Report

This version of our code does not import and use g2p, but functions with a built-in dictionary. We have created this code as a failsafe for the one that uses the g2p python package to transduce the texts into their phonetic equivalents.

We built the code thus:

* We first imported sys and io to set the system encoding to UTF-8.
* We wrote the dictionary that maps the Amharic letters to their respective phonemes taken from the Amharic letters’ SERA phoneme representatives.
* We assigned two variables to act as placeholders for our Amharic and Tigrigna sample texts.
* We defined a function that processes the sample texts and removes the punctuations and the leading/trailing whitespaces, and used it to clean the sample texts and assigned these clean texts to new clean-text variables.
* We created a function which processes the grapheme-to-phoneme conversion of these clean-text variables
* We defined a function that counts the word frequency and computes the word frequency overlap.
* We assigned variables to the word-frequency-processed and returned Amharic and Tigrigna texts.
* We processed the clean-text variables with the g2p-converting function we defined and created placeholders for these phonetically represented sample texts.
* We created a function that removes whitespaces from the processed phoneme texts.
* We cleaned the phoneme texts from whitespaces using the whitespace- removing function.
* We assigned these clean phoneme texts new variables.
* We defined a new function to count the phonemes in the phoneme texts, and assigned the return values to new variables.
* We defined a function to compute the phoneme overlap score, and used it to process these new variables that hold the counted phonemes.
* Finally, we wrote print statements to return the variables to the word- frequency-processed Amharic and Tigrigna texts, the word overlap score, the Amharic and Tigrigna phoneme distributions, and the phoneme overlap score.

Sample result:

Amharic Word Frequency: Counter({'ላይ': 3, 'አቤቱ': 2, 'በኃይልህ': 2, 'ንጉሥ': 2, 'ደስ': 2, 'በማዳንህ':

2, 'ሁሉ': 2, 'ታግኛቸው': 2, 'ከፍ': 2, 'ይለዋል': 1, 'እጅግ': 1, 'ሐሤትን': 1, 'ያደርጋል': 1, 'የልቡን': 1, 'ፈቃድ': 1,

'ሰጠኽው': 1, 'የከንፈሩንም': 1, 'ልመና': 1, 'አልከለከልኸውም': 1, 'በበጎ': 1, 'በረከት': 1, 'ደርሰህለታልና': 1,

'ከክቡር': 1, 'ዕንቍ': 1, 'የሆነ': 1, 'ዘውድን': 1, 'በራሱ': 1, 'አኖርህ': 1, 'ሕይወትን': 1, 'ለመነህ': 1, 'ሰጠኸውም': 1,

'ለረጅም': 1, 'ዘመን': 1, 'ለዘላለሙ': 1, 'ክብሩ': 1, 'ታላቅ': 1, 'ነው': 1, 'ክብርንና': 1, 'ምስጋናን': 1, 'ጨመርህለት':

1, 'የዘላለም': 1, 'በረከትን': 1, 'ሰጥተኸዋልና': 1, 'በፊትህም': 1, 'ደስታ': 1, 'ታሰኘዋለህ': 1, 'በእግዚአብሔር': 1,

'ተማምኖአልና': 1, 'በልዑልም': 1, 'ምሕረት': 1, 'አይናወጥም': 1, 'እጅህ': 1, 'ጠላቶችህን': 1, 'ቀኝህም': 1,

'የሚጠሉህን': 1, 'በተቈጣህም': 1, 'ጊዜ': 1, 'እንደ': 1, 'እሳት': 1, 'እቶን': 1, 'አድርጋቸው': 1, 'እግዚአብሔር': 1,

'በቍጣው': 1, 'ያጠፋቸዋል': 1, 'እሳትም': 1, 'ትበላቸዋለች': 1, 'ፍሬአቸውን': 1, 'ከምድር': 1, 'ዘራቸውንም': 1,

'ከሰው': 1, 'ልጆች': 1, 'ታጠፋለህ': 1, 'ክፋትን': 1, 'በአንተ': 1, 'ዘርግተዋልና': 1, 'የማይቻላቸውንም': 1, 'ምክር': 1,

'አሰቡ': 1, 'ወደ': 1, 'ኋላቸው': 1, 'ትመልሳቸዋለህ': 1, 'ፍላጻን': 1, 'በፊታቸው': 1, 'ታዘጋጃለህ': 1, 'በል': 1,

'ጽናትህንም': 1, 'እናመሰግናለን': 1, 'እንዘምርማለን': 1})

Tigrigna Word Frequency: Counter({'ኢኻ': 8, 'እዩ': 5, 'እውን': 5, 'ኣብ': 4, 'ኸኣ': 3, 'ሂብካዮ': 3,

'ጐይታይ': 2, 'ኺሕጐስ': 2, 'በረኸት': 2, 'ድማ': 2, 'ኢድካ': 2, 'ንዅሎም': 2, 'ኽትፈልጥ': 2, 'እያ': 2, 'ሓዊ': 2,

'ካብ': 2, 'ከኣ': 2, 'ኦ': 1, 'ንጉስ': 1, 'ብሓይልኺ': 1, 'ብድሕነትኪ': 1, 'ኣዝዩ': 1, 'ትምኒት': 1, 'ልቡ': 1, 'ንልማኖ':

