PH Language Similarity Analysis

Baybayon, Anthony[1], Catignas, Ronald Dawson[2], Cruzado, Jose Paolo[3], and Tamondong, Mariel[4]

De La Salle University

[1]anthony_baybayon@dlsu.edu.ph, [2]dawson_catignas@dlsu.edu.ph, [3]jose_paolo_cruzado@dlsu.edu.ph, [4]mariel_tamondong_a@dlsu.edu.ph

1. Introduction

This paper presents a computational analysis exploring the relationships among 16 languages, 13 of which are native to the Philippines. The primary objective is to investigate how these languages relate to one another by quantifying their orthographic (character-based textual) similarity. The languages analyzed are Adasen, Bikolano, Cebuano, Chavacano, English, Ilokano, Ilonggo, Kinaray-A, Masbatenyo, Paranan, Romblomanon, Spanish, Tagalog, Tausug, Waray, and Yami.

2. Data and Preprocessing

2.1 Data Collection

To build the multilingual corpora that are comparable, which is essential when performing language similarity analysis, we collected all our data from bible texts in the 16 different languages from a single source utilizing a custom web scraper [1]. It is also noted that to ensure textual consistency across all languages, the corpora were restricted to three specific books of the New Testament: Mark, Matthew, and Luke. The total size of the corpora is 1,200,405 words.

Table 1. Corpora Word Distribution

Language	Word Count	Percentage
Adasen	92,218	7.7%
Bikolano	64,540	5.4%
Cebuano	68,897	5.7%
Chavacano	97,055	8.1%
English	66,890	5.6%
Ilokano	59,445	5.0%
Ilonggo	74,632	6.2%
Kinaray-A	77,204	6.4%
Masbatenyo	71,226	5.9%
Paranan	74,123	6.2%
Romblomanon	75,652	6.3%
Spanish	62,683	5.2%
Tagalog	67,231	5.6%
Tausug	90,659	7.6%
Waray	71,798	6.0%
Yami	86,152	7.2%

2.2 Preprocessing

After collection, all 16 corpora were cleaned using a uniform preprocessing pipeline to prepare them for the feature extraction step. This process consisted of four steps:

1. **Convert to Lowercase:** All text was converted to lowercase.
2. **Remove Numbers:** All numerical digits were removed. While the initial goal was to target verse numbers, it was determined that a general removal of all digits was simpler and equally effective. Since the analysis relies on character n-grams, this naive removal sufficiently cleans the text of numerical noise.
3. **Remove Punctuation and Special Characters:** All non-alphanumeric characters were removed.
4. **Normalize Whitespace**: All newlines, tabs, and excess spaces were normalized to a single space.

This cleaning process results in a normalized corpus ready for character-level feature extraction.

3. Methodology

3.1 Feature Engineering

3.1.1 N-Gram Profile Creation

To convert the cleaned corpora into a measurable, quantitative format, we employed a feature engineering process based on character n-grams. The core of this process was to create a comprehensive n-gram profile for each language.

This was executed as follows:

1. **N-Gram Generation:** We selected a character n-gram size of n = 3 (trigrams). Before generating the n-grams, all spaces in the text were converted to underscores (_) to ensure that word boundaries were preserved as distinct trigrams while making them more explicit and readable (e.g., ang vs. ng_). Overlapping trigrams were then extracted from the entire corpus of each language.
2. **Frequency Counting:** For each of the 16 languages, the raw frequency of every unique trigram was counted. This resulted in 16 separate frequency distributions, one for each language.
3. **Master Vocabulary Creation:** A single master vocabulary was created by compiling all unique trigrams found across all 16 languages. The master vocabulary consists of 6,378 unique 3-grams.
4. **Frequency Matrix Construction:** A master matrix was constructed. In this matrix, each row represents one of the 6,378 unique trigrams, and each column represents one of the 16 languages. The cells were populated with the raw frequency of that trigram in that language. If a trigram from the master vocabulary did not appear in a specific language, its value was set to 0.

This matrix of raw counts serves as the input for the next stage of the methodology: TF-IDF vectorization and similarity computation.

3.1.2 TF-IDF Transformation

The raw frequency matrix is not ideal, as it can skew similarity. Common n-grams (e.g., ang, _a_) that appear frequently in all languages would dominate the comparison, even though they are not linguistically unique identifiers. To address this, we applied the TF-IDF (Term Frequency-Inverse Document Frequency) transformation to re-weight the matrix and highlight the n-grams that are most unique and informative for each language [2]. This was calculated in two parts:

1. **Term Frequency (TF)**: To account for the varying sizes of the language corpora, the raw frequencies were normalized. For each language, the raw count of every n-gram was divided by the total number of n-grams in that language's document. This gives the relative frequency of an n-gram within a single language.
2. **Inverse Document Frequency (IDF):** To measure how unique an n-gram is, we first calculated the Document Frequency (DF) which is the number of languages (documents) in which each n-gram appeared. The IDF was then computed by taking the logarithm of the total number of languages (16) divided by this Document Frequency. This calculation gives a high score to rare n-grams (that appear in few languages) and a low score to common n-grams.

