Graph Modeling of Lexical Distance in the Austronesian Language Family Subgroups

Ananda Aulia Nurramadhan - 13525135
Program Studi Teknik Informatika
Sekolah Teknik Elektro dan Informatika
Institut Teknologi Bandung, Jalan Ganesha 10 Bandung
E-mail: nandanurau@gmail.com, 13525135@std.stei.itb.ac.id

***Abstract*—The Austronesian language family is one of the largest language families, comprising 1,200 languages, extending from Madagascar to Easter Island. This language family is then divided into two main subgroups, Formosan and Malayo-Polynesian. This paper integrates the concepts of graph theory and lexical distance to assess the extent of how different the vocabularies are among languages in each Austronesian subgroups. This paper utilizes Swadesh lists of the two subgroups, containing common vocabulary, as the basis for calculating lexical distance. Then, weighted graphs for both subgroups were modelled based on lexical distance, with nodes representing languages, and vertices representing lexical distance.**

Keywords—graph theory; Austronesian language family; lexical distance

I. Introduction

The Austronesian language family is a language family with a massive expanse, geographically extending from the island of Madagascar situated in the Indian Ocean to Easter Island in the southeastern portion of the Pacific. This language family includes languages of small groups of inner hill people practicing shifting agriculture in Mindanao, Borneo, and Sulawesi, languages of large, stratified societies such as Balinese and Javanese, and trade languages such as Malay. The common ancestor of the Austronesian languages has been thought to have originated 5,000 years ago in Taiwan.

The language family diverges into two main subgroups or branches: Formosan and Malayo-Polynesian. The Formosan languages are found in Taiwan, most of them being extinct, with 14 languages still having a presence in the island, albeit endangered. These languages used to flourish on the island before the settlement of the Chinese. Meanwhile, the Malayo-Polynesian is a behemoth that encompasses every other Austronesian language outside of Taiwan. It encompasses languages found mostly in the Madagascar, Malay Archipelago, Micronesia, Melanesia, and Polynesia. The largest Austronesian languages belong to this subgroup, such as Malay, Tagalog, and Javanese.

The lexical distance between two languages is a metric of how different their vocabularies are. The definition of lexical distance used in this paper is the average of the sum of a normalized Levenshtein distance between cognates across two languages. A lexical distance of 0 would mean a perfect overlap in vocabulary, while a lexical distance of 1 would mean no overlap in vocabulary whatsoever.

To model this metric, this paper utilizes graph theory, which is a concept of the representation of discrete objects and how they are related. Graphs represent a discrete object as a node, and the relationship between two nodes as a vertex, connecting the two. A vertex can have a value (weight), or even a set direction from one node to another.

In this paper, undirected weighted graphs will be modeled according to the lexical distance of languages in the two main Austronesian language subgroups Formosan and Malayo-Polynesian. The purpose of this paper is to assess the divergence of the Austronesian language family and demonstrate the application of graph theory in linguistics, specifically lexicostatistics.

II. Theoretical Framework

A. Graph

Graphs are used to represent discrete objects and their relationships, where the former is represented using vertices and the latter using edges. Mathematically, a graph G is defined as $G = (V, E)$, where V is a set of vertices and E is a set of pairs of vertices (edges). V cannot be empty, while E can. In other words, a graph must have vertices, while edges are optional.

A graph's edges can have a value and orientation. If each of a graph's edges has a value, it is considered a weighted graph. Meanwhile, if each of a graph's edges has orientation, it is considered a directed graph (or alternatively, digraph). These two traits are not mutually exclusive, so there can exist a weighted digraph.

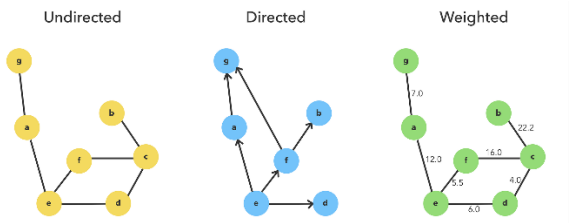


Fig. 1. Undirected, directed, and weighted graphs side by side. Adapted from https://fahadsultan.com/csc223/datastructs/graphs.html.

