area 5 (Poco Rutang) and area 6 (Siru) in the Lembor District, exhibit notable differences in speech. On the other hand, the comparison of other observation areas indicates no significant differences. To simplify, the map below illustrates the mentioned areas.

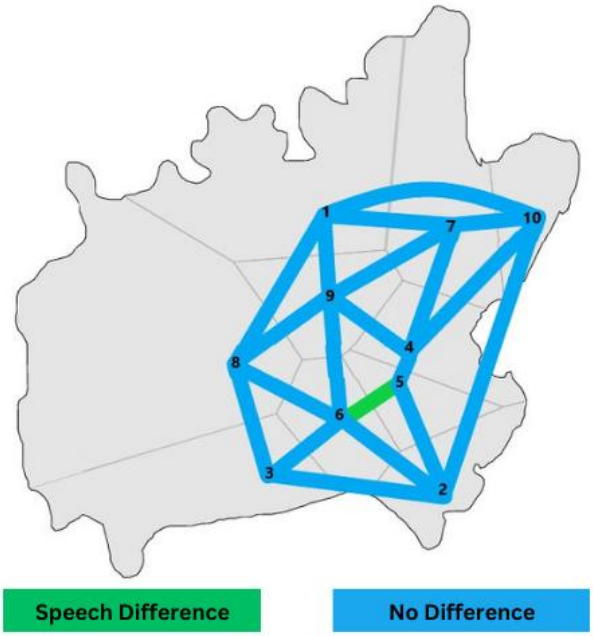

Figure 7. Isogloss Line for Lexical Distinction Computation

The map above indicates that lexical differences between observation areas have little influence on using the Manggarai language in the ten areas.

Supporting Evidence for Language Classification by Gabmap

Clustering Analysis

Clustering is a technique that involves dividing a set of objects, such as geographic regions, into distinct groups based on their similarity levels (Everitt, 2011). The primary purpose of clustering is to identify data patterns by grouping objects with similar traits and characteristics while highlighting differences between groups. Each subgroup in the clustering analysis has a unique set of features and is often closely related based on specific distance metrics (Prokić, 2010: 25). The findings from the regional grouping and calculations using the dialectometry method were compared with the results of this clustering analysis. This research employs the agglomerative clustering method, in which every object is placed in a separate cluster and progressively grouped into larger clusters until a single set is obtained.

