Analyzing Language Endangerment Patterns Across Families and Genera

Term Project Report

Mehmet Sundu

November 6, 2024

Contents

Abstract 1

1 Introduction 2

2 Research Question 2

3 Data Sources 2

4 Methodology 3
4.1 Data Cleaning and Preparation . . . 3
4.1.1 Preprocessing CSV Files . . . 3
4.1.2 Enhanced CSV Loading with Logging . . . 3
4.1.3 Data Standardization . . . 3
4.1.4 Data Merging . . . 3
4.2 Operational Data Layer . . . 3
4.2.1 Families Table . . . 3
4.2.2 Genera Table . . . 3
4.2.3 Languages Table . . . 4
4.2.4 Entity-Relationship Diagram . . . 4
4.3 Analytical Data Layer . . . 4
4.4 ETL Pipeline . . . 4
4.4.1 Extraction . . . 4
4.4.2 Transformation . . . 4
4.4.3 Loading . . . 5
4.4.4 Automation . . . 5
4.5 Data Mart Creation . . . 5
4.6 Visualization . . . 5

5 Analytics Plan 5

6 Analysis and Findings 5
6.1 Endangerment Rates per Family . . . 5
6.1.1 High-Risk Families . . . 5
6.1.2 Interpretation . . . 6
6.2 Endangerment Rates per Genus . . . 6
6.2.1 High-Risk Genera . . . 6
6.2.2 Insights . . . 6
6.3 Geographical Patterns of Endangerment . . . 6
6.3.1 Macroareas with Highest Endangerment Rates . . . 6
6.3.2 Lower Endangerment Rates . . . 6
6.3.3 Moderate Endangerment . . . 6
6.3.4 Interpretation . . . 6
6.4 Visualization of Results . . . 7
6.4.1 Top 10 Genera by Language Endangerment Rate . . . 7
6.4.2 Geographical Distribution of Endangered Languages . . . 8

7 Conclusion 8
7.1 Answer to Research Question . . . 8
7.2 Key Findings . . . 8

8 Recommendations **9**

8.1 Targeted Preservation Efforts . . . 9

8.2 Further Research . . . 9

8.3 Policy and Support . . . 9

9 Challenges and Solutions **9**

9.1 Data Cleaning and Integration . . . 9

9.1.1 Challenges . . . 9

9.1.2 Solutions . . . 9

9.2 Mapping and Visualization . . . 9

9.2.1 Challenges . . . 9

9.2.2 Solutions . . . 9

10 Project Delivery and Reproducibility **10**

10.1 Naming Conventions and Structure . . . 10

10.2 Documentation . . . 10

10.3 Version Control . . . 10

11 References **10**

Abstract

Language endangerment poses a significant threat to cultural diversity worldwide. This project investigates patterns of language endangerment across different language families and genera to identify high-risk groups and geographical areas. By integrating data from Glottolog and the World Atlas of Language Structures (WALS), as well as supplementary datasets from Kaggle, we developed an end-to-end data analysis pipeline using MySQL and Python. The analysis reveals that smaller language families and certain genera exhibit higher rates of endangerment, with notable regional vulnerabilities in South America and Australia. The findings aim to inform targeted preservation efforts and stimulate further research into the factors contributing to language loss.

1 Introduction

Language is a fundamental aspect of human culture and identity. The loss of a language signifies not just the disappearance of words but also the erosion of cultural heritage, knowledge systems, and historical narratives. Currently, many languages are endangered, facing the threat of extinction due to various socio-political, economic, and environmental factors. This project seeks to analyze patterns of language endangerment across different families and genera to understand which groups are most at risk and to identify any geographical patterns that may exist.

2 Research Question

Is there a pattern of endangerment within specific families or genera of languages?

Understanding whether language endangerment is concentrated within particular families or genera can help in prioritizing preservation efforts and in comprehending the underlying causes of language loss.

3 Data Sources

The analysis relies on data from the following sources:

- **Glottolog Database**[1]: Provides comprehensive information on languages, dialects, families, and their classifications.
  - Files used:
    - `languoid.csv`
    - `language.csv`
    - `languages-and-dialects-geo.csv`
- **World Language Family Map Dataset**[2]: A dataset that maps languages to their respective families and provides geographical coordinates.
- **World Atlas of Language Structures (WALS)**[3]: Offers data on the structural properties of languages.
  - Dataset used:
    - `wals-data.csv`[4]

Key Notes:

- The datasets from Kaggle were used to supplement and cross-validate data from Glottolog and WALS.
- Data integration from multiple sources ensured completeness and accuracy.
- All datasets were preprocessed and cleaned to resolve inconsistencies and missing values.

[1] https://glottolog.org/

[2] https://www.kaggle.com/datasets/rtatman/world-language-family-map/data

[3] https://wals.info/

[4] https://www.kaggle.com/datasets/rtatman/world-atlas-of-language-structures?select=wals-data.csv

4 Methodology

4.1 Data Cleaning and Preparation

Data cleaning was a critical step to ensure the integrity and consistency of the datasets.

4.1.1 Preprocessing CSV Files

The raw CSV files had inconsistencies in quoting styles and delimiters, leading to parsing errors. A Python script was developed to standardize quotes and escape delimiters, ensuring consistent formatting across all files.

4.1.2 Enhanced CSV Loading with Logging

To handle malformed lines during CSV loading, an error handler was implemented. This allowed the script to log and skip over problematic lines without halting execution, thus maintaining data integrity.

4.1.3 Data Standardization

Fields such as `status`, `family`, `genus`, and `name` were cleaned by trimming whitespace and converting text to lowercase. Duplicate records were identified and removed based on `Glottocode` to prevent data redundancy.

4.1.4 Data Merging

Datasets were merged on `Glottocode` to consolidate information:

- `language.csv` was merged with `languoid.csv` to include the `status` field.
- The merged dataset was then combined with `languages-and-dialects-geo.csv` and Kaggle datasets to incorporate geographical data.

4.2 Operational Data Layer

A normalized database schema was designed in MySQL to store the operational data, consisting of three main tables:

4.2.1 Families Table

- `Family_ID` (Primary Key)
- `Family_Name`

4.2.2 Genera Table

- `Genus_ID` (Primary Key)
- `Genus_Name`
- `Family_ID` (Foreign Key referencing `Families(Family_ID)`)

4.2.3 Languages Table

- `Language_ID` (Primary Key)
- `Language_Name`
- `ISO_Code`
- `Family_ID` (Foreign Key)
- `Genus_ID` (Foreign Key)
- Other attributes: `Status`, `Macroarea`, `Latitude`, `Longitude`, `Countrycodes`

4.2.4 Entity-Relationship Diagram

```
+------------------+            +-----------------+            +------------------------+
|     Families     |            |      Genera     |            |       Languages        |
+------------------+            +-----------------+            +------------------------+
| *Family_ID* (PK) |<---------+ | *Genus_ID* (PK) |<---------+ | *Language_ID* (PK)     |
| Family_Name      |          | | Genus_Name      |          | | Language_Name          |
+------------------+          | | Family_ID (FK)  |          | | ISO_Code               |
                              | +-----------------+          | | Family_ID (FK)         |
                              |                              | | Genus_ID (FK)          |
                              |                              | | Status                 |
                              |                              | | Macroarea              |
                              |                              | | Latitude               |
                              |                              | | Longitude              |
                              |                              | | Countrycodes           |
                              |                              | +------------------------+
                              |                                   ^
                              |                                   |
                              +-----------------------------------+
```

Figure 1: Entity-Relationship Diagram of the Database Schema

4.3 Analytical Data Layer

A denormalized table, `LanguageEndangerment`, was created to consolidate data for analysis. This table includes measures such as total languages and the count of endangered languages per family and genus.

4.4 ETL Pipeline

An Extract, Transform, Load (ETL) pipeline was implemented using MySQL stored procedures and triggers to automate data processing.

4.4.1 Extraction

Data was extracted from the operational tables: `Families`, `Genera`, and `Languages`.

4.4.2 Transformation

Data was transformed by cleaning and standardizing fields, particularly the `Status` field, to ensure consistent categorization of language endangerment levels.

4.4.3 Loading

Transformed data was loaded into the `LanguageEndangerment` table, preparing it for analytical queries.

4.4.4 Automation

Stored procedures and triggers were used to automate the ETL process upon data changes, ensuring the analytical data remained up-to-date with the operational data.

4.5 Data Mart Creation

Views were created to serve as data marts for analytical queries, simplifying data access for analysis:

- `FamilyEndangermentDataMart`
- `GenusEndangermentDataMart`
- `MacroareaEndangermentDataMart`

4.6 Visualization

Python scripts were used to generate visualizations, including bar charts and an interactive map, to aid in interpreting the analysis results.