Adjacency and incidence are two properties in graphs that assess the connectivity of a graph. Two vertices are considered adjacent when they share an edge. A vertex that is not adjacent to any other vertices is considered an isolated vertex. For an edge $e = (v_j, v_k)$, e is considered incident with vertices v_j and v_k. The degree of a vertex v, $d(v)$, is the number of edges that are incident with v.

Other than with a set, a graph can be represented using a matrix, in two ways. The first is using an adjacency matrix, where the rows and columns are the vertices of a graph. For a graph G and its adjacency matrix A, if G is unweighted, a_{ij} has a value of 1 if vertices v_i and v_j are adjacent, 0 if otherwise. However, if G is weighted, then the value of a_{ij} would be the weight of the edge that is incident with both v_i and v_j. If the weight of edge (v_i, v_j) is undefined, its corresponding matrix elements are substituted with a value of 0 or ∞.

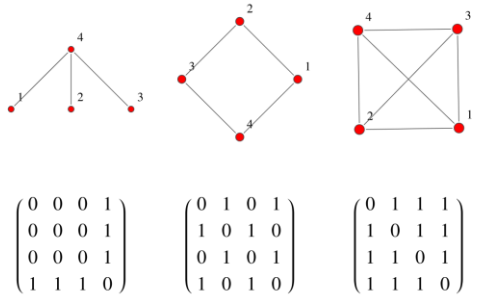


Fig. 2. Undirected graphs and their adjacency matrix representation. Adapted from https://mathworld.wolfram.com/AdjacencyMatrix.html.

The second is in the form of an incidence matrix, where the rows represent the vertices of a graph and the columns represent the edges. For a graph G and its incidence matrix A, a_{ij} has the value of 1 if vertex v_i is incident with edge e_j, 0 if otherwise.

B. Austronesian Language Family

The Austronesian language family is a language family that consists of roughly 1,200 languages, constituting 20% of the world's languages, and encompasses Malagasy (in Madagascar) in the western part of the Indian Ocean to Easter Island in the southeastern part of the Pacific Ocean, and the Formosan languages in Taiwan and Hawaiian in the northern Pacific to Māori in New Zealand. The Austronesian languages are spoken by around 270 million people worldwide, in Madagascar, Malaysia, Indonesia, the Philippines, Taiwan, coastal New Guinea, and in the archipelagos of Melanesia, Micronesia, and Polynesia. It is also spoken in pockets on mainland Asia, situated in southern Vietnam and Cambodia and on Hainan Island in southern China. In the western regions where Austronesian is spoken, some languages are spoken by millions, while many in the eastern regions are spoken only by one thousand or less people on average. Javanese, Sundanese, Malay, and Tagalog are the largest Austronesian languages.

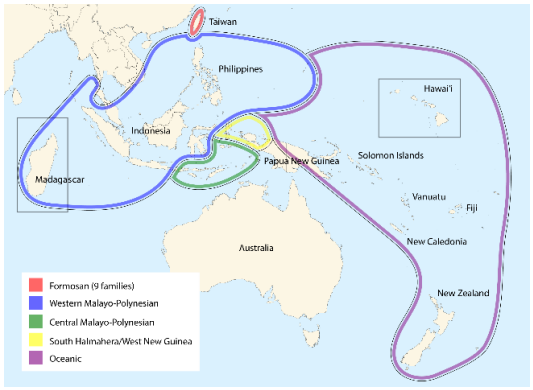


Fig. 3. Geographical distribution of the Austronesian language family. Adapted from https://en.wikipedia.org/wiki/Austronesian_languages.