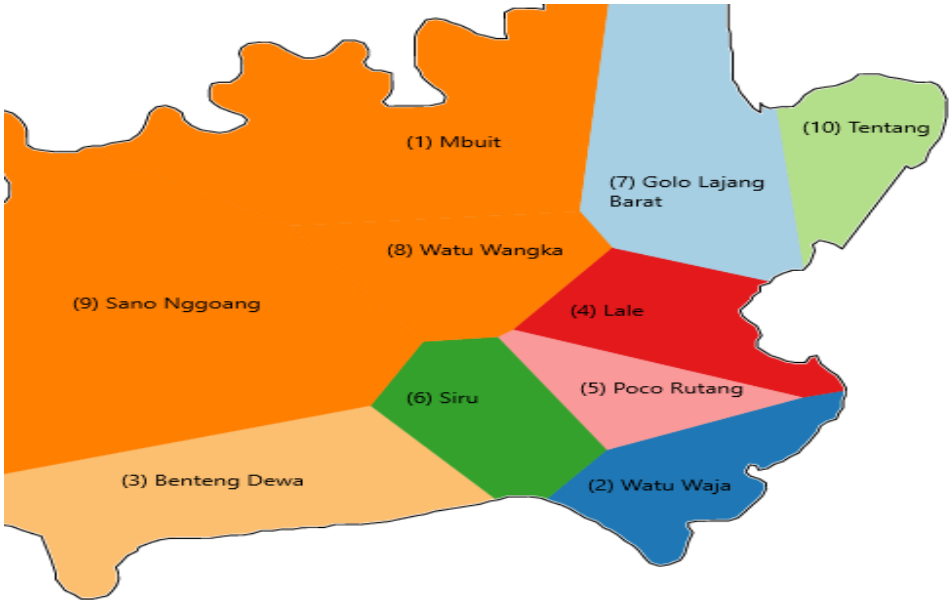

Figure 8. Clustering Analysis

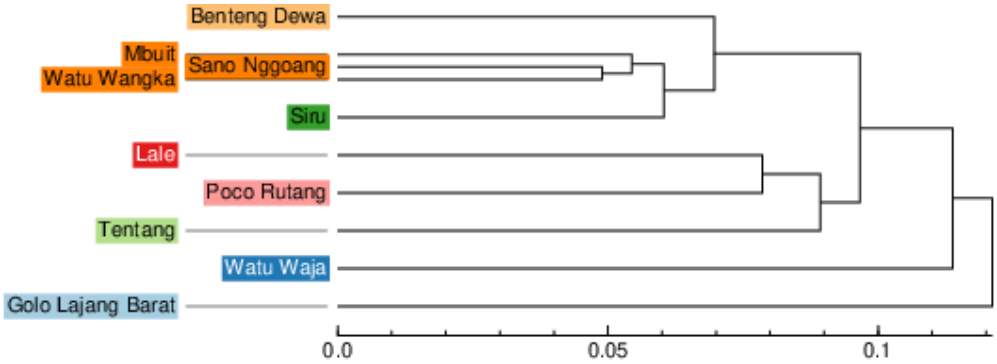

Figure 9. Dendrogram for Clustering Analysis Results

The process involves combining two or more places with the least linguistic difference into a single group. It's important to note that colour similarity does not indicate linguistic similarity, and each colour symbolizes only one group. When creating a dialect classification, start from the right and move towards the left on the dendrogram until you identify a breakpoint with the desired number of branches for the groups in the classification. The clustering map and dendrogram can be understood as follows.

Table 5. Language classification of Manggarai language based on hierarchical clustering

Group	Subgroup	Survey Sites
Group A	Subgroup A_1	Sano Nggoang + Watu Wangka
	Subgroup A_2	Subgroup A_1 + Mbuit
	Subgroup k A_3	Kelompok A_2 + Siru
	Subgroup A_4	Subgroup A_3 + Benteng Dewa
Group B	Subgroup B_1	Lale + Poco Rutang
	Subgroup B_2	Subgroup B_1 + Tentang
Group C	-	Group A + B
Group D	-	Group C + Watu Waja
Group E	-	Group D + Golo Lajang Barat

Group A1 composed of Sano Nggoang and Watu Wangka, and Group B1, comprising Lale and Poco Rutang, are two villages grouped due to their high similarity or low difference level, merging them into one group. These two groups become the same at a higher level when Mbuit is included as a group member. In this higher classification, Golo Lajang Barat stands out as the survey site with the most significant linguistic distance compared to other areas.

This classification highlights a difference between traditional grouping using dialectometric triangular permutations and the computational method produced by Gabmap. While standard calculations result in an MSdK group consisting of Sano Nggoang, Watu Wangka, Mbuit, Siru, and Benteng Dewa, Gabmap only groups Mbuit, Watu Wangka, and Sano Nggoang into one group and separates Siru and Benteng Dewa into a higher group. Similarly, with transitional regional groups, Gabmap only groups Lale and Poco Rutang at a lower level due to their higher level of similarity and excludes Tentang and Watu Waja at a higher level. Golo is isolated but at the top of the hierarchy, indicating a high linguistic difference from other variations, followed by Watu Waja at a lower level. This suggests that geographical distance does not significantly influence the level of linguistic differences in language variation between survey sites.