5 Analytics Plan

The analytics focused on:

- Calculating endangerment rates across families and genera.
- Identifying families and genera with the highest endangerment rates.
- Analyzing geographical patterns of language endangerment.
- Visualizing the findings for enhanced interpretation.

6 Analysis and Findings

6.1 Endangerment Rates per Family

The analysis revealed that many language families exhibit a high percentage of endangered or extinct languages.

6.1.1 High-Risk Families

- Families such as **Aikaná**, **Alacalufan**, **Andoke**, and **Atakapa** show **100% endangerment**, indicating that all known languages in these families are either endangered or extinct.
- Larger families such as **Arawakan** and **Pama-Nyungan** also display significant endangerment rates of 78.57% and 54.21%, respectively.

6.1.2 Interpretation

Smaller language families or those with limited resources tend to have higher endangerment rates, possibly due to less exposure and support for preservation efforts.

6.2 Endangerment Rates per Genus

6.2.1 High-Risk Genera

- **Athapaskan**: 95.45% endangerment.
- **Yuman**: 85.71% endangerment.
- Genera such as **Northern Iroquoian** and **California Uto-Aztecan** exhibit endangerment rates exceeding 80%.

6.2.2 Insights

Certain genera within larger families are disproportionately vulnerable, suggesting that localized factors such as regional policies, socio-economic conditions, and cultural assimilation pressures significantly contribute to language loss.

6.3 Geographical Patterns of Endangerment

6.3.1 Macroareas with Highest Endangerment Rates

- **South America**: 68.27% endangerment rate.
- **Australia**: 55.49% endangerment rate.
- **North America**: 50.69% endangerment rate.

6.3.2 Lower Endangerment Rates

- **Africa**: 7.29% endangerment rate.
- **Eurasia**: 23.13% endangerment rate.

6.3.3 Moderate Endangerment

- **Papunesia**: 8.93% endangerment rate.

6.3.4 Interpretation

Geographic regions with high linguistic diversity and substantial indigenous populations are more susceptible to language endangerment, often due to historical colonization, cultural assimilation policies, and lack of support for minority languages.

