Predlog poboljšanja korišćenja etiopskog pravopisa kreiranjem standardizovane virtualne tastature

Gerawork Aynekulu

Mentor: Vladimir Flipović (PhD)

Matematički fakultet Univerziteta u Beogradu

Beograd, Srbija, 2016.

Improving Ethiopic-based Orthography by standardized virtual keyboard

The case of Amharic

Gerawork Aynekulu

Advisor: Vladimir Flipović (PhD)

Submitted to Belgrade University, Faculty of Mathematics
In partial fulfillment of the requirements for the degree of
Master of Science in computer science

Belgrade, Serbia, 2016.

Acknowledgment

I first would like to send my gratitude to the "World in Serbia~ Svet u Srbiji" scholarship project for granting the scholarship that covers tuition fee, housing cost, meals and stipend.

Second, I send my thanks to the mathematics faculty professors, assistants and other staffs for your collaboration. Your help by arranging resource materials and exams in English language just for a single student deserves so much credit.

In addition, I say thank you to the "MSc R & I 2013" colleagues of mine for your support, tips, tricks and making my campus stay unforgettable.

Last but not least, I would like to express my gratitude to my supervisor Vladimir Flipović for introducing the concept of Software development. I thank him for the follow ups, remarks, inputs and sharing his precious time.

Loved ones and new friends I made in Belgrade, I say thank you for supporting through the entire process.

Table of Contents

List of Figures

List of Tables

Apstrakt

Ovaj rad opisuje kako standardizacija metoda za unos teksta na mobilnim uređajima može da poboljša ortografiju etiopskog pisma, koje ima više od 350 karaktera. U radu je predstavljena tastatura sa rasporedom tipki kao na standardnoj QWERTY tastaturi, Analizirana je učestalost karaktera na jeziku sa pismom zasnovanom na etiopskom, a opcija automatskog popunjavanja sa čistim anotiranim rečnikom omogućava brzo i precizno kucanje. U ovom radu je jezik amharik izabran za studiju slučaja i pokazano je da se standardizacija neće zaustaviti na odbacivanju veoma različitih alatki za kucanje teksta koje su razvili poznati proizvođači softvera, već i doprinosi poboljšanju pisanja amharika, s obzirom na to da zbog nepostojanja strogih pravila pisanja, redundantnosti simbola, fonoloških nedoslednosti, i nekontrolisanog usvajanja pozajmljenica. Predstavljena je i priprema korpusa za jezik amharik, analiza učestalosti karaktera i reči, metrički rezultati pritiska na tipke po karakteru za postojeće dizajne tastatura, Osmišljen je i predlog novog dizajna tastature. Rezultat procene pokazuje da dizajn tastature za svaki lokalni jezik zasnovan na etiopskom rešava transfer iz jednog jezika u drugi kod bilingvalnih korisnika.

Abstract

This paper describes how standardization of text input methods for handheld devices can improve Orthography in Ethiopic, a more than 350 characters script. It assets a keyboard layout design based on the standard QWERTY keyboard, character frequency in Ethiopic-based language and auto-complete from clean tagged dictionary could allow to type quickly and precisely. It choses Amharic as a case study and proclaims standardization will not stop at eradicating very dissimilar key layout typing tools, developed by distinct software vendors, but improves Amharic spelling caused by lack of strict spelling conventions, symbol redundancy, phonology clashes and unsupervised adoption of loan words as well. Amharic corpus preparation, character and word frequency analysis, keystrokes per characters metric result of existing keyboard layouts and proposal of new keyboard layout is presented. Evaluation result shows design of keyboard layout for each local Ethiopic-based language solves cross-language transfer in bilingual authors.