The final TF-IDF score for each n-gram in each language was then calculated by multiplying its TF value by its IDF value. This resultant TF-IDF matrix, which prioritizes unique and characteristic n-grams, is the final set of features used for computing similarity.

3.2 Similarity Computation

Once the final TF-IDF feature matrix was engineered, the next step was to compute a quantitative similarity score between every pair of languages. For this, we used Cosine Similarity. This method was chosen because it is highly effective for comparing TF-IDF vectors.

Cosine similarity measures the angle between two vectors in a multi-dimensional space [3]. This evaluates how similar the direction of their n-gram profiles is, regardless of the magnitude or the total size of their respective corpora. A score closer to 1 indicates that the n-gram distributions of two languages align closely, while a score closer to 0 indicates they are orthographically dissimilar.

To perform this computation, the TF-IDF matrix, which had n-grams as rows and languages as columns, was first transposed so that each language was represented as a row vector. We then used *sklearn's* cosine_similarity function to compute the pairwise scores between all 16 language vectors. The final output of this process is a 16x16 similarity matrix, where each cell (i, j) contains the similarity score between language i and language j. This matrix provides the quantitative foundation for the subsequent clustering analysis.

3.3 Language Clustering

The 16x16 similarity matrix from the previous step provides the raw data for clustering, but it must be transformed into a format suitable for hierarchical analysis. This methodology was performed in three stages:

1. **Similarity Matrix to Distance Matrix Conversion:** Hierarchical clustering algorithms operate on distances rather than similarities. Therefore, the first step was to convert our similarity matrix. To do this, we subtract each similarity score from 1, which inverts the scale. The resulting distance matrix now quantifies the orthographic distance between each language pair.
2. **Hierarchical Linkage:** To build the language tree, we used Agglomerative Hierarchical Clustering. This method required the 16x16 distance matrix to be converted into a condensed distance vector, which was done using *scipy.spatial.distance.squareform*. This vector was then fed into the linkage function from *scipy.cluster.hierarchy*. We chose the average linkage method, which defines the distance between two clusters as the average of all pairwise distances between the languages in one cluster and the languages in the other. This linkage matrix (Z) mathematically defines the structure of the dendrogram.
3. **Determining Optimal Clusters (k):** The linkage matrix creates a full tree, but it does not tell us the "optimal" number of distinct clusters (k). To find this, we used the Silhouette Score method. We programmatically cut the tree (Z) into k clusters (for every k from 2 to 10) using *fcluster* and calculated the average silhouette score for each cut. The analysis showed that k=6 produced the highest score (0.4746), suggesting it is the most statistically significant grouping for this dataset. This value was used to set the color threshold for the final dendrogram visualization.

The Cophenetic Correlation Coefficient was also calculated to be 0.992, indicating that the dendrogram is a highly faithful and non-distorted representation of the original distances [4]. The resulting visualizations are presented and interpreted in Section 4.

4. Results and Interpretation

4.1 Similarity Matrix

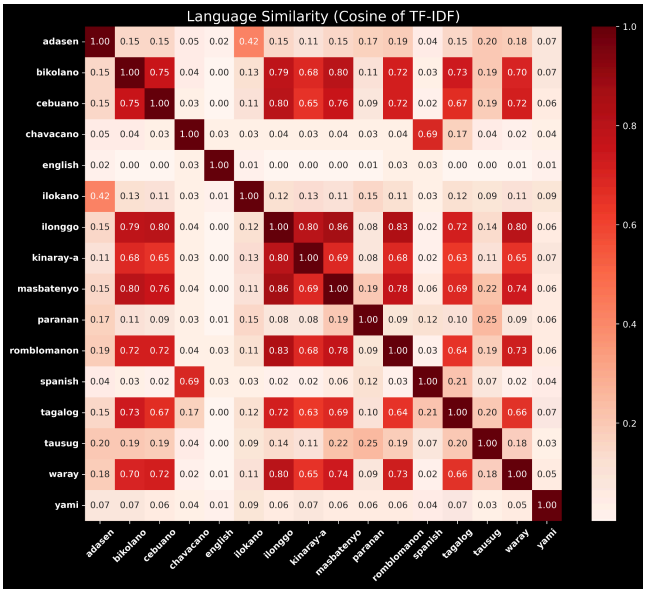

Figure 1. Similarity Matrix of the 16 Languages

Based on the similarity matrix shown in Figure 1, here are the top and bottom 5 most similar and dissimilar language pairs.

