From Data Points to Polygons: An Innovative Approach to Geolinguistic Mapping

Hiromi Oda, Tokyo Online University, Japan

The IAFOR International Conference on Arts & Humanities in Hawaii 2025
Official Conference Proceedings

Abstract

This paper introduces a software tool that implements an innovative type of geolinguistic map, originally proposed by Hideo Suzuki in the 1980s. The software, developed using the R language, automates the map creation process based on word forms and geographic coordinates. Unlike traditional maps, this new type not only displays the locations of word occurrences but also outlines their distribution areas, allowing us to visualize how similar word forms are used in geographically distant regions, such as parts of India and Scandinavia. In Suzuki's time, collecting word data required extensive library research, and mapping involved manually plotting data on large physical maps. Today, with advancements in computing and AI, this process can be fully automated. The software analyzes word similarities using metrics like edit distance, groups the words accordingly, and generates two layers for each group: one displaying the geographic data points, and the other showing a smoothed polygon encircling the distribution area of each group. Both layers can be toggled independently, and multiple-word groups can be displayed simultaneously for comparison. The map is interactive, allowing users to zoom in or out and quickly focus on specific regions or word groups. Thanks to recent advances in AI and the availability of online linguistic resources, data collection has also become significantly more efficient. This tool opens up new possibilities for studying language distribution patterns on a global scale. This software tool is implemented using R and its geographic computing packages, all of which are generously available under open licenses.

Keywords: Geolinguistic Map, GIS (Geographic Information System), R Language, Hideo Suzuki

Introduction

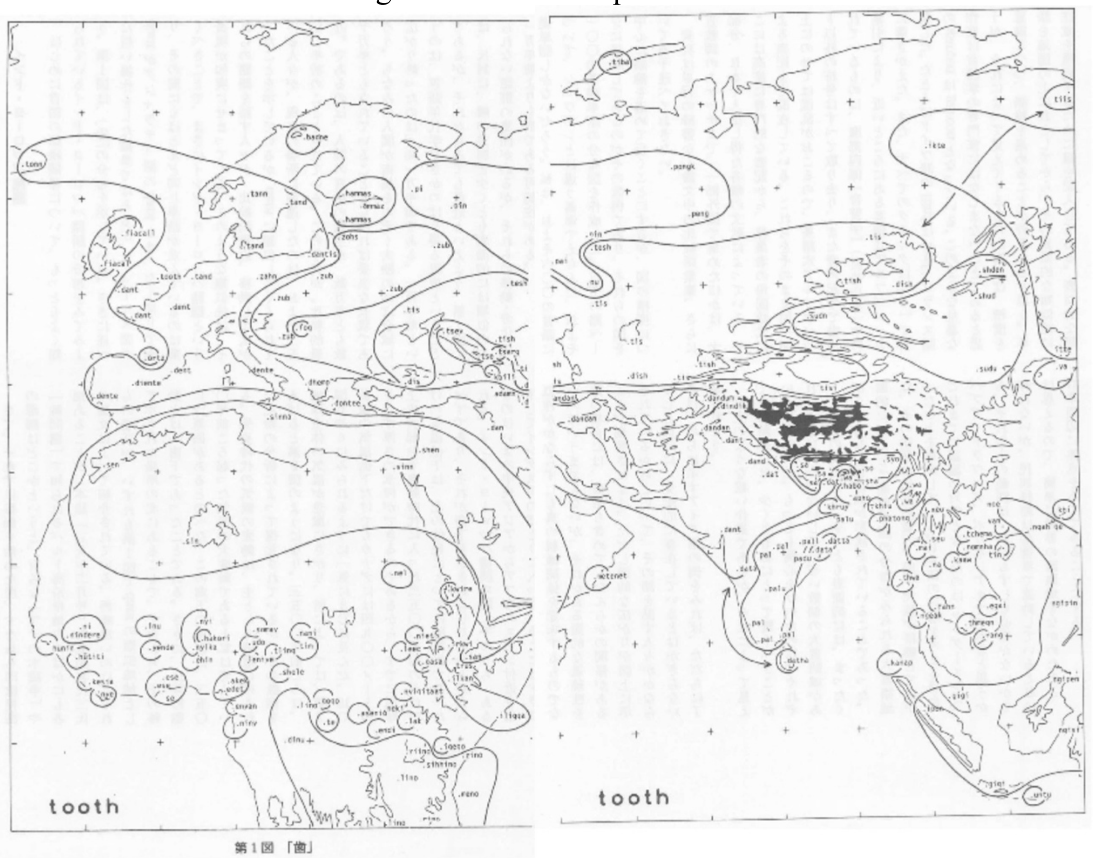
Hideo Suzuki, a climatologist, hypothesized that long-term climate changes drive human migrations, which, in turn, influence language evolution. To analyze this further, he spent a decade collecting Swadesh-style basic vocabulary items from approximately 600 languages and plotted word forms on large maps. A unique aspect of his maps is that he drew encircling lines to indicate the distribution areas of similar word forms.