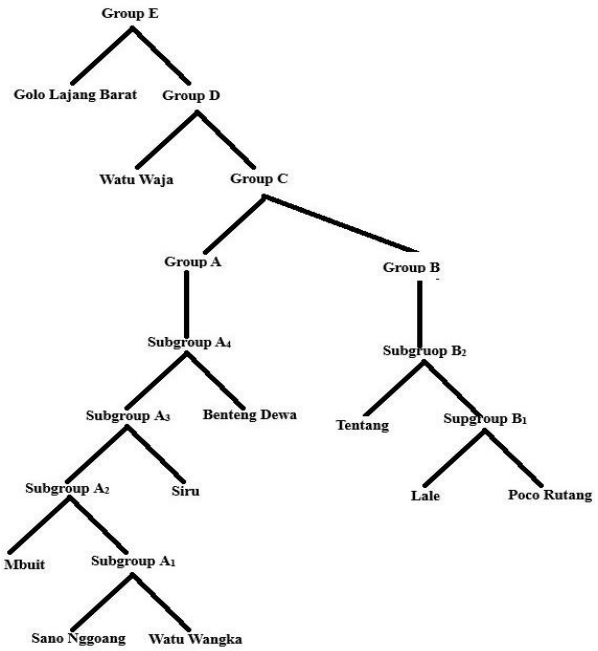

Figure 10. Hierarchical Clustering (Bottom-Up Clustering)

In this classification, two groups, namely group A1, consisting of Sano Nggoang and Watu Wangka, and group B1, consisting of Lale and Poco Rutang, are joined together because they share a high level of similarity or a low level of difference. When Mbuit is included, these two groups become the same at a higher level. Additionally, West Golo Lajang is the observation area with the highest linguistic distance compared to other sites in a higher classification.

It is important to note that this classification differs from the traditional grouping method, which uses dialectometric triangular permutations. Traditional calculations consistently group Sano Nggoang, Watu Wangka, Mbuit, Siru, and Benteng Dewa. However, Gabmap, a

computational method, only groups Mbuit, Watu Wangka, and Sano Nggoang into one group while separating Siru and Benteng Dewa into a higher group. Regional groups in transition areas are treated similarly. Gabmap only groups Lale and Poco Rutang at a lower level because they have a higher level of similarity and exclude Tentang and Watu Waja at a higher level.

Furthermore, Golo Lajang Barat is at the top of the hierarchy and is considered a variation with a high level of difference from other variations, followed by Watu Waja at a lower level. This indicates that geographical distance does not influence the level of differences in language variation between observation areas.

Fuzzy Clustering Analysis

The grouping results are consistent with those derived from the clustering analysis. In fuzzy clustering, random noise is added to the original distance matrix multiple times. Clustering is then performed on the altered matrix, and the frequency of occurrence of each cluster is determined. The most reliable groupings frequently appear in multiple runs with additional noise.