Top 5 Similar Language Pairs:

1. Ilonggo & Masbatenyo: **0.86**
2. Ilonggo & Romblomanon: **0.83**
3. Bikolano & Masbatenyo: **0.80**
4. Ilonggo & Kinaray-a: **0.80**
5. Ilonggo & Waray: **0.80**

The high similarity among the top 5 language pairs can be justified and seen as non-coincidental as all are members of the Greater Central Philippine language family [5]. This strongly indicates that the character n-gram methodology successfully captured the significant orthographic overlap shared by these linguistically related languages.

Top 5 Dissimilar Language Pairs:

1. English & Kinaray-a: **0.003**
2. English & Ilonggo: **0.003**
3. English & Tagalog: **0.003**
4. English & Masbatenyo: **0.003**
5. English & Cebuano: **0.004**

Looking at the dissimilar pairs, we can notice an evident pattern wherein all 5 are English & a particular language. Looking deeper, we can see that English is dissimilar to every language in the analysis, with all of its respective scores being less than 0.05 in the matrix.

The other two non-Philippine languages (Spanish & Yami) also showed high dissimilarity among all languages with one outlier being Spanish and Chavacano, with a similarity of 0.69, which can be attributed to Chavacano being a Spanish-based creole [6].

4.2 Dendrogram

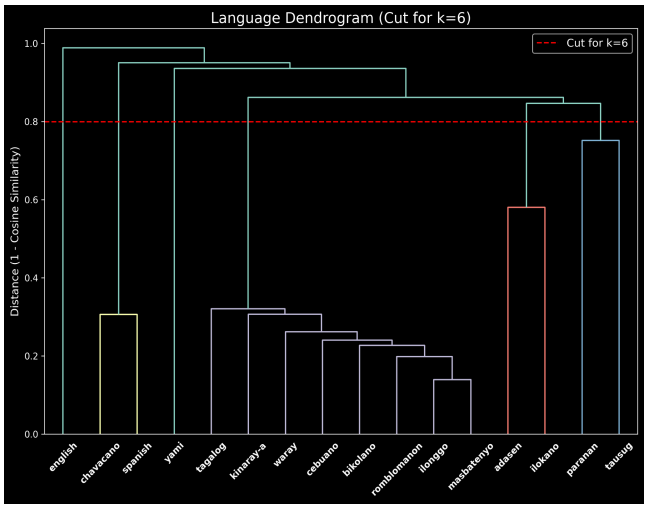

Figure 2. Dendrogram of the 16 Languages

As seen in Figure 2, we have the following clusters when k is equal to 6.

Cluster 1: Chavacano & Spanish

This cluster correctly pairs Chavacano with Spanish. This aligns with the gold standard, as Chavacano is a Spanish-based creole [6], and this pairing was also the strongest non-Philippine link in the similarity matrix (0.69).

Cluster 2: Bikolano, Cebuano, Ilonggo, Kinaray-A, Masbatenyo, Romblomanon, Tagalog, Waray

This is the largest cluster, grouping eight languages. This finding is highly accurate, as all eight are classified as Greater Central Philippine languages in our gold standard [5].

Cluster 3: Adasen & Ilokano

This cluster correctly groups Adasen and Ilokano as they are both classified as Northern Luzon languages in our gold standard [7].

Cluster 4: Paranan & Tausug

This is the only cluster that diverges from the gold standard, grouping Paranan (Northern Luzon) with Tausug (Greater Central Philippine) [7] [5].

Cluster 5: Yami

Cluster 6: English

These clusters correctly isolate Yami and English as distinct families. The dendrogram visualizes this by showing them branching off at a very high distance, confirming their dissimilarity from all other groups.

5. Evaluation

To quantitatively assess the quality of our computationally derived clusters, we evaluated them against an external gold standard. Our predicted model identified 6 clusters (k=6) using hierarchical clustering and silhouette analysis. This was compared against a gold standard we created based on established linguistic classifications from Ethnologue, which resulted in 5 distinct genealogical families:

1. **Greater Central Philippine**
2. **Northern Luzon**
3. **Bashiic (Yami)**
4. **Creole / Spanish-based**
5. **Germanic (English)**

Notably, Chavacano and Spanish were intentionally grouped into a single "Creole / Spanish-based" family for this gold standard, as Chavacano is documented as a Spanish-based creole [6].

We used four standard metrics from *scikit-learn* to compare our 6 predicted cluster assignments against the 5 gold standard labels:

1. **Adjusted Rand Index (ARI)**
2. **Homogeneity**
3. **Completeness**
4. **V-measure**

Our analysis yielded the following results: an ARI of **0.781**, Homogeneity of **0.930**, Completeness of **0.786**, and a V-measure of **0.852**.