Figure 1 presents a sample map for the word "tooth" (Suzuki, 1990). By following the encircling line that covers the lower part of Scandinavia, one can trace a group of words stretching as far as parts of India. Suzuki simply collected word forms from various languages and plotted them on a large map, making it easy to observe connections between geographically distant regions in the East and West, even without prior knowledge of the Indo-European language family.

Although Suzuki's map-making process was laborious and time-consuming, advancements in computer technology and the availability of online resources now make it feasible to implement this method computationally. This project aims to reproduce Suzuki's maps using the R programming language and its GIS packages. Since some implementation details will be discussed elsewhere, this report primarily focuses on the usage and benefits of the automatic mapping tool, the main outcome of this project.

The remainder of this paper is structured as follows: first, Suzuki's original map-making process is summarized. Next, the implementation principles are explained. Finally, two sample datasets are used to verify the implementation.

Figure 1: Suzuki Map "Tooth"



Suzuki Map Creation Procedure

In this section, we summarize the procedure for creating the Suzuki Map, following Suzuki (1990). The process can be divided into three major steps: (1) collecting basic vocabulary data, (2) plotting the data on a map, and (3) grouping similar words.

(1) Collection of Basic Vocabulary Data

Basic vocabulary data is gathered from a variety of sources, including dictionaries and records left by missionaries. The term "basic vocabulary" refers to a list proposed by Morris Swadesh (Swadesh, 1950), which includes words related to body parts (e.g., "eye," "ear") and common concepts (e.g., "come," "cold") that are considered historically stable. Lexicostatistics assumes that these words change at a constant rate over time. The collected data comes in various formats, and for consistency, all transcriptions are standardized into English-style Romanization. Suzuki reports collecting data on approximately 100 modified Swadesh list items from about 600 languages worldwide.

(2) Map Creation

A large world map, approximately the size of a single bed mattress, is prepared, and the collected vocabulary data is plotted onto it.

(3) Grouping Similar Words

This step is the defining feature of the Suzuki Map. Regarding lexical similarity assessment, Suzuki states:

> The determination of similarity is subjective, but I believe that most readers will find the examples presented below to be similar. In fact, even words that are deemed dissimilar here may be considered similar by linguists. (p. 16)

Thus, the judgment of similarity is primarily intuitive. If others attempt to recreate this map, a key concern is whether their assessments will align with Suzuki's, raising issues of consistency and the need for methodological refinement.

Implementation of Suzuki Map Using R and GIS Packages

This section outlines the approach used to reproduce the Suzuki Map.

Basic Approach to Reproduction

Certain aspects of the original Suzuki Map cannot be implemented exactly as they were, necessitating adaptations for a computer-based application. Additional steps are required to ensure compatibility with modern mapping tools while preserving the core methodology.

System Used

To implement the Suzuki Map digitally, we utilize the R programming language along with specialized packages for map creation. R is well-suited for visualization and offers a variety of user-friendly geographic information system (GIS) packages essential for linguistic mapping. Below, we introduce two key packages used in this project.

The sf Package

The sf package is a toolkit for handling geospatial data in R. The name sf refers to Simple Features, a standard format for geospatial data representation. This package enables seamless integration of Simple Features within the R environment.

Simple Features is an internationally recognized standard that provides a uniform representation of geospatial data. It defines models for concisely representing geographical entities such as points and regions. This project adopts the following core concepts:

- Point: A single geographic location
- Linestring: A sequence of connected line segments representing features such as rivers or roads
- Polygon: A closed area defined by a collection of points

By using the sf package, these objects can be manipulated in a manner similar to conventional data frames in R.

The tmap Package

The tmap package is a toolkit for creating and visualizing maps in R. While it supports the generation of static maps, its key functionality in this project is its ability to create interactive maps. This feature allows for dynamic exploration of linguistic data distributions.

Data Collection

Data collection is a fundamental challenge in linguistic mapping. However, the availability of online linguistic datasets and AI tools such as ChatGPT significantly reduces the workload. In this project, sample data was exclusively collected using ChatGPT for verification purposes.

Sample Data

To evaluate whether the proposed method produces results comparable to those of Suzuki, we prepared a sample dataset on the Indo-European language family. The data collection process followed these steps:

1. *Generating a list of Indo-European languages* and obtaining the approximate latitude and longitude of their central regions where the languages are spoken.
2. *Identifying and listing words for "mouth" and "tooth"* in each language. While Suzuki's Figure 1 focuses on tooth, this project verifies the approach using two sample datasets.

The collected data underwent minimal manual editing, primarily for formatting into CSV files.

The resulting dataset for the words "mouth" includes 58 lexical items. Table 1 presents the first few entries of this sample data. The data file must include columns for the word form, longitude, and latitude. Additionally, an optional column labeled "Language" below can be included to provide supplementary information. This information will be displayed in a small pop-up window when a data point is clicked on the generated map.

Table 1: Sample Data Prepared in the CSV Text Format

```
Word, x, y, Language
mond,18.4241,-33.9249,Afrikaans
gojë,19.8187,41.3275,Albanian
beran,44.4991,40.1792,Armenian
aho,-2.9340,43.2630,Basque[1]
rot,27.5615,53.9045,Belarusian
mukh,90.4125,23.8103,Bengali
usta,18.4131,43.8563,Bosnian
genou,-1.6778,48.1173,Breton
usta,23.3219,42.6977,Bulgarian
```

Word Grouping

The words listed in the *Word* column of the sample dataset are classified into groups. While Suzuki originally performed this classification intuitively, a more objective and consistent method is preferable to ensure reproducibility. Therefore, we employ a widely recognized approach known as *edit distance* (also referred to as *Levenshtein distance*) to calculate linguistic similarity and categorize words accordingly.