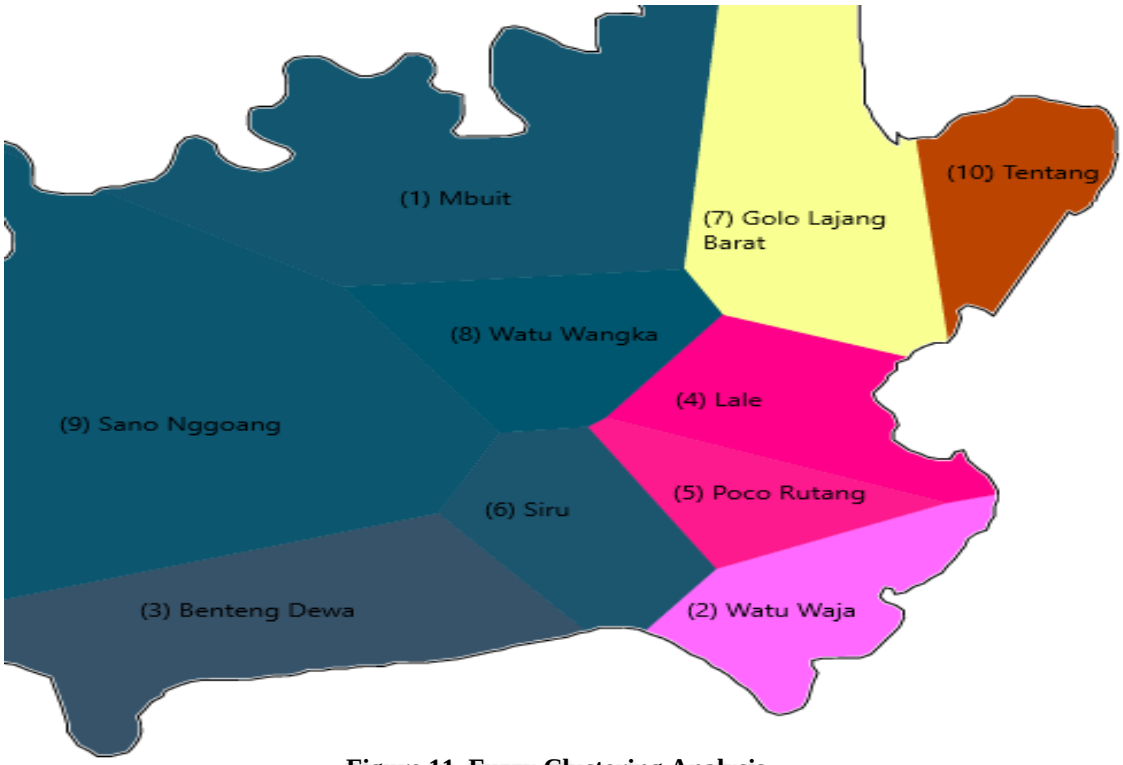

Figure 11. Fuzzy Clustering Analysis

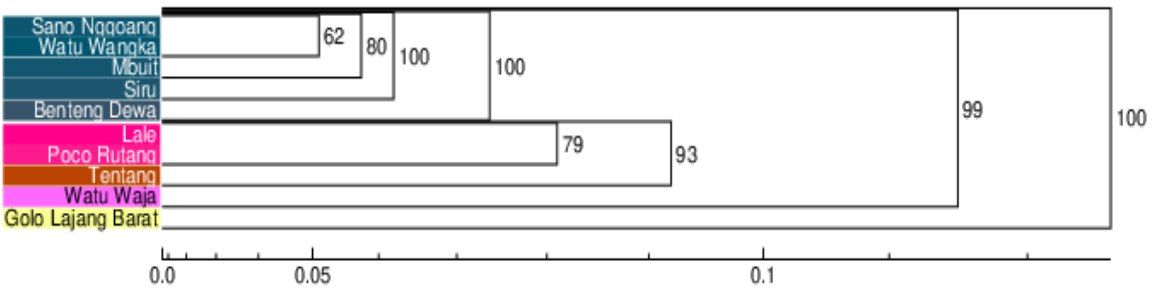

Figure 12. Dendrogram for Fuzzy Clustering Results

The dendrogram displays the percentages that indicate the likelihood of each cluster appearing with noise. Larger groups have a higher probability and are more likely to be real groups, whereas smaller groups have a lower probability and are less specific. To assess the accuracy of the groupings, one has to look from left to right (i.e., from smallest to largest groups). If a significant grouping appears in all repetitions, it has a probability level of 100%, indicating that the group is correct. If the probability presentation is smaller than that, the grouping is not entirely reliable.

Figure 12 shows that the Sano Nggoang and Watu Wangka clusters have a low % probability level of 62%, whereas the Lale and Poco Rutang clusters have a probability level of 79%. However, if the Sano Nggoang and Watu Wangka clusters include Mbuit, Siru, and Benteng Dewa in the group, the probability level becomes stable and correct at 100%. Similarly, if all observation areas are grouped into one cluster in the order of the dendrogram, the probability level will also be 100%.

In other words, Sano Nggoang, Watu Wangka, Mbuit, Siru, and Benteng Dewa are classified into a separate group of language variations called the MSdK group. Meanwhile, the grouping of Lale, Poco Rutang, and Regarding still needs further proof because they have a smaller probability level, so they cannot be included in the same cluster. Additionally, Watu Waja and West Golo Lajang appear at the top level, meaning they have the most different levels from other observation areas. Both locations, however, show a higher level of closeness.