Introduction

In an attempt to propose the role of typing tools in improving Ethiopic-based orthography, a more than 350 characters ancient script used for writing many Ethiopian languages, I learned that lack of publicly available character and word frequency analysis has created a huge knowledge gap. Hence, this project is started by developing the frequency analysis of Amharic, the most popular and official working language of Ethiopia, from text documents available on the Internet to make an informed decision. Therefore, the new proposed keyboard is based on two fundamental backgrounds:

1. Amharic letters and words frequency analysis
2. Users' typing experience in widely used QWERTY keyboard.

Although Amharic has been the de facto language of literature and governance for modern Ethiopia since 19th century, there are some fundamental spelling faults to date that need more research from language academies. Even if lack of strict spelling conventions, symbol redundancy, phonology clashes and unsupervised adoption of loan words from both Semitic and Cushitic sister languages make up lion's share, I believe the recent non-foolproof and very dissimilar key layout typing tools has exacerbated the inconsistency and misspelling in Amharic orthography. Incompetence to accept all Amharic characters and over ability to accept non Amharic characters make text editing very cumbersome. Searching, replacing, sorting and other text query features from local editors to search engine results are complicated due to lack of easy-to-use typing software. Consequently, users are forced to express their ideas in unofficial and dissimilar Latin transliteration in text messaging or online interaction.

Therefore, the project prepares Amharic corpus collected from different online resources, character and word frequency analysis, shows the fault of the existing layouts and proposes a keyboard design exclusively for Amharic, while not rejecting acceptable namesake characters. In addition to speeding up human-computer interaction, it strives to show the impact of autocomplete from clean tagged dictionary in tackling unacceptable spelling errors. Finally, it proposes improvement on the Amharic alphabet sequence and designated Unicode table to improve Ethiopic text entry performance.

I. Project Description

1. Project description

a) Purpose

The main purpose of this project is to show how standardization of virtual keyboards can improve orthography in Ethiopic script user languages. The following are specific objectives.

- Prepare Amharic character and word frequency analysis
- Find out the problem of existing keyboard layouts
- Design a standardized Amharic virtual keyboard layout for mobile devices that will enable to type quickly and precisely.
- Suggest improvement to researchers working on Unicode and key layout standardization.

b) Scope

The scope of the project is designing Android keyboard layout which can be used to accept all Amharic alphabets, punctuation marks and autocomplete for the most frequently used keywords. Choosing bottom-up approach, the project attempts to tackle the misspelling in the whole Ethiopic-based literature by taking Amharic as a case study. In the era where localization is taken as a way forward, I don't believe only one for all Ethiopic typing tool languages is practical. Amharic, Geez, Tigirigna etc. need their own specific keyboards. Therefore although the conclusion made from Amharic can be translated and implemented in other languages that use Ethiopic script, the scope of this project is only limited to Amharic language.

The final product can be installed on any android device with a minimum SDK version 7 and Nyala font that is used to support Ethiopic Unicode (UTF-16) scripts.

c) Stakeholders

The intended users of the app will be any Amharic speaker living in Ethiopia or abroad who uses smart phones. Poverty, 30% adult literacy rate [1], state monopolized internet service provider results in Ethiopia having 1.9% internet penetration rate according to global internet report 2015 [2]. Buddle.com estimated 34% mobile and 2.3% internet penetration with massive improvement in international bandwidth fiber infrastructure and 3G mobile brand services. [3]. Only 4% of Ethiopians are the lowest smart phone owners, and an additional 44% own cellphone that is not smartphone. [4] Most stakeholders are urban and metropolitan area dwellers that potentially can speak the federal working language. In addition, there are more than a million men in Diaspora community, living in USA, Middle East and Europe.

The Ge'ez Frontier Foundation, Ethiopian Information and Communication Technology Development Agency (EICTDA) could find the suggestions and improvements as input to their standardization effort and supervisory role of Ethiopic.

d) Naming conventions and Definitions

Amharic (አማርኛ/amarəñña)/: The official working language of the Federal Democratic Republic of Ethiopia

Ethiopic Script: The script used for writing many Ethiopian languages, mainly those languages that derived from Geez, such as Amharic, Tigrigna and Agew.

Abugida/abjad (አቡጊዳ/*äbugida*): Type of writing system whose basic characters denotes consonants followed by a particular vowel, and in which diacritics denote other vowels.

Fidel (ፊደል/*fidäl*): Full table set of Ethiopic script used for many languages in Ethiopia

Namesake characters: Characters with the same sound but different symbol.

EICTDA: Ethiopian Information and Communication Technology Development Agency

Amharic alphabet: A subset of Ethiopic script (*ሀ-ፐ* and its 8 transformations), which is currently used in Amharic Orthography.