Edit distance measures the minimum number of operations (insertion, deletion, or substitution) required to transform one string into another. A smaller edit distance indicates a higher degree of similarity between words.

R provides the adist function for computing edit distances, which we use to calculate the distances between words in the dataset. For example, calculating the distances between six sample words yields the following results.

Table 2: Word Distance Matrix Based on Edit Distance Measure

```
[1] "boca"    "bejeth" "usta"    "sta"     "mund"
"mond"

[,1] [,2] [,3] [,4] [,5] [,6]
[1,]    0    5    3    3    4    3
[2,]    5    0    5    5    6    6
[3,]    3    5    0    1    4    4
[4,]    3    5    1    0    4    4
[5,]    4    6    4    4    0    1
[6,]    3    6    4    4    1    0
```

Table 2 shows that the distances between words 3 and 4, as well as words 5 and 6, are both 1. By setting the distance threshold to 1, we obtain the following word groups:

[1] *Basque* is not an Indo-European language, but it will be left as is since this is part of the results produced by ChatGPT.

Table 3: Word Groups of the Sample Word List With the Threshold 1

```
[1] "boca"

[2] "bejeth"

[3] "usta" "sta"

[4] "mund" "mond"
```

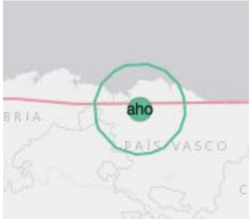
POINT Objects

Words that do not belong to any group remain as *point objects*. In fact, all data points are initially treated as point objects during the initial data import phase.

In Suzuki's original map, even isolated data points are enclosed by boundary lines. Following this convention, we also enclose point objects within boundary lines. This approach highlights when a data point does not belong to a surrounding word group, as demonstrated in Figure 2. Although group membership can be distinguished by color, enclosing individual points with boundary lines enhances visual clarity.

In geospatial terminology, the region surrounding an object is called a *buffer*. Following this practice, we refer to the line that encircles an object as the *buffer line*. In this project, we consistently apply a 30 km buffer around objects and perform smoothing operations on the buffer lines.

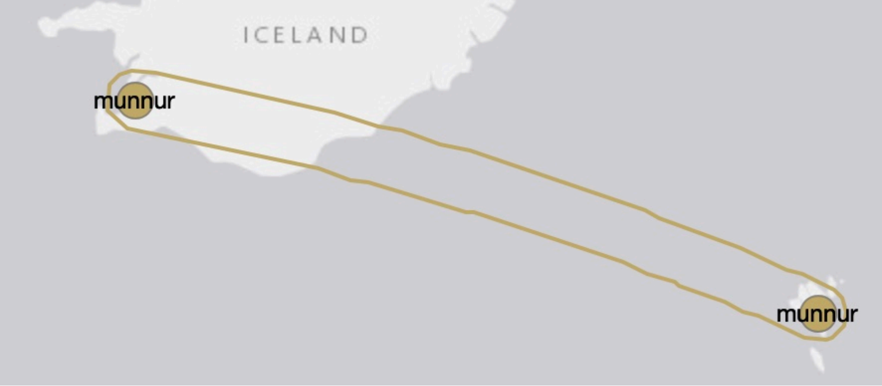
Figure 2: A POINT Object With the Buffer Line



LINESTRING Objects

When word grouping results in a group containing exactly two words, the connection is represented as a *linestring object*, as illustrated in Figure 3. The two words are connected by a line, and a 30 km buffer is applied around it.

Figure 3: A LINESTRING Object



POLYGON Objects

If a group consists of three or more words, it is represented as a *polygon object*. The same 30 km buffer and smoothing operations are applied, ensuring that the display renders data points along with their enclosing buffer lines.

However, when multiple groups exist and buffer lines overlap, users may wish to hide certain buffer lines and display only the data points of a specific group. To accommodate this, we separate the *data points layer* and the *buffer line layer*, allowing users to toggle their display independently.

Additionally, to maintain consistency, each data point and its corresponding buffer line are assigned the same color and are managed as a paired list structure.

To demonstrate the use of these two layers, we examine the following group of words from the sample dataset.

```
"usta" "usta" "usta" "sta" "usta"
"usta" "usta"  "sta"  "usta"
```

In Figure 4, both the *data points layer* and the *buffer line layer* are displayed, demonstrating that the buffer lines effectively enclose the corresponding word groups. If necessary, the display can be adjusted to show only data points (Figure 5) or only the buffer lines (Figure 6).