CONCLUSION

The groupings resulting from the gabmap clustering analysis function provide validation for dialectometric calculations. The Kempo dialect group, which consists of Sano Nggoang, Watu Wangka, Mbuit, Siru, and Benteng Dewa, is the most accurately grouped dialect area between the two methods. However, other observation areas appear inconsistent. In dialectometric calculations for phonological differences, Golo Lajang Barat cannot be grouped with Watu Wangka, Mbuit, and Tentang because they show dialect differences. A similar case also applies to the Watu Waja area, which cannot be classified into the same group as Benteng Dewa and Siru because all three speak different dialects. Because the Mbuit, Watu Wangka, Sano Nggoang, Siru, and Benteng Dewa areas do not show phonological differences, these areas can be included in the same group.

Looking at the hierarchical grouping in Figure 10, it can be seen that West Golo Lajang is the observation area with the highest differences compared to other sites and is followed by Watu Waja, which has a higher level of similarity but is not on the same level as Golo Lajang Barat. After these two areas, eight other observation areas were classified into two groups. The first group consists of the Mbuit, Watu Wangka, Sano Nggoang, Siru, and Benteng Dewa areas, where the Manggarai Kempo (BMK) language is spoken. The second group refers to the area where the transitional dialect of the Manggarai language is spoken, namely the observation area whose linguistic features are absorbed from the Kempo and Kolang dialects. This group includes the Lale, Poco Rutang, and Tentang area groups.

NOTE

I would like to thank an anonymous reviewer for very helpful comments on the earlier draft of this paper.

REFERENCES

Ayatrohaedi. (1978). *Bahasa Sunda Di Daerah Cirebon Sebuah Kajian Lokabahasa*. Jakarta: Universitas Indonesia Library.

Ayatrohaedi. (1979). *Dialektologi: sebuah pengantar*. Jakarta: Pusat Pembinaan dan Pengembangan Bahasa.

Baru, D. V. D.; & Saripudin, A. (2022). Variasi Dialek Bahasa Manggarai Nusa Tenggara Timur. Undergraduate thesis, Sriwijaya University.

Boberg, C.; Nerbonne, J.; & Watt, D. (2018). *The Handbook of Dialectology*. United States: Wiley Blackwell.

Burger, A. (1946). Voorlopige Manggaraise Spraakkunst. *Bijdragen tot de Taal-, Land- en Volkenkunde 103*(1):15–265. https://doi.org/10.1163/22134379-90001214

Chelliah, S. L.; & Reuse, W. J. (2010). *Handbook of Descriptive Linguistic Fieldwork*. New York City: Springer Publishing.

Everitt, B. S.; Landau, S.; Leese, M.; & Stahl, D. (2011). *Cluster Analysis* (5th edition). John Wiley & Sons. https://doi.org/10.1002/9780470977811

Feleke, T. L.; Gooskens, C.; & Rabanus, S. (2020). Mapping the Dimensions of Linguistic Distance: A Study on South Ethiosemitic Languages. *Lingua* 243:102893. https://doi.org/10.1016/j.lingua.2020.102893

Fernandez, I. Y. (1993). *Relasi historis kekerabatan bahasa Flores: Kajian linguistik historis komparatif terhadap sembilan bahasa di Flores*. Kupang: Nusa Indah.

Goebl, H. (1984). *Dialektometrie, Prinzipien und Methoden des Einsatzes der Numerischen Taxonomie im Bereich der Dialektgeographie*. Vienna: Austrian Academy of Sciences Press.

Grimes, C. E. (1997). *A Guide to the People and Languages of Nusa Tenggara*. Artha Wacana Press.

Heeringa, W. & Nerbonne J. (2001). Dialect Areas and Dialect Continua. *Language Variation and Change 13*(3):375–400. https://doi.org/10.1017/S0954394501133041

Heeringa, W. & Nerbonne, J. (2004). Dialect Areas and Dialect Continua. *Language Variation and Change 13*(3):375–400. https://doi.org/10.1017/S0954394501133041

Jeszenszky, P.; Steiner, C.; & Leemann, A. (2021). Reduction of Survey Sites in Dialectology: A New Methodology Based on Clustering. *Front. Artif. Intell. 4*:642505. https://doi.org/10.3389/frai.2021.642505

Kiki, R. (2019). *Variasi Dialek Bahasa Manggarai Kajian: Dialektologi Diakronis*. Skripsi, Universitas Muhammadiyah Mataram.