1, 'ኸናፍሩ': 1, 'ኣይከሓድካዮን': 1, 'ጽቡቕ': 1, 'ሒዝካ': 1, 'ናብኡ': 1, 'መጺእካ': 1, 'ርእሱ': 1, 'ኣኽሊል': 1,

'ክቡር': 1, 'ኣእማን': 1, 'ኣቐሚጥካዮ': 1, 'ህይወት': 1, 'ኪወሃበካ': 1, 'ሓቲትካዮ': 1, 'ንነዊሕ': 1, 'እዋንን': 1,

'ንዘለኣለምን': 1, 'ክብሩ': 1, 'ብድሕነትካ': 1, 'ዓብዪ': 1, 'ክብርን': 1, 'ምስጋናን': 1, 'ወሲኽካሉ': 1, 'ዘለኣለማዊ': 1,

'ብሓጐስ': 1, 'ቅድሜኻ': 1, 'ክትሕጐስ': 1, 'ንጉስሲ': 1, 'እግዚኣብሄር': 1, 'ይውከል': 1, 'ብምሕረት': 1, 'እቲ': 1,

'ልዑል': 1, 'ኣይናወጽን': 1, 'ጸላእትኻ': 1, 'የማነይቲ': 1, 'እቶም': 1, 'ዚጸልኡኻ': 1, 'ብግዜ': 1, 'ቝጥዓኻ': 1, 'ከም':

1, 'እቶን': 1, 'ኽትገብሮም': 1, 'ኣሎካ': 1, 'የሆዋ': 1, 'ብቝጥዓኡ': 1, 'ኼጥፍኦም': 1, 'ኪበልዖም': 1, 'ፍሪኦም': 1,

'ምድሪ': 1, 'ዘርኦም': 1, 'ደቂ': 1, 'ሰብ': 1, 'ክትቈርጽ': 1, 'ከመይሲ': 1, 'ልዕለኺ': 1, 'ኽፉእ': 1, 'ዘስፋሕፍfh': 1,

'ዓቕሞም': 1, 'ዘይፈቕደሎም': 1, 'ምኽሪ': 1, 'ኣዳልዮም': 1, 'እዮም': 1, 'ክትመልሶም': 1, 'ገጾም': 1, 'ፍላጻታት': 1,

'ከተንብሮም': 1, 'ብሓይልኻ': 1, 'ልዕል': 1, 'ኣብል': 1, 'ስለቲ': 1, 'ጽንዓትካ': 1, 'ውዳሰ': 1, 'ኽንዝምር': 1, 'ኢና':

1})

Word Overlap Score: 1.1363636363636365

Amharic Phoneme Distribution: Counter({'e': 126, 'a': 83, 'l': 48, 'n': 45, 't': 43, 'h': 43, 'm': 33,

'w': 28, 'b': 27, 'r': 22, 's': 21, 'k': 19, 'u': 17, 'y': 17, 'c': 16, 'g': 15, 'd': 13, 'ə': 11, 'f': 11, 'z': 11,

'o': 8, 'q': 6, 'i': 6, 'j': 5, 'ñ': 4})

Tigrigna Phoneme Distribution: Counter({'a': 96, 'h': 79, 'k': 67, 'e': 60, 'i': 44, 't': 43, 'n': 35,

'm': 35, 'o': 33, 'b': 33, 'y': 32, 'l': 32, 's': 25, 'ə': 25, 'u': 21, 'w': 21, 'r': 17, 'g': 13, 'd': 13, 'z': 11,

'f': 10, 'q': 9})

Phoneme Overlap Score: 0.88

References:

* [g2p · PyPI](https://pypi.org/project/g2p/)
* [Geʽez script - Wikipedia](https://en.wikipedia.org/wiki/Ge%CA%BDez_script)
* [Amharic-Revision.pdf (internationalphoneticassociation.org)](https://www.internationalphoneticassociation.org/sites/default/files/handbookfiles/Amharic-Revision.pdf)
* [PyPI · The Python Package Index](https://pypi.org/)
* [Python - Count and split/strip words in strings - Stack Overflow](https://stackoverflow.com/questions/49892743/python-count-and-split-strip-words-in-strings)
* [Python Counter Module - Python Geeks](https://pythongeeks.org/python-counter-module/)
* [Python's Counter: The Pythonic Way to Count Objects – Real Python](https://realpython.com/python-counter/)
* [Python Classes and Objects - GeeksforGeeks](https://www.geeksforgeeks.org/python-classes-and-objects/)
* [Python re.sub() – Replace using Regular Expression (pythonexamples.org)](https://pythonexamples.org/python-re-sub/)
* [A Complete Guide to Redis Keys - GeeksforGeeks](https://www.geeksforgeeks.org/a-complete-guide-to-redis-keys/)
* [g2p\_transformer/preprocessing/text.py at master · sophiayk20/g2p\_transformer](https://github.com/sophiayk20/g2p_transformer/blob/master/preprocessing/text.py) [(github.com)](https://github.com/sophiayk20/g2p_transformer/blob/master/preprocessing/text.py)