These scores indicate a high degree of accuracy. An ARI of 0.781 shows a strong alignment between our predicted clusters and the known language families. The extremely high Homogeneity (0.930) demonstrates that our 6 predicted clusters were pure, meaning they did not incorrectly mix languages from different gold-standard families. This high score is largely because only Cluster 4 (Tausug & Paranan) had incorrectly mixed languages based on the gold standard while the other clusters were accurate. The Completeness score of 0.786 is slightly lower because of

Cluster 4 (Paranan & Tausug) once again, which led to the Greater Central Philippine and Northern Luzon clusters being incomplete. Overall, a V-measure (harmonic mean of homogeneity and completeness) of 0.852 confirms that our n-gram-based methodology was highly effective at capturing and reproducing the established relationships between these languages.

6. Conclusion and Limitations

6.1 Conclusion

This project successfully demonstrated the effectiveness of computational methods in exploring and quantifying the relationships between 16 languages. By creating comparable Bible-based corpora, a feature set of character trigrams, and applying TF-IDF weighting, we were able to compute a textual similarity matrix.

The subsequent hierarchical clustering produced a dendrogram that aligns significantly with established linguistic classifications. Our model correctly identified a large Greater Central Philippine cluster, a distinct Northern Luzon pair, and the strong lexical link between Chavacano and Spanish. It also successfully isolated outliers like Yami and English, demonstrating its ability to distinguish between different language families.

The high evaluation scores, an Adjusted Rand Index (ARI) of 0.781 and a V-measure of 0.852, provide quantitative validation that our computationally derived clusters strongly correlate with our Ethnologue-based gold standard. In summary, the project achieved its objective by showing that even a relatively simple character-level n-gram model can effectively map relationships between languages.

6.2 Limitations

Despite the successful results, this project has several limitations that should be acknowledged:

1. **Corpus Domain:** The most significant limitation is the nature of our corpus. All 16 corpora were sourced exclusively from Bible texts. This is a single, highly specialized domain that does not capture the full breadth of vocabulary used in modern, everyday language. The resulting similarity scores are therefore based only on this specific lexical set.
2. **Method vs. Gold Standard:** A core limitation is the mismatch between our method and our evaluation. Our model measures orthographic (textual) similarity using character n-grams. Our gold standard, Ethnologue, classifies languages based on genealogical (historical) descent. While these two are often correlated (as seen in our high ARI score), they are not the same. This fundamental difference explains both our model's successes and its errors.

Future work could address these limitations by incorporating more diverse corpora from multiple domains and by exploring more advanced features, such as word-level n-grams or word embeddings, to capture deeper semantic relationships.

7. AI Usage Declaration

Anthony Baybayon

AI was only used to aid in learning the essential Python libraries and its functions for the dendrogram and similarity matrix. AI was used to ensure that the code syntax for these statistical approaches in Python are correct.

Dawson Catignas

AI was used to gather pros and cons of the different approaches of creating the feature sets such as using n-grams, levenshtein distance, and others. It was also used as an aid in grammar checking and paraphrasing certain segments in the paper. Each response was then double-checked and validated ensuring that the final decision on what to use was made by me.

Paolo Cruzado

AI was used as an aid in summarizing and comparing various language similarity approaches and verifying preprocessing pipeline descriptions. It also helped clarify technical explanations for improved readability. It was also utilized for grammar checking, sentence restructuring, and formatting consistency across the document. All generated or suggested content was carefully reviewed, validated, and finalized by the authors, ensuring that every methodological and analytical decision remained fully human-directed.

Mariel Tamondong

I did not use AI in this project. My rationale for this is that the results and interpretations following the methodology are straightforward and easy to understand from my point of view. Thus, the derived conclusions and interpretations were directly lifted from the information in the figures and the notebook.

8. References

[1] Life.Church. 2025. Bible.com. YouVersion. https://www.bible.com/

[2] GeeksforGeeks. 2025. Understanding TF-IDF (Term Frequency–Inverse Document Frequency). https://www.geeksforgeeks.org/understanding-tf-idf-term-frequency-inverse-document-frequency/

[3] GeeksforGeeks. 2025. Cosine Similarity. https://www.geeksforgeeks.org/cosine-similarity/

[4] MathWorks. 2025. Cophenetic correlation coefficient. https://www.mathworks.com/help/stats/cophenet.html

[5] SIL International. 2025. Greater Central Philippine (Subgroup 1481). Ethnologue. https://www.ethnologue.com/subgroup/1481/

[6] SIL International. 2025. Chavacano (cbk). Ethnologue. https://www.ethnologue.com/language/cbk/

[7] SIL International. 2025. Northern Luzon (subgroup 2518). Ethnologue. https://www.ethnologue.com/subgroup/2518/