6.4 Visualization of Results

6.4.1 Top 10 Genera by Language Endangerment Rate

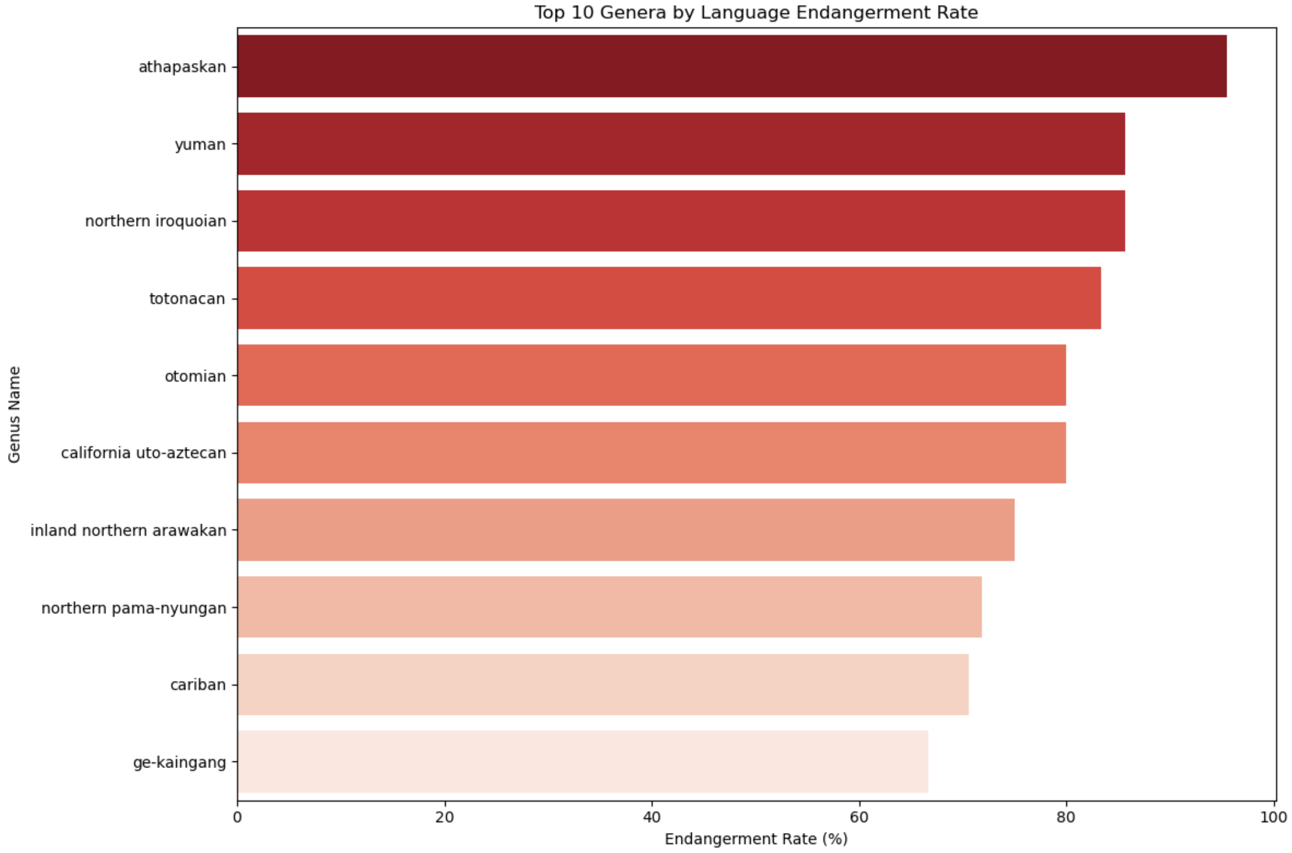

Figure 2: Top 10 Genera by Language Endangerment Rate

Genera such as **Athapaskan** and **Yuman** are prominently featured, highlighting their critical status and need for immediate preservation efforts.