Austronesian languages are thought to descend from a single ancestor, likely spoken in Taiwan around 5,000 years ago. This ancestor is divided into the Formosan languages of Taiwan, and Malayo-Polynesian.

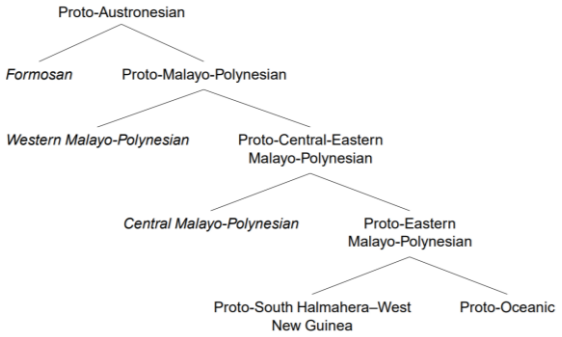


Fig. 4. The Austronesian language family tree. Adapted from D. C. Kamholz, "Austronesians in Papua: Diversification and Change in South Halmahera–West New Guinea," Ph.D. dissertation, University of California, Berkeley, California, USA, 2014.

1) Formosan

Formosan includes several extinct and fourteen living languages in Taiwan, belonging to nine phylogenetic groups. Formosan languages are spoken by about 300,000 people in the valleys of the central mountain chain that divides the country vertically. The Formosan languages used to dominate the island before the arrival of the Chinese. Now, they all face endangerment.

2) Malayo-Polynesian

Malayo-Polynesian is divided into Western Malayo-Polynesian and Central-Eastern Malayo-Polynesian.

a) Western Malayo-Polynesian

Western Malayo-Polynesian languages are spoken in the Philippines, Malaysia, most of Indonesia, and Madagascar. This group also includes the Chamic languages which are spoken in mainland Southeast Asia and China.

b) Central-Eastern Malayo-Polynesian

Central-Eastern Malayo-Polynesian languages are spoken in eastern Indonesia, select areas in New Guinea, and the archipelagos of Micronesia, Melanesia, and Polynesia. Central-

Eastern Malayo-Polynesian is composed of Central Malayo-Polynesian and Eastern Malayo-Polynesian, the latter of which is further divided into South Halmahera-West New Guinea and Oceanic.

Central Malayo-Polynesian includes around 50 languages spoken in the Lesser Sunda Islands of Flores, Sumba, and Timor as well as the Moluccas. South Halmahera-West New Guinea refers to the 45 or so languages spoken in the southern half of the Moluccan island of Halmahera and in the Doberai Peninsula of western New Guinea. Oceanic languages are spoken by around 2 million people scattered around Polynesia, Micronesia, and Melanesia. Around 400 languages are spoken in Melanesia alone, which is considered one of the most linguistically diverse regions in the world.

C. Cognate

Cognates are words that share a common ancestor, deriving from the same root in a proto-language, which tend to have similar meanings and pronunciations across different languages, e.g. English *mother*, German *Mutter*, and Spanish *madre*, which are all cognates, tracing back to Proto-Indo-European **méh₂tēr*. However, not all words that look or sound similar are cognates. Words that sound similar but have different meanings and origins are referred to as false cognates or false friends.

Cognates provide evidence of the relationships between languages. By comparing cognates across languages, linguists can infer properties of the common ancestral language. Cognates also help in classifying languages into families. For instance, the existence of cognates like English *night*, German *Nacht*, and Spanish *noche* supports the classification of these languages into the Indo-European family. Additionally, regular sound changes, such as Grimm's Law in Germanic languages, show systematic phonetic shifts that can be traced through cognate sets.

Cognates also highlight the influence of language contact and borrowing. Languages often adopt words from each other, leading to the creation of cognates. Many English words are borrowed from Latin and French due to historical events such as the Roman occupation and the Norman Conquest. This is evident by words like *library* (from Latin *libraria*) and *government* (from Old French *governer*). Borrowing can occur due to trade, conquest, and cultural exchange.