2. Overview

There is no consensus among scholars whether Ethiopic [***U+1200–U+137F***] is an alphabetic[1] or syllabic[2] writing system. It has 350 characters that are used for writing many Ethiopian languages, mainly those languages that are derived from classic Ge'ez, such as Amharic, Tigrigna, Sebatbeit and Agew etc. It is believed that those languages were crafted in the current form in 4th century AD. Ethiopic writing system is used by Ethiopians, Eritreans, Ethiopian Jews in Israel, Rastafarians and millions of east African origin emigrants throughout the world. Ge'ez, currently serving only as the main language of liturgy in

[1] " … A syllabic writing system recognizes only the syllables and creates symbols (diagrams) for each of the syllable. It neither recognizes the consonants and the vowels nor creates symbols to represent them. In that sense the Ethiopic writing system is not syllabic. Most symbols of the Ethiopian writing system, that is, most signs in the "fidel", represent a consonant and a vowel, not a syllable. For example, bu, bi, ba, be, bo, are each a pair of a consonant and a vowel. In the Ethiopian writing system, each of these pairs is represented by one symbol. You have to identify these symbols as signs representing a consonant and a vowel. But they are not necessarily syllables; if they are it is just accidental, which happens only when a combination of a consonant and a vowel produces a syllable, which is not always the case. For the sake of illustration, let us have these one-syllable English words: "bay", "bed", "sun", "girl", "sink", and "germ". As we can see, the first three have three letters and the last three have four letters. But they are all words of one syllable. See how the dictionary divides them. If so, a syllabic writing system would use one symbol for each. But since the English (or the Latin) writing system is not syllabic, but consonantal, it used a combination of consonant and vowel symbols to represent them. This is true also of the Ethiopian writing system. Even though each of the six words ("bet", "bed", "sun", "girl", "sink", and "germ") is of one syllable, one would use more than one sign to reproduce them in Ethiopian fidel . In Ge'ez and Amharic, "sar" ("grass"), "bet" ("house"), "ayn" "("eye") are words of one syllable. If you can write each of these with one symbol of the Ethiopic or Amharic alphabet, then the system is syllabic; if not, it is not. As you know, you need two letters for "sar" and "bet", and three letters for "ayn". Let us have more examples: the words "meskel" ("cross"), "sibket" ("preaching") and "sigdet" ("prostration") are composed of two syllables: mes.qel, sib.ket, and sig.det. If you can write each of these words with only two symbols of the Ethiopic or Amharic alphabet (i.e. one symbol for each syllable) then the system is syllabic; if not, it is not. You need four (not two) symbols (i.e. letters) for each of these words. Our system is consonantal like the English, not syllabic." [12]

[2] "The Ethiopians are the only people that differ from the users mentioned above. The difference lies in the fact that our writing system, unlike the Greek and the Latin which use the alphabetic system, is the prior syllabic system... In the same way that Geez took and modified Sabean scripts to become a full-fledged written language, Amharic also took the Geez scripts and became a written language. However, unlike Geez, Amharic took all the 26 characters of its predecessor, among which we find the "extra" characters that make similar sounds in the "ha", "se", "Se", etc, family. It is believed that it is because of the pressure from the church and the state that all these characters were maintained.
On top of the 26 characters, Amharic also needed additional characters to represent sounds that it acquired from Cushitic and other foreign languages. This was done by placing a small bar (or hat) on top of 7 characters that were inherited from Geez. Examples are "she", "che", "Ce", "je", "Gne" or "Ne", "He" and "zhe". Amharic had now, by this time, 33 characters. The total number of syllables is accordingly 231 (33x7).

There haven't been changes in the rather large number of shapes and variations in the script adopted by Amharic, mainly due to the influence of the church and the state." [9]

Ethiopian and Eritrean orthodox churches, played the same role as Latin has been playing for European languages in drafting new words and being the origin of most characters.