6.4.2 Geographical Distribution of Endangered Languages

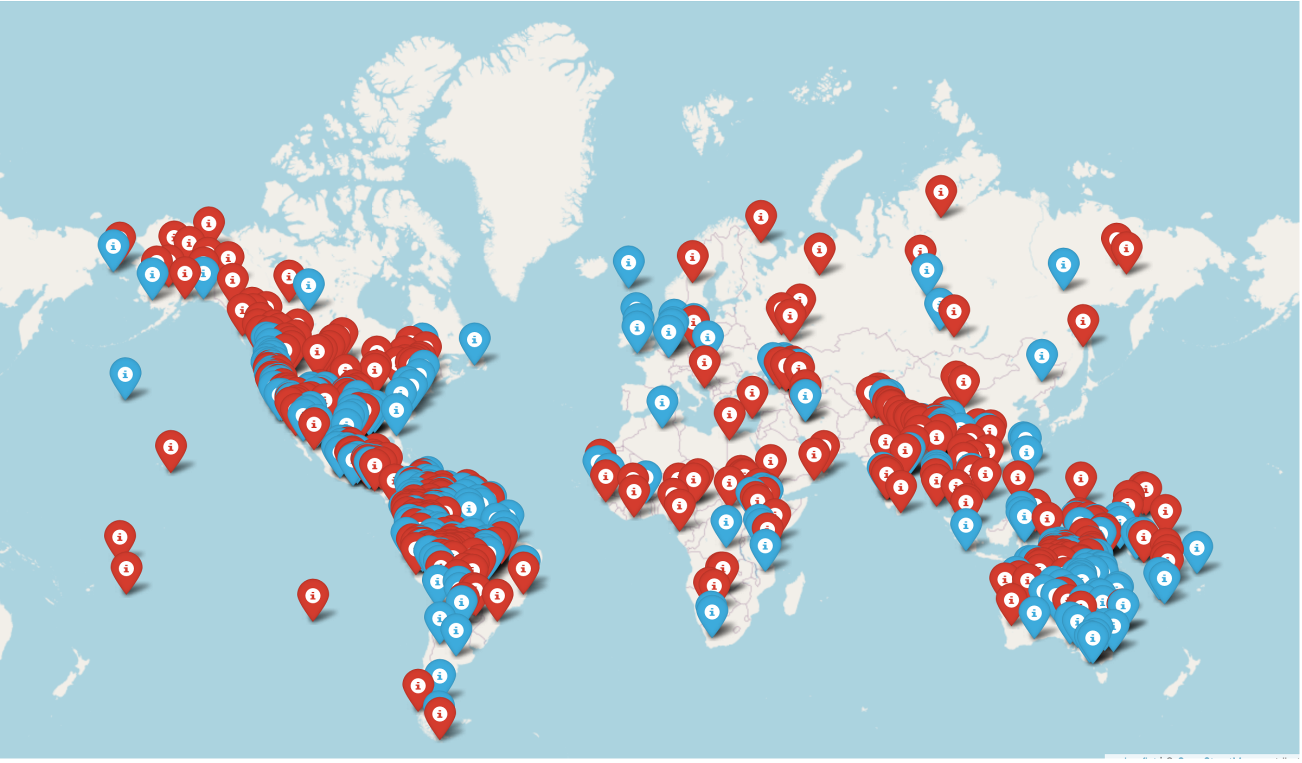

Figure 3: Geographical Distribution of Endangered Languages

This map visualizes the global distribution of endangered languages, emphasizing regions with higher concentrations of at-risk languages. The screenshot represents the interactive map generated using Folium, showcasing the spatial patterns of language endangerment.

7 Conclusion

7.1 Answer to Research Question

The analysis confirms that there is a discernible pattern of endangerment within specific families and genera of languages. Smaller families and certain genera within larger families exhibit higher rates of language endangerment.

7.2 Key Findings

- **Smaller Families at Higher Risk**: Families with fewer languages are more susceptible to endangerment due to limited resources and lesser global recognition.
- **Regional Vulnerabilities**: Macroareas like South America and Australia have a higher concentration of endangered languages, indicating regional factors at play.
- **Genus-Level Insights**: Certain genera are disproportionately affected, highlighting the need for genus-specific preservation strategies.

8 Recommendations

8.1 Targeted Preservation Efforts

Focus language preservation initiatives on highly endangered families and genera, particularly in the most affected macroareas, to maximize the impact of conservation efforts.

8.2 Further Research

Investigate the cultural, political, and ecological factors contributing to language endangerment in high-risk areas to develop informed and effective preservation strategies.

8.3 Policy and Support

Encourage local and international support for language revitalization programs that involve native speaker communities, ensuring that preservation efforts are culturally sensitive and sustainable.

9 Challenges and Solutions

9.1 Data Cleaning and Integration

9.1.1 Challenges

- Inconsistencies in CSV formatting, leading to parsing difficulties.
- Missing or malformed data entries, affecting data completeness.
- Duplicate records causing potential data skew.

9.1.2 Solutions

- Implemented preprocessing steps to standardize data formats.
- Enhanced error handling during data loading to manage malformed entries.
- Cleaned and normalized data fields, removing duplicates to ensure data integrity.

9.2 Mapping and Visualization

9.2.1 Challenges

- Handling missing geographical coordinates for accurate mapping.
- Ensuring accurate plotting of languages on the map to reflect true geographical distribution.

9.2.2 Solutions

- Consolidated geographical data from multiple sources to fill in missing coordinates.
- Handled missing values by prioritizing more reliable data sources.
- Utilized robust mapping libraries (e.g., Folium) for accurate and interactive visualizations.

10 Project Delivery and Reproducibility

10.1 Naming Conventions and Structure

The project follows a structured directory layout to enhance reproducibility and ease of navigation:

```
Term1/
 data/
    Families.csv
    Genera.csv
    Languages.csv
 sql_scripts/
    create_database.sql
    create_tables.sql
    import_data.sql
    create_etl_procedures.sql
    create_triggers.sql
    create_views.sql
    analysis_queries.sql
    create_materialized_views.sql
 python_scripts/
    data_cleaning.py
    data_visualization.py
 docs/
    README.md
 results/
    endangerment_rates_per_family_genus.csv
    language_endangerment_map.html
    analysis_report.txt
 .git/
```

10.2 Documentation

- **README.md**: Contains setup instructions, project overview, and explanations of each component.
- **Comments in Scripts**: All SQL and Python scripts include comments explaining their functionality and usage.

10.3 Version Control

The entire project is tracked using Git, with a detailed commit history for transparency and collaboration purposes.

11 References

- Glottolog: https://glottolog.org/
- WALS: https://wals.info/

- World Language Family Map Dataset: https://www.kaggle.com/datasets/rtatman/world-language-family-map/data
- WALS Data on Kaggle: https://www.kaggle.com/datasets/rtatman/world-atlas-of-language-select=wals-data.csv
- MySQL Documentation: https://dev.mysql.com/doc/
- Pandas Documentation: https://pandas.pydata.org/docs/
- Folium Documentation: https://python-visualization.github.io/folium/