Koile, E.; Chechuro, I.; Moroz, G.; & Daniel, M. (2022). Geography and language divergence: The case of Andic languages. *PLoS ONE 17*(5): e0265460. https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0265460

Kessler, B. (1995). Computational dialectology in Irish Gaelic. In *Seventh Conference of the European Chapter of the Association for Computational Linguistics*, Dublin, Ireland. Association for Computational Linguistics. https://aclanthology.org/E95-1009/

Lauder, M.R.M.T. (1993). *Pemetaan dan Distribusi Bahasa-Bahasa di Tangerang*. Jakarta: Pemetaan dan Distribusi Bahasa-Bahasa di Tangerang.

Leinonen, T.; Çöltekin, Ç.; & Nerbonne, J. (2016). Using Gabmap. *Lingua 178*(1): 71–83. https://doi.org/10.1016/j.lingua.2015.02.004

Mahsun. (2004). *Dialektologi Diakronis: Sebuah Pengantar*. Yogyakarta: Gadjah Mada University Press.

McDavid, R. I. (1946). Dialect Geography and Social Science Problems. *Social Forces 25*(2): 168–72. https://doi.org/10.2307/2571555

Nadra & Reniwati. (2009). *Dialektologi: Teori dan Metode*. Yogyakarta: Elmatera Publishing.

Nerbonne, J. (2010). Measuring the Diffusion of Linguistic Change. *Philosophical Transactions of the Royal Society B: Biological Sciences 365*(1559): 3821–28. https://doi.org/10.1098/rstb.2010.0048

Nerbonne, J. & Kretzschmar, W. (2023). Introducing Computational Techniques in Dialectometry. *Computers and the Humanities 37*(1): 245-255. https://doi.org/10.1023/A:1025064105053

Parera, J. D. (1991). *Kajian Linguistik Umum Historis Komparatif dan Tipologi Struktural*. Jakarta: Erlangga.

Prokić, J. (2010). *Families and Resemblances* [Doctoral Dissertation, University of Groningen]. Groningen Dissertations in Linguistics. https://research.rug.nl/en/publications/families-and-resemblances

Séguy, J. (1973). La dialectométrie dans l'Atlas linguistique de la Gascogne. *Revue de Linguistique Romane*, Vol. 37: 145–146. https://doi.org/10.5169/seals-658403

Verheijen, J. A. J. (1967). *Kamus Manggarai*. The Hague: Nijhoff.

Verheijen, J. A. J & Grimes, C. E. (1995). Manggarai. In D.T. Tyron (Ed.). *Comparative Austronesian Dictionary: An Introduction to Austronesian Studies*, 585-592. De Gruyter Mouton. https://doi.org/10.1515/9783110884012

Wieling, M. & Nerbonne, J. (2015). Advances in Dialectometry. *Annual Review of Linguistics 1*(1): 243–64. https://doi.org/10.1146/annurev-linguist-030514-124930

APPENDIX

Appendix 1. Example of a Coding Model with Python Automation

	Mbuit	Watu Waja	Benteng Dewa	Lale	Poco Rutang	Siru	Golo Lajang Barat	Sano Nggoang	Watu Wangka	Tentang
Mbuit	debu	1	0	1	1	1	1	0	0	0
Watu Waja	1	debu	1	0	0	1	0	1	1	1
Benteng Dewa	0	1	debu	1	1	1	1	0	0	0
Lale	1	0	1	debu	0	1	0	1	1	1
Poco Rutang	1	0	1	0	debu	1	0	1	1	1
Siru	1	1	1	1	1	debu	1	1	1	1
Golo Lajang Barat	1	0	1	0	0	1	debu	1	1	1
Sano Nggoang	0	1	0	1	1	1	1	debu	0	0
Watu Wangka	0	1	0	1	1	1	1	0	debu	0
Tentang	0	1	0	1	1	1	1	0	0	debu

	Mbuit	Watu Waja	Benteng Dewa	Lale	Poco Rutang	Siru	Golo Lajang Barat	Sano Nggoang	Watu Wangka	Tentang
Mbuit	air	0	0	0	0	0	0	0	0	0
Watu Waja	0	air	0	0	0	0	0	0	0	0
Benteng Dewa	0	0	air	0	0	0	0	0	0	0
Lale	0	0	0	air	0	0	0	0	0	0
Poco Rutang	0	0	0	0	air	0	0	0	0	0
Siru	0	0	0	0	0	air	0	0	0	0
Golo Lajang Barat	0	0	0	0	0	0	air	0	0	0
Sano Nggoang	0	0	0	0	0	0	0	air	0	0
Watu Wangka	0	0	0	0	0	0	0	0	air	0
Tentang	0	0	0	0	0	0	0	0	0	air

Appendix 2. Example of a summary of the comparison of all words between one particular region and another region

Observation area 1-7 (Mbuit and Golo Lajang Barat)					
Gloss	**Value**	**Gloss**	**Value**	**Gloss**	**Value**
debu	1	flower	1	hit	1
air	0	kill	0	erase	1
root	0	hunt	1	liver	0
1st Singular "I"	0	bad	0	nose	1
stream	0	bird	1	life	1
child	1	rotten	1	green	1
wind	0	earthworm	1	suck	1
dog	1	kiss	0	black	0
what	1	wdebu	1	count	0
fire	0	meat	1	rain	1
float	0	and	0	forest	0
smoke	1	lake	1	3rd person singular	1
cloud	0	blood	0	mother	0
how	1	come	0	fish	1
good	0	leaf	1	tie	0
burn	0	dust	1	wife	0
reverse	1	near	0	this	1
plenty	0	with	0	that	1
father	0	hear	1	sew	1
lie down	1	in (inside)	1	street	1
new	0	in (location)	0	heart	1
wet	1	where	0	fall	0
stone	0	cold	1	far	1
how many	1	stand	1	fog	0
split	1	here	0	feet	0
right/correct	0	there	0	if	1

Observation area 1-7 (Mbuit and Golo Lajang Barat)					
seed	1	push	1	we	1
swollen	0	two	1	you (polite)	0
swim	0	sit	1	right	0
walk	0	tail	0	because	1
wight	0	four	0	talk	0
give	1	2nd Singular	1	small	1
big	1	dig	1	fight	1
when	1	salt	1	head	1
animal	0	scratch	1	dry	0
star	0	fat	1	left	1
fruit	0	teeth	1	dirty	1
moon	0	bite	0	nail	0
feather	0	rub	1	skin	0
flower	1	mountain	0	yellow	1

Appendix 3. Geographic distance between regions

	Mbuit	Watu Waja	Benteng Dewa	Lale	Poco Rutang	Siru	Golo Lajang Barat	Sano Nggoang	Watu Wangka	Tentang
Mbuit	0	0.1421	0.0714	0.0988	0.1205	0.0636	0.1242	0.0595	0.0494	0.0981
Watu Waja	0.1421	0	0.1037	0.0836	0.099	0.139	0.1169	0.1437	0.1416	0.1194
Benteng Dewa	0.0714	0.1037	0	0.0861	0.0927	0.0661	0.1295	0.0736	0.0767	0.0828
Lale	0.0988	0.0836	0.0861	0	0.0786	0.1103	0.0919	0.1057	0.1006	0.081
Poco Rutang	0.1205	0.099	0.0927	0.0786	0	0.1163	0.1352	0.1229	0.1217	0.0976
Siru	0.0636	0.139	0.0661	0.1103	0.1163	0	0.1377	0.0554	0.059	0.103
Golo Lajang Barat	0.1242	0.1169	0.1295	0.0919	0.1352	0.1377	0	0.1338	0.1214	0.1263
Sano Nggoang	0.0595	0.1437	0.0736	0.1057	0.1229	0.0554	0.1338	0	0.0489	0.1074
Watu Wangka	0.0494	0.1416	0.0767	0.1006	0.1217	0.059	0.1214	0.0489	0	0.1011
Tentang	0.0981	0.1194	0.0828	0.081	0.0976	0.103	0.1263	0.1074	0.1011	0