D. Levenshtein Distance

Given two strings u and v over an alphabet Σ, the Levenshtein distance from u to v (or vice versa) is the number of edit steps needed to change u to v, where an edit step encompasses the action of deleting, inserting, or replacing a symbol.

For example, the Levenshtein distance from “WARM” to “BEAR” is 3, since “WARM” can be changed to “BEAR” with three edit steps: “WAR” (delete M, step 1), “BAR” (replace W with B, step 2), “BEAR” (insert E, step 3).

The Levenshtein distance between two strings can be computed using dynamic programming. Let n be the length of u and m the length of v. Let $u[i]$ be the prefix of u of length i and $v[j]$ be the prefix of v of length j, for $0 \le i \le n$ and $0 \le j \le m$. Subproblem (i, j) is defined to be the computation of the Levenshtein distance from $u[i]$ to $v[j]$, which is referred to as $L[i, j]$. There is a total of $(n + 1)(m + 1)$ subproblems. The Levenshtein distance is $L[n, m]$. The algorithm to calculate the Levenshtein distance is as follows:

```
1  for all i let L[i, 0] = i
2  for all j let L[0, j] = j
3  for all 1 ≤ i ≤ n
4      for all 1 ≤ j ≤ n
5          if (u[i] = v[j])
6              L[i, j] = min{L[i - 1, j] + 1, L[i, j - 1] + 1,
   L[i - 1, j - 1]}
7          else
8              L[i, j] = min{L[i - 1, j] + 1, L[i, j - 1] + 1,
   L[i - 1, j - 1] + 1}
```

			B	E	A	R
		0	1	2	3	4
	0	0	1	2	3	4
W	1	1	1	2	3	4
A	2	2	2	2	2	3
R	3	3	3	3	3	2
M	4	4	4	4	4	3

Fig. 5. Subproblem matrix of the Levenshtein distance of “WARM” and “BEAR” according to the provided algorithm. Notice that $L[n, m] = 3$. Adapted from [6].

E. Lexical Distance

The lexical distance between two languages is the degree of the difference between the vocabulary of the two languages. One way to assess the lexical distance between two languages is by using the Levenshtein distance between pairs of words across two languages. Let α and β be languages, and α_i and β_j be words from α and β respectively. The lexical distance between α_i and β_j, $D(\alpha_i, \beta_j)$, is given by

$$D(\alpha_i, \beta_j) = \frac{D_l(\alpha_i, \beta_j)}{L(\alpha_i, \beta_j)} \tag{1}$$

where $D_l(\alpha_i, \beta_j)$ is the Levenshtein distance between two words and $L(\alpha_i, \beta_j)$ is the number of characters of the longer of the two words α_i and β_j. Therefore, the distance can have a value between 0 and 1. If two words are the same, then the value of the lexical distance between them is 0.

Now, assume the list of words for any language that is being compared contains M items. Any word of a language α is α_i with $1 \le i \le M$. Then, two words α_i and β_j have the same meaning if $i = j$. The distance between two languages is

$$D(\alpha, \beta) = \frac{1}{M} \sum_i D(\alpha_i, \beta_i) \tag{2}$$

where the sum goes from 1 to *M*. Only pairs of words with the same meaning are used in this definition.

III. Methodology

For this paper, an undirected weighted graph will be modelled based on the lexical distance of only a select few Austronesian languages. Two versions of the graph will be created, according to the two main branches of the Austronesian language family: Formosan and Malayo-Polynesian. For both graphs, nodes represent languages and vertices represent the lexical distance between two languages. To improve the readability of the graph, a color spectrum will be used to illustrate lexical distance on each vertex and the thickness of a vertex will scale based on lexical distance.