It has left to right writing direction in horizontal lines unlike other Semitic languages. Ethiopic is a syllabic system, unlike Latin i.e. "a symbol represents not a sound but pairing or groups of phonemes. According to professor Baye Yimam [5] "Syllabic, one of the three writing systems [alphabetic, syllabic, logographic], correlates a symbol to combination of vowel and consonant". For example, a language which has 26 basic sounds and 5 vowels will have 26*5=130 groups of phonemes which directly means 130 different symbols. So each symbol represents consonant(s) plus vowel and is the result of 8 or more different modifications to indicate the vowels. Peter T. Daniels (in paper [6]) categorizes Ethiopic under alpha-syllabary or abugidas i.e. a type of writing system whose basic characters denotes consonants followed by a particular vowel, and in which diacritics denote other vowels. Below is a simple example to show how the transformation is done.

Order	1st	2nd	3rd	4th	5th	6th	7th	8th
Ge'ez	ግዕዝ/ ge'ez	ካዕብ/ ka`Ib	ሣልስ/ sals	ራብዕ/ rab`I	ኃምስ/ hams	ሳድስ/ sads	ሳብዕ/ sab`I	ሳምን/ samin
Ethiopic	በ	ቡ	ቢ	ባ	ቤ	ብ	ቦ	ቧ
Vowel	з	u	i	a	e	ə	o	'wa
Latin	Bз	Bu	Bi	Ba	Be	B	Bo	B'wa
Example	Bird	Bull	Bill	Bat	Bet	Brown	Bob	Quality[3]
Ethiopic	በrd	ቡll	ቢll	ባt	ቤt	ብrown	ቦb	ኳlity

Table 1 Example of transliteration of syllables with vowels and transformation

Therefore, a single Ethiopic character is equivalent to monograph, digraph, trigraph, sometimes more Latin characters combination. There is no standard way of transliterating Ethiopic into Latin alphabet though there are individual attempts (see [7]) like the EAE transliteration system developed by Encyclopedia Aethiopica, and the BGN/PCGN* system,

[3] 'wa' is a diphthong i.e. a vowel with two different targets

which was designed for use in Romanizing names written in Amharic characters adopted by the UN in 1967.

In this project I took Amharic, the second most spoken Semitic language next to Arabic with more than 27 million native speakers (according to [1]) and learned by a large number of citizens as second language throughout the country. Amharic is chosen as a case study because it is popular and the official working language of the Federal Democratic Republic of Ethiopia. Amharic is written using Ethiopic, which has inherited the learning materials of Geez language with little modification.

Amharic has extant records dating back 14th century and has been the de facto language of literature and governance of modern Ethiopia since 19th century.However, there are some fundamental spelling faults that need more research (detailed information about that issue are presented in [8]).

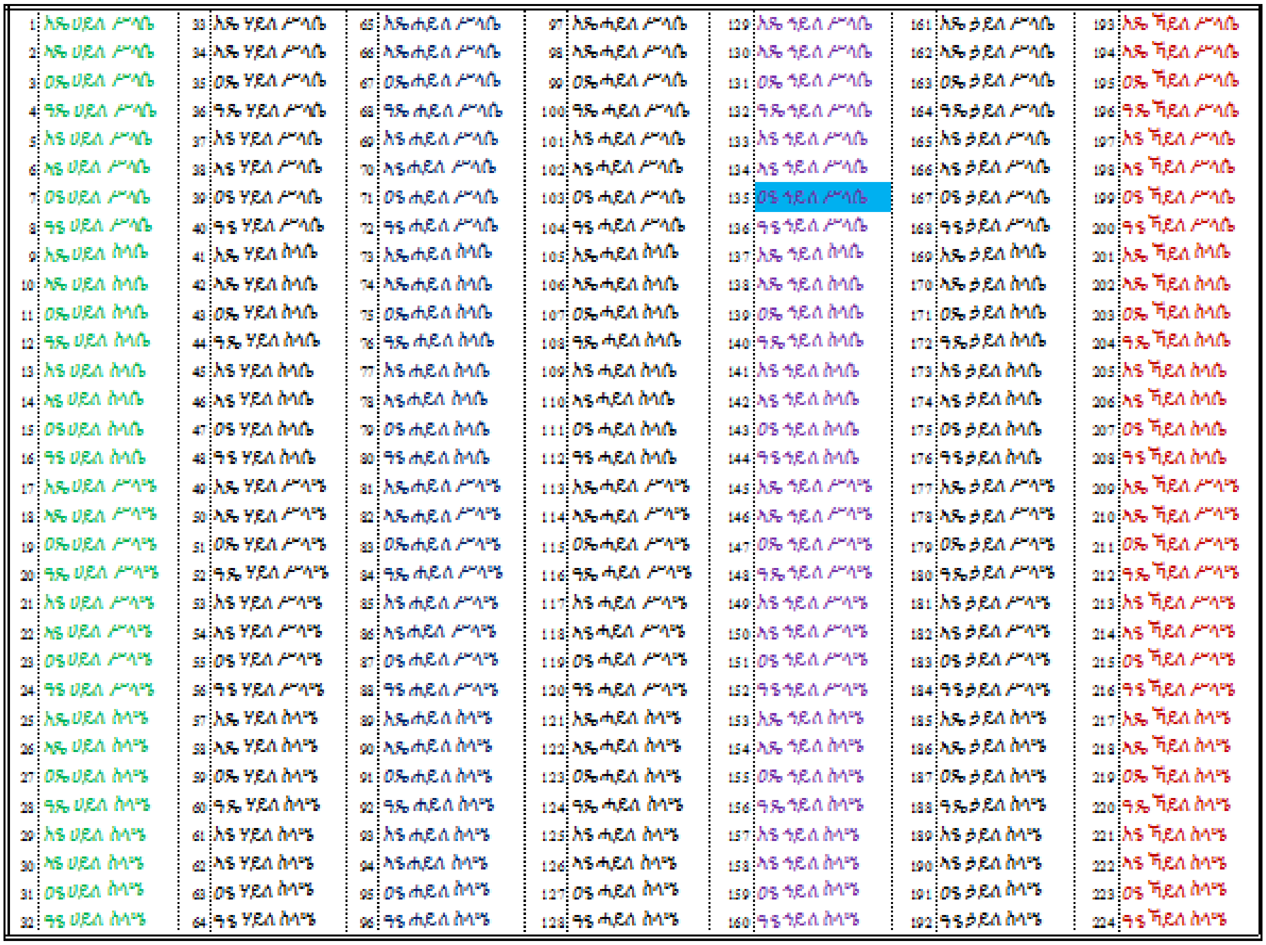
Amharic sentence doesn’t have uppercase letters in the beginning of a sentence or anywhere for that matter. However, there are additional soft and stressed sounds which are usually used as analogous to cases in English. Therefore the English equivalent 26 upper and 26 lower cases are more than enough to contain the 34 root characters of Amharic and its 8 order transformation.

3. Problem statement

Amharic has less than 220 sounds and due to its Ethiopic background it is supposed to need a symbol for each sound. These symbols are not random rather a transformation of 28 main roots and 8 orders column. However, due to phonologically lost yet graphically present symbols (ዐ, ሐ, ኸ, ሰ, ፀ, and ኀ), loan words foreign words and symbols (ቪ in Visa) and reserve symbols of modification table (ቿ, ቮ, ቯ, ጿ) result in the script having additional characters. Consequently it currently has, including non-mandatory namesake and extended characters, more than 280 symbols in addition to punctuation marks. Most linguistic scholars (elaborated in [9]) argue the problem of namesake characters dates back when Amharic inherited main characters from its predecessor Ge’ez. The process left extra characters with similar sound and the difficulty of designing computer or typewriter keyboard gets worth when the extended symbol is more than two for a single sound. These are ሐ,ኸ,ኀ, ሠ, ,ዐ ,ጸ each of them having 8 transformations to create 48 redundant symbols. Daniel Yakob inquires further the complexity added by 4 ''Diqqala' or extended characters of Ge'ez, each having 5 transformations (in [10]), as wrong substitution "Glypheme Misidentification".e.g. መኮንን Vs መኰንን/ mekōnnin and ኅንዳር vs ጐንዳር/*Gondär*.

Ethiopian Languages Study and Research Center published a dictionary in February 1993. (Paper [11]) with a proposal to reform the alphabet by "reducing" 80 namesake characters including 8th column. There is no linguistic or mathematical justification on why particularly namesake characters are still in Amharic orthography, which makes the research center's decision commendable. For example, there are two characters which are used to write the first character of the English word 'sun' that literally means 'Tsehay' in Amharic. For the 'Ts' sound Amharic has 'