

Addis Ababa Institution of Technology

Department of pre-Engineering

Computer programming

Group assignment

Members ID.No

1. Ygeramu Sisay UGR/9603/16
2. Yohannes Alemneh UGR/8514/16
3. Eyob Dereje UGR/8159/16
4. Getasew Yayew UGR/5578/16
5. Michael Ashenafi UGR/4141/16

Submitted to:

Tech. Lemlem Hagos

Submission date:

August 12, 2024

Acknowledgment

Frist of all, we want say thank you for everyone who participated in this project mostly all the group members.

We also want to say thank you our programing teacher Tech. Lemlem Hagos because despite its toughness, the assignment is ingesting and it allows us to practice things that we have been learning in the class. It also tests our programing skill hardly.

And also our representative Yafiet Gelila

In general, small droplet of water can make a difference, so everyone who participate in this project directly or indirectly, we want to say

Thank you!

# Text Processing Mini Project Report

## 1. Introduction

This mini project focuses on text processing for a pair of Ethiopian languages from the Semitic families. The selected language pair for this project is Amharic and Tigrigna. The tasks include text pre-processing, word frequency analysis, grapheme to phoneme conversion, and statistical analysis of phonemic distribution using g2p algorithm.

## 2. Project Tasks and Implementation

### 2.1 Getting Text Data

We collected parallel text data for the selected language pair from Amharic and Tigrigna. The data consists of texts on similar topics in both languages.

The selected texts are:

Amharic : በአዲስ አበባ እየተመዘገበ የመጣው ሁለንተናዊ ለውጥ እያደገ ካለው የገቢ አሰባሰብ ሥርዓት የሚመነጭ መሆኑን ከንቲባ አዳነች አቤቤ ገለጹ። ከከተማ እስከ ወረዳ ያለው የአዲስ አበባ ከተማ አስተዳደር አመራሮችን ያሳተፈ የ2016 አፈጻጸም ግምገማ እና የ2017 ዕቅድ ውይይት መድረክ እየተካሄደ ይገኛል። የአዲስ አበባ ከተማ ከንቲባ አዳነች አቤቤ መድረኩን ሲያስጀምሩ እንደገለጹት፥ በአዲስ አበባ ሁለንተናዊ ለውጥ እንዲመዘገብ ካስቻሉ በርካታ ጉዳዮች መካከል ገቢ ከማሳደግ አንጻር የተከናወኑ ውጤታማ ተግባራት ወሳኝ ሚና አላቸው።

Tigrigna : ብመሰረት እቲ ንሱ ዝበሎ እቲ ኣብ ኣድዲስ ኣቤባ ዘሎ ዅነታት ለውጢ ካብቲ እናወሰኸ ዝኸይድ ዘሎ ኣገባብ ምእካብ ኣታዊ ዝመነጨ ኢዩ ። ባጀት 2016ን ባጀት 2017ን ኣብ ቤት ተወከልቲ ይመያየጡ ኣለዉ። ከንቲባ ኣድኣድነኽ ኣቢበ እቲ ዋዕላ ኣብ እተጀመረሉ እዋን ኣታዊ ንምርካብ ዚግበር ውጽኢታዊ ንጥፈታት ሓደ ኻብቲ ኣብ ኣድዲስ ኣቤባ ሓፈሻዊ ለውጢ ንኺግበር ዘኽኣሎ ብዙሕ ረቛሒታት ምዃኑ ገለጸ ።

### 2.2 Text Pre-processing

The first task to mapping of the function that contain mapping of character and tore them in the dictionary. This means converting one scripte to the other. In our case converting from Amharic to English or Tigrigna to English.

import string

def load\_mapping(file\_path):

    mapping =

    with open(file\_path, 'r', encoding='utf-8') as file:

        for line in file:

            if '=' in line:

                key, value = line.strip().split('=')

                mapping[key] = value

    return mapping

### 2.3 Removal of punctuatuion

The next task is remove the punctuation mark from both languages,

def remove\_punctuation(languages):

    punctuation\_mark = [':', ';', '::', '/','\*','()',

                    '?','።' , '፣' , '፥' , '፦' , '፧' , '፨','፤', '፠ ', '፦']

    for mark in punctuation\_mark:

        languages = languages.replace(mark, '')

    return languages

### 2.4 Input conversion

This code is used to convert the input text in to the mapping dictionary. For our project the input was Amharic and Tigrigna to the text of words(sera words).

def convert\_text(input\_text, mapping):

    converted\_text = []

    for char in input\_text:

        converted\_text.append(mapping.get(char, char))

    return ''.join(converted\_text)

### 2.5 Word overlap

After mapping the text with the mapping text the next task is analyzing word overlap. For our case analyzing the word overlap between Tigrigna and Amharic.

def calculate\_word\_overlap(text1: str, text2: str) -> float:

    words1 = set(text1.split())

    words2 = set(text2.split())

    shared\_words = words1.intersection(words2)

    total\_unique\_words = len(words1.union(words2))

    return (len(shared\_words) / total\_unique\_words) \* 100 if total\_unique\_words > 0 else 0

### 2.6 Reading text

The main purpose this code is to read the enter content of the file. For our project reading the Amharic, Tigrigna and words from ‘UTF-8’.

if \_\_name\_\_ == "\_\_main\_\_":

    # Load the mapping from the file

    mapping\_dict = load\_mapping('words.txt')

    # Read the Amharic text

    with open('Amharic text.txt', 'r', encoding='utf-8') as file:

        amharic\_text = file.read()

    # Read the Tigrigna text

    with open('Tigrigna text.txt', 'r', encoding='utf-8') as file:

        tigrigna\_text = file.read()

### 2.7 Calling what is defined

The last task is calling what is defined in the above code and run the code.

print("Converted Amharic Text:")

    print(converted\_amharic\_text)

    print("\nConverted Tigrigna Text:")

    print(converted\_tigrigna\_text)

    print("\nShared Characters:", shared\_characters)

    print("Unique Amharic Characters:", unique\_amharic)

    print("Unique Tigrigna Characters:", unique\_tigrigna)

    print(f"Similarity Percentage: {similarity\_percentage:.2f}%")

## 3. Results

Unique Amharic words = {‘c’,’n’}

Unique Tigrigna words = {‘x’,’w’}

Word overlap = 1.03%

Phoneme ovderlap = 89.47%

## 4. Conclusion

This project demonstrated the process of text pre-processing, word frequency analysis, grapheme to phoneme conversion, and statistical analysis for a pair of Ethiopian languages. The results show the extent of word and phoneme overlap between the selected languages.

## 5. References

Natural Language Processing with Python (NLTK)

Python Text Processing with NLTK 2.0 Cookbook

<https://www.w3schools.com/python/ref_set_intersection.asp>

<https://pypi.org/project>

<https://codewithmosh.com>

YouTube video “Python Tutorial - Python Full Course for Beginners” by mosh hamedani

<https://en.wikipedia.org/wiki/Ge%CA%BDez_script>