Figure 4: A POLYGON Object

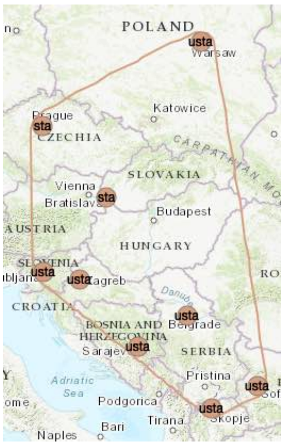


Figure 6: Only the Data Points

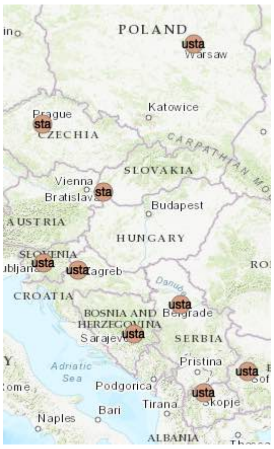
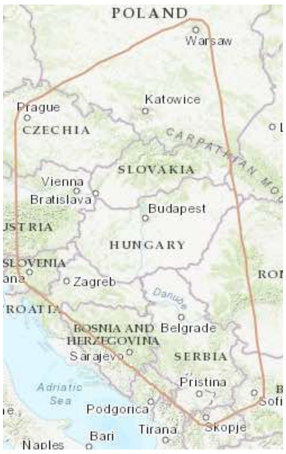


Figure 5: Only the Buffer Line



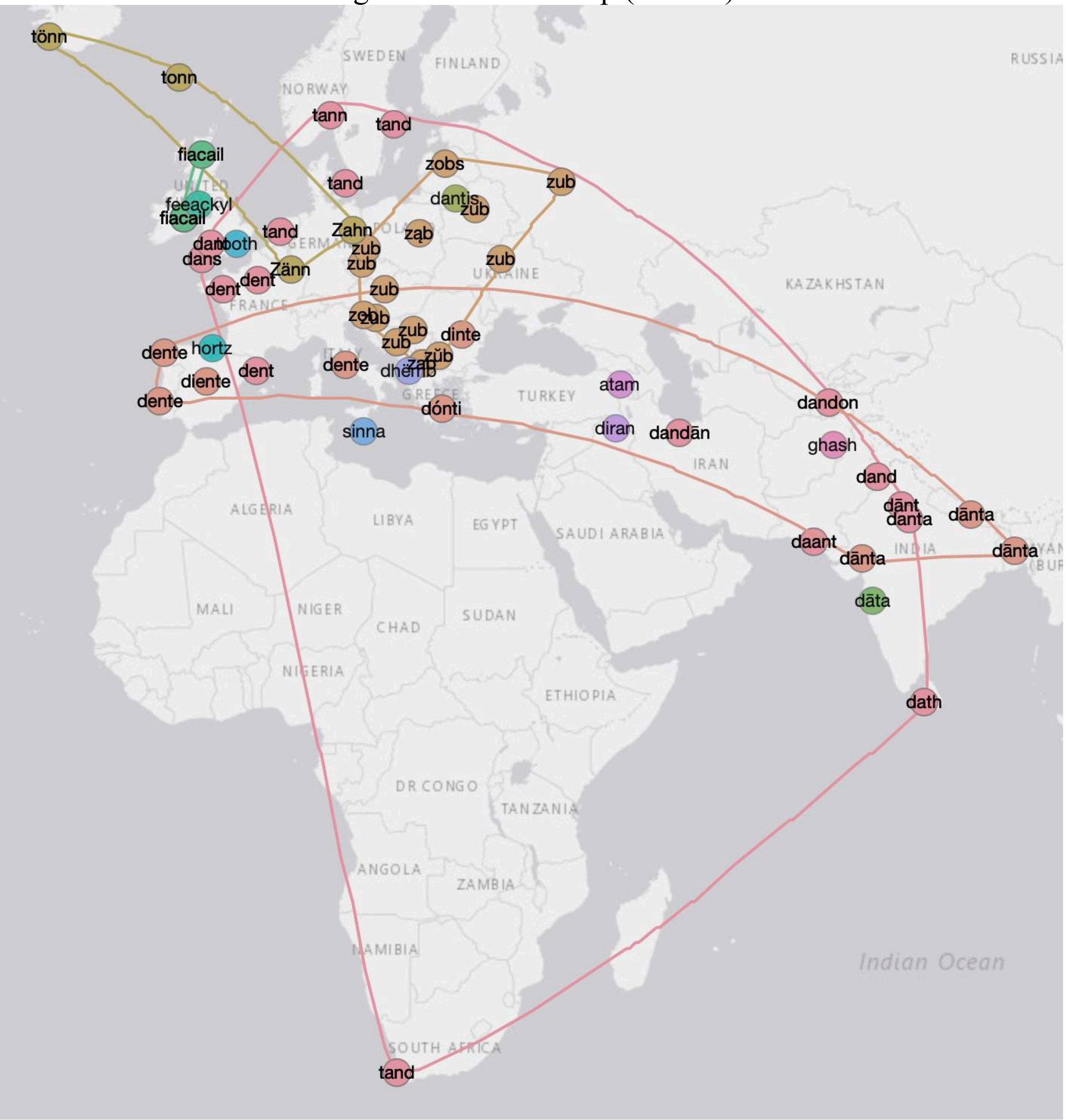
Final Form: Reproducing the Suzuki Map Using R and GIS Packages

The process of visualizing the results of word grouping based on edit distance and faithfully reproducing Hideo Suzuki's map on a computer screen has been outlined above.

In Suzuki's original work, the first example introduced was the word *"tooth."* We collected a sample dataset using the same procedure for *"tooth."* The grouping process can be controlled by adjusting the edit distance threshold. For example, when words with an edit distance of 2 or less are grouped together, 15 groups are formed. Figure 7 presents these groups using the *tmap* package. The resulting map reveals linguistic connections spanning vast regions from east to west.

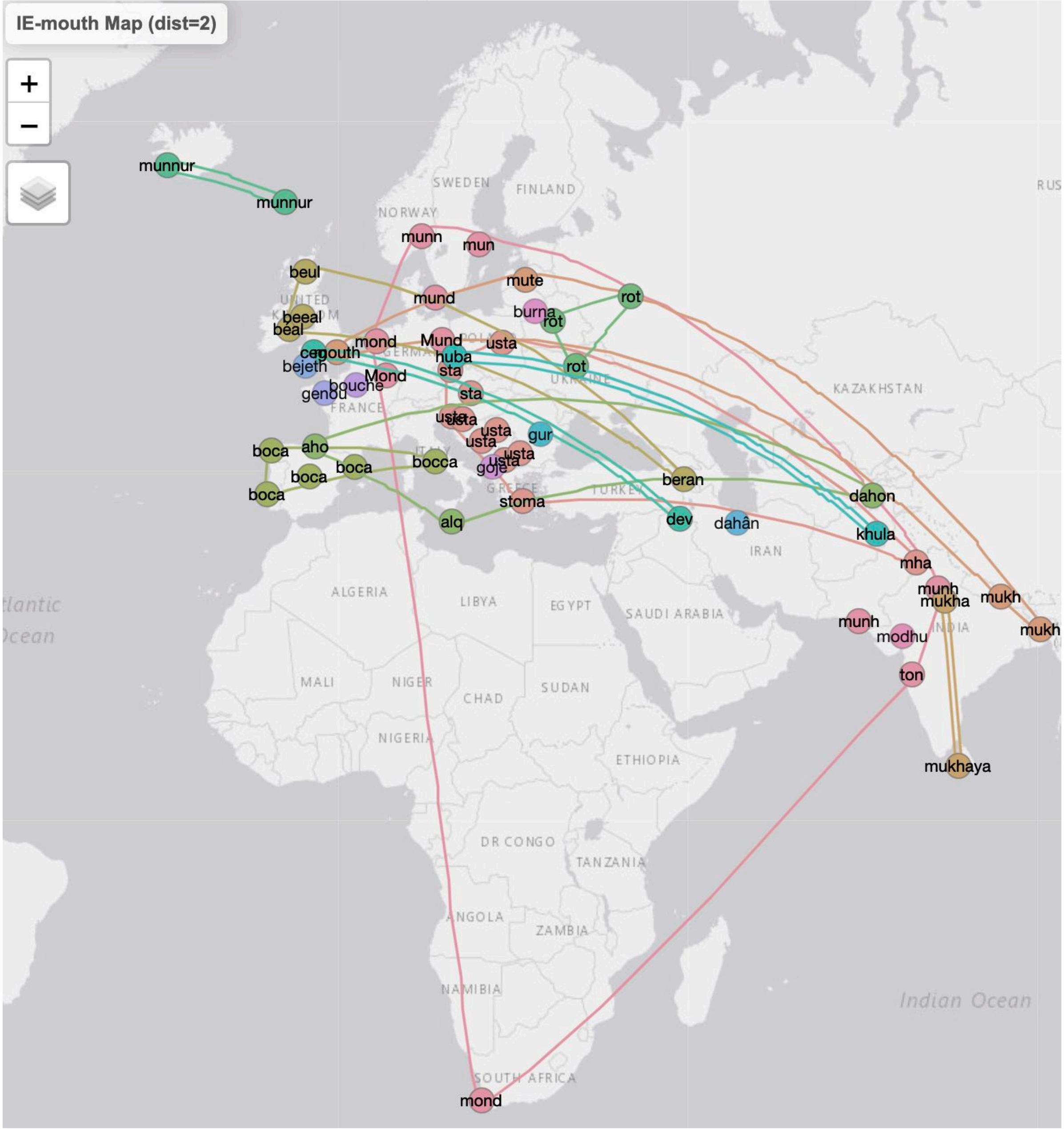
The largest group extends as far as *Afrikaans*, historically linked to Dutch due to colonial influences, demonstrating a strong connection with Dutch and related languages. This group also spans a vast area, connecting regions from *India to the Scandinavian Peninsula*. This visualization replicates Suzuki's original map, illustrating the expansive distribution of basic vocabulary across a wide geographical range.

Figure 7: "Tooth" Map (ED = 2)



We can draw a similar map using the data for "mouth," as shown in Figure 8, to verify the similar patterns of word groups that have wide areas of distribution.

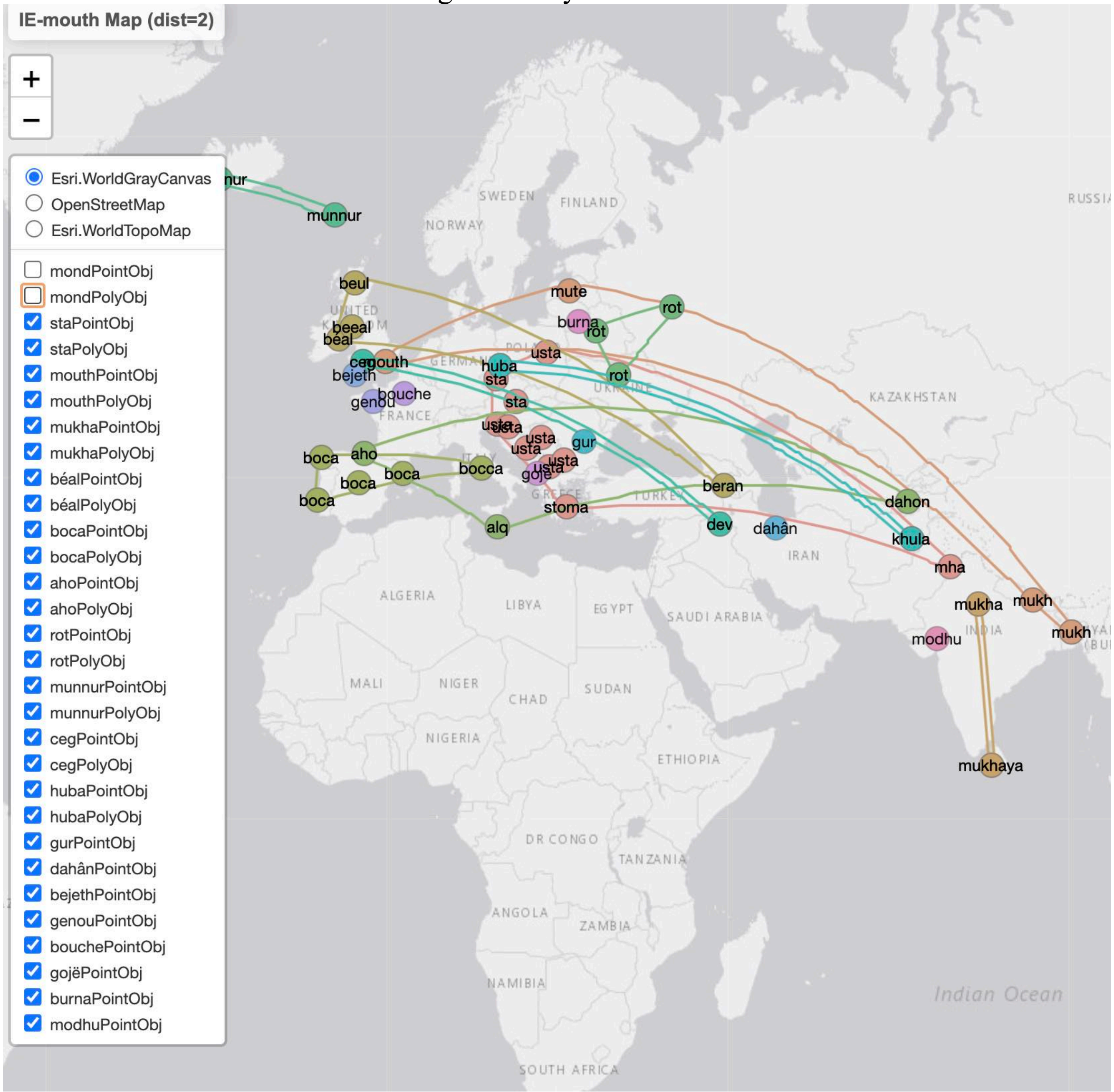
Figure 8: "Mouth" Map



Zooming, Scaling, and Layer Control

When displaying maps in the interactive mode of the *tmap* package, clicking on the icon resembling stacked diagonal sheets in the top-left corner reveals a list of layers. Users can manipulate the map by zooming in and out or adjusting the position using the + and - icons, or through mouse operations.

Figure 9: Layer Control



IE-mouth Map (dist=2)

● Esri.WorldGrayCanvas
○ OpenStreetMap
○ Esri.WorldTopoMap

☐ mondPointObj
☐ mondPolyObj
☑ staPointObj
☑ staPolyObj
☑ mouthPointObj
☑ mouthPolyObj
☑ mukhaPointObj
☑ mukhaPolyObj
☑ béalPointObj
☑ béalPolyObj
☑ bocaPointObj
☑ bocaPolyObj
☑ ahoPointObj
☑ ahoPolyObj
☑ rotPointObj
☑ rotPolyObj
☑ munnurPointObj
☑ munnurPolyObj
☑ cegPointObj
☑ cegPolyObj
☑ hubaPointObj
☑ hubaPolyObj
☑ gurPointObj
☑ dahânPointObj
☑ bejethPointObj
☑ genouPointObj
☑ bouchePointObj
☑ gojëPointObj
☑ burnaPointObj
☑ modhuPointObj

To focus on the distribution of word groups from East to Europe, we can hide the groups extending into Africa (see Figure 9). This visualization highlights the presence of groups connecting regions from *India to Europe, England*, and the *Scandinavian Peninsula*. Through this process, the relationships within the Indo-European language family can be visually confirmed.

This project adopted a methodology that minimizes manual intervention, relying solely on *ChatGPT* for data collection. The process involved preparing CSV files for data input and evaluating how accurately Suzuki's linguistic map could be reproduced and validated. By following the outlined steps, we successfully generated a *Suzuki Map*, allowing for broad geographic analysis and the identification of long-distance linguistic connections.

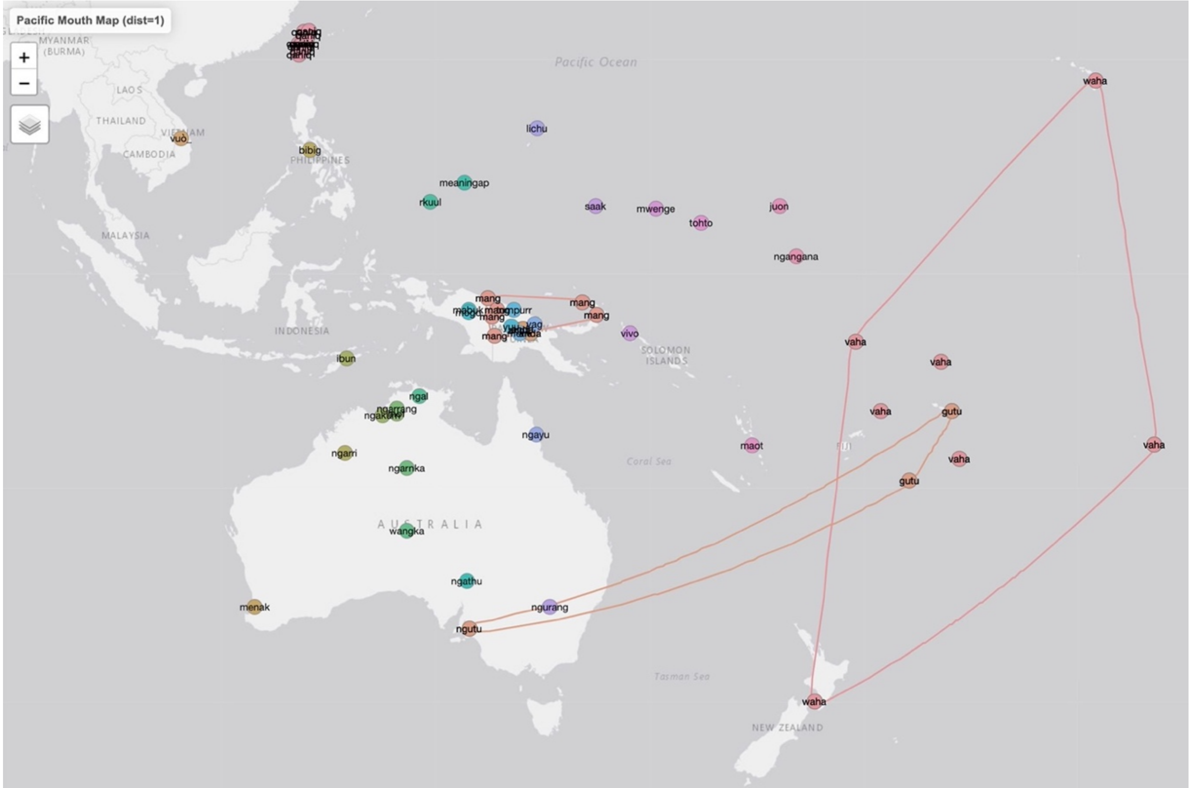
The primary objective of this project was to reconstruct and verify the *Suzuki Map*, and the results successfully support Suzuki's claims while introducing minor refinements. The interactive maps generated in this study can be exported as *HTML files*. These maps can be displayed and manipulated using a web browser, enabling further analysis and exploration.

Languages in the Pacific Ocean

To apply the *Suzuki Map* to languages in the Pacific Ocean, about 100 languages were selected, roughly following the size of the number of speakers. *ChatGPT* found words for *"mouth"* in about 80 languages. The *Edit Distance (ED)* thresholds of 1 and 2 were used to create interactive maps.

Figure 10 shows the *Suzuki Map* with *ED = 1*. It is possible to see the connections between languages in *New Zealand, Tuvalu, Tahiti,* and *Hawaii*. Additionally, an *aboriginal language*, *Kaurna*, spoken in South Australia, is connected to *Samoan* and *Tongan*.

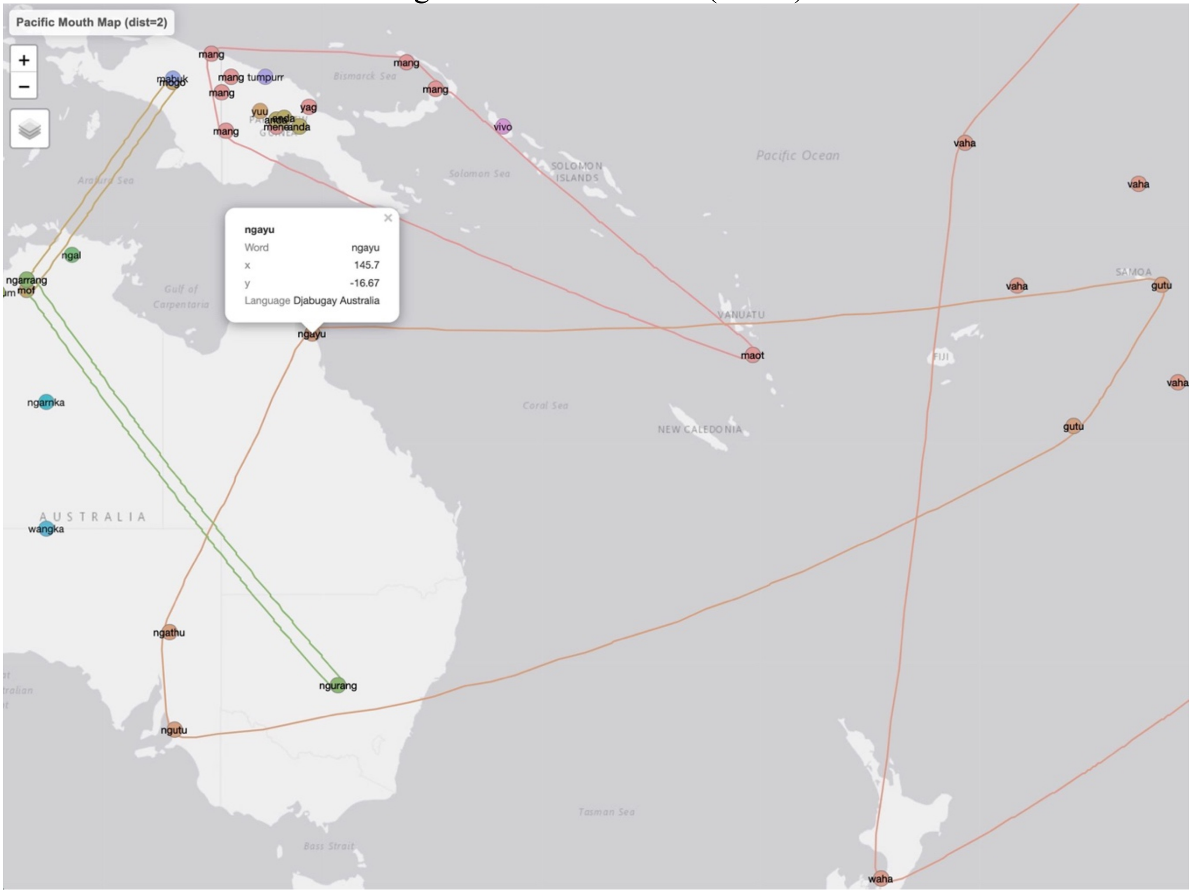
Figure 10: Pacific Ocean (ED=1)



ED=1 does not capture the similarity between *Kaurna's "ngutu"*, *"ngathu"* of the neighboring *Adnyamathanha* language, and *"ngayu"* of the *Djabugay* language. However, if we choose *ED=2*, a larger group can be formed, as shown in Figure 11. Although the results here do not align with Suzuki's original work due to the differing data points, the usefulness of the map is evident.

Note the pop-up information when a data point is clicked in the window. It is possible to add notes in the CSV data for reference while inspecting the interactive map.

Figure 11:Pacific Ocean (ED=2)



ngayu
Word ngayu
x 145.7
y -16.67
Language Djabugay Australia

Conclusion

This paper summarizes a project that aimed to implement the unique geolinguistic map proposed by *Hideo Suzuki* in the 1980s. Interactive *Suzuki maps* on a computer replicated Suzuki's basic results while significantly enhancing the usability of his map. These improvements allow users to zoom in, zoom out, and compare multiple layers of groups simultaneously. The *SF* and *tmap* packages in the *R* language made it possible to implement the map in a relatively straightforward manner.

Through the implementation, a couple of enhancements were proposed. First, there is a need for an objective similarity measure for grouping. Although the *Edit Distance* measure was used in this report, there is room for improvement in this respect.

Second, the use of *ChatGPT* and other possible online resources can greatly assist in preliminary data collection. While the results may not be fully reliable, having an initial working data set facilitates the initiation of a data collection project.

Acknowledgment

I would like to express my gratitude to those who provide R and various packages as open-source software. In particular, for this project, I am especially grateful for the availability of the user-friendly *sf* and *tmap* packages. This tool is also planned to be publicly released after further validation.

References

Muenchow, R., Lovelace, J., & Nowosad, J. (2019). *Geocomputation with R*. Routledge. https://geocompr.robinlovelace.net/

Suzuki, H. (1987). Ethnic migration and language distribution. *Gekkan Gengo, 17*(7), 24–40. (In Japanese)

Suzuki, H. (1990). *Climate change motivating language change: An approach by linguistic dendrochronology*. NHK Publishing Co. (In Japanese)

Swadesh, M. (1950). Salish internal relationships. *International Journal of American Linguistics, 16*(4), 157–167. https://doi.org/10.1086/464084

Contact email: oda.hiromi@internet.ac.jp