This paper uses [8] and [9], comparative vocabulary lists (also referred to as Swadesh lists), providing common words that are cognates across a few languages of the Formosan and (mostly) Malayo-Polynesian branch respectively. The Formosan list includes 12 unique languages, for a total of 24 items as some languages are represented in a number of dialects, such as the language of Amis, being represented through two dialects, Fata'an and Farang. This list includes 206 basic words, such as *hand*, *left*, *right*, *leg*, and *walk* in the respective languages in the list, with a few missing entries here and there, most notably with the Sakizaya language. Meanwhile, the Malayo-Polynesian list includes a total of 22 languages, and includes some of the largest Austronesian languages, such as Malay, Tagalog, and Javanese. However, one of the languages included, Taivoan, is not a Malay-Polynesian language. So, for the purposes of this paper, only 21 of the 22 languages in the list are considered. The list of words used is pretty much identical to the one used in the Formosan list, with 207 words in the respective languages. This list also has some missing entries, most notably with the Itawis language.

Both vocabulary lists were adapted to make calculations feasible. Hyphens, semicolons, apostrophes, and any notes were removed from each entry, leaving "pure" letters behind. Additionally, the lists were turned into TSV (tab-separated values) files as some entries contain more than one word, which are then each separated by commas.

Calculations of lexical distance utilize (1) and (2). For entries that encompass more than one word, the smallest possible lexical distance is chosen. As stated before, there are several missing entries. To mitigate this, any calculations involving missing entries are skipped. This means that the amount of word pairs may potentially not be constant. In order to make calculations of lexical distance easier, a Python script was created, implementing (1) and (2). (The Python script, as well as the adapted vocabulary lists are available in the GitHub repository provided in the appendix.)

IV. Results and Discussion

Using the adapted dataset for both branches and using the Python script, we obtain the lexical distance for each pair of languages in each branch, in the form of an adjacency matrix, with each element representing the lexical distance between the language at row *i* and column *j*. Note that the elements of the main diagonal in each matrices are undefined. Notice that for elements of the main diagonal, $i = j$. In other words, the language is the same. (Technically, the lexical distance is still defined, i.e. with a value of 0, but it is insignificant. Therefore, it is not included in the modeling.) Due to the full tables being way too large to fit in these pages, only a few notable snippets are provided below.

TABLE I. Closest Formosan Languages According to Lexical Distance

Language 1	Language 2	Lexical Distance
Seediq (Hecuo)	Seediq (Truku)	0.092525405
Paiwan (Butanglu)	Paiwan (Tjubar)	0.092660457
Saisiyat (Tungho)	Saisiyat (Ta'ai)	0.102952487
Seediq (Toda)	Seediq (Truku)	0.102964643
Seediq (Toda)	Seediq (Hecuo)	0.106558928

TABLE II. Furthest Formosan Languages According to Lexical Distance

Language 1	Language 2	Lexical Distance
Tsou (Duhtu)	Atayal (Squliq)	0.896593915
Tsou (Duhtu)	Atayal (Skikun)	0.879691634
Tsou (Duhtu)	Saisiyat (Ta'ai)	0.877443025
Rukai (Mantauran)	Atayal (Skikun)	0.876090849
Rukai (Tana)	Atayal (Skikun)	0.875229415

TABLE III. Closest Malayo-Polynesian Languages According to Lexical Distance

Language 1	Language 2	Lexical Distance
Ibanag	Itawis	0.293404118
Indonesian	Malay	0.330346285
Kangeanic	Madurese	0.340143134
Tagalog	Central Bikol	0.365857975
Cebuano	Central Bikol	0.370433632

TABLE IV. Furthest Malayo-Polynesian Languages According to Lexical Distance

Language 1	Language 2	Lexical Distance
Sasak	Tahitian	0.849800479
Ilocano	Tahitian	0.840857915
Tahitian	Kangeanic	0.83557371
Tahitian	Malay	0.834165218
Chamorro	Tahitian	0.827679002

With the calculated lexical distances in the form of adjacency matrices, we can construct the graph for both branches. The vertices in the graphs follow a scale based on its

weight (lexical distance). The thicker and purpler a vertex is, the lexical distance between the incident nodes (languages) is smaller. The thinner and bluer, ditto, but the other way. (Unfortunately, the graphs are too large to display here with decent quality and readibility. For the reader's sake, both graphs are provided in the appendix for clearer viewing.)

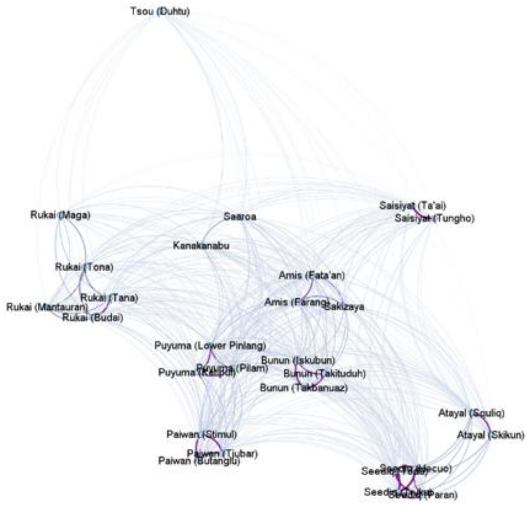


Fig. 6. Graph representation of the lexical distance among the Formosan languages. Created by the author using Gephi. (This graph is available in the appendix.)

As we can observe, from Fig. 6, most Formosan languages, standalone and dialectical alike, form their own distinct clusters, such as the Seediq and Puyuma languages. While most languages still stay within a general region, even if they belong to their respective clusters, the Tsou language is a notable exception. This is further backed up by Table 2.

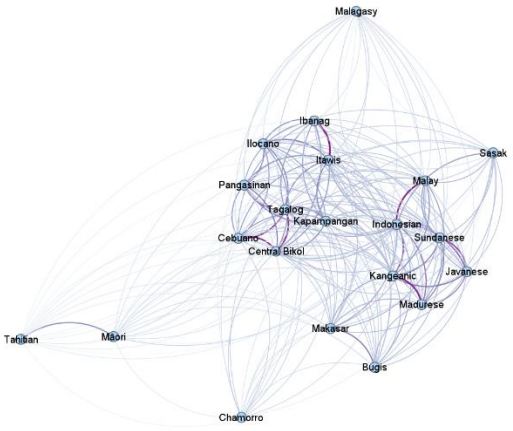


Fig. 7. Graph representation of the lexical distance among the Malayo-Polynesian languages. Created by the author using Gephi. (This graph is available in the appendix.)

Looking at Fig. 7, we can find that the Malayo-Polynesian languages form two main clusters. The first of which consists of 8 languages found in the Philippines, such as Tagalog and Cebuan. The second cluster is made up of languages found in Indonesia and Malaysia, such as Malay and Sundanese. We can notice that one language strays the furthest away, that being Tahitian, which is accurate considering Table 4.

V. Conclusion

Graph theory is a handy tool to help visualize the lexical distance between languages. By using a weighted graph and modeling it based off lexical distances, we can get a general idea of which languages have similar vocabulary, in terms of spelling (orthography). Based on the results of this paper, we can see that languages that are part of the same family can still have differences in their vocabulary. This shows that the concept of graph theory has its place in other fields, specifically linguistics.

Orthography alone isn't the only aspect to consider when we're talking about similarities between languages. This paper still hasn't considered the differences between how words are pronounced and each language (phonetics), as well as the nuance of meaning they could have in different languages (semantics). A more comprehensive list of both languages and vocabulary (aside from basic concepts) could also be considered, in order to yield more conclusive findings. Other than that, a comparison among languages of both the Formosan and Malayo-Polynesian subgroups could be conducted, to further assess the divergence between the two subgroups.

Appendix

Video on this paper by the author, uploaded on YouTube: https://youtu.be/JINfxuLOYNo

Higher quality versions of Fig. 6 and Fig. 7 on Google Drive: https://drive.google.com/drive/u/1/folders/1ET56D_Xz6aIdzPPOUIdDpmUBrNiVeXAc

GitHub repository of this paper, containing lexical distance implementation in Python and adapted vocabulary lists: https://github.com/nandanurau/Austronesian-Language-Subgroup-Lexical-Distance

Acknowledgment

The author would like to express their deepest gratitude towards Mr. Dr. Ir. Rinaldi Munir, M.T., for all the invaluable insight and knowledge he has bestowed upon the author throughout his IF1220 (Discrete Mathematics) course. The author also would like to cast light to their close friends, dearest to the author, who have given the author the will to keep going.

REFERENCES

[1] R. Munir, “Graf (Bagian 1),” IF1220 (Discrete Mathematics) course notes, Institut Teknologi Bandung, Bandung, Indonesia, 2026. [Online]. Available: https://informatika.stei.itb.ac.id/~rinaldi.munir/Matdis/2025-2026/20-Graf-Bagian1-2026.pdf. [Accessed: June 17, 2026]

[2] R. Munir, “Graf (Bagian 2),” IF1220 (Discrete Mathematics) course notes, Institut Teknologi Bandung, Bandung, Indonesia, 2026. [Online]. Available: https://informatika.stei.itb.ac.id/~rinaldi.munir/Matdis/2025-2026/21-Graf-Bagian2-2026.pdf. [Accessed: June 17, 2026]

[3] A. Adelaar, "The Austronesian languages of Asia and Madagascar: A historical perspective," in *The Austronesian Languages of Asia and Madagascar*, A. Adelaar and N. P. Himmelmann, Eds. London, UK: Routledge, 2005, pp. 1-42.

[4] A. Gutman and B. Avanzati, "Austronesian Languages," *The Language Gulper*, 2013. [Online]. Available: https://languagesgulper.com/eng/Austronesian.html. [Accessed: June 17, 2026]

[5] C. Roy, "Cognates in Linguistic Analysis: Examing the Interconnections of Language Similarities," *Mathematica Eterna*, vol. 14, no. 2, pp. 1-2, 2024. [Online]. Available: https://www.longdom.org/articles/cognates-in-linguistic-analysis-examing-the-interconnections-of-language-similarities-110344.html. [Accessed: June 17, 2026]

[6] L. Larmore, "Levenshtein Edit Distance," CSC 477 (Analysis of Algorithms) course notes, University of Nevada, Las Vegas, Nevada, USA, 2021. [Online]. Available: https://web.cs.unlv.edu/larmore/Courses/CSC477/S21/Topics/levenshtein.pdf. [Accessed: June 18, 2026]

[7] F. Petroni and M. Serva, "Measures of lexical distance between languages," arXiv preprint arXiv:0912.0884, 2009. [Online]. Available: https://arxiv.org/pdf/0912.0884. [Accessed: June 18, 2026]

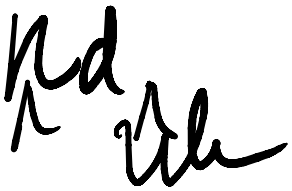
[8] Wiktionary contributors, “Appendix:Cognate sets for Formosan languages”, *Wiktionary*, https://en.wiktionary.org/wiki/Appendix:Cognate_sets_for_Formosan_languages. [Online]. [Accessed: June 18, 2026].

[9] Wiktionary contributors, "Appendix:Austronesian Swadesh lists," *Wiktionary*, 2024. [Online]. Available: https://en.wiktionary.org/wiki/Appendix:Austronesian_Swadesh_lists. [Accessed: June 18, 2026].

STATEMENT

With this, I hereby declare that the paper that I’ve written is my own work, not an adaptation nor translation of another paper, and is not the result of plagiarism.

Bandung, 19 June 2026



Ananda Aulia Nurramadhan (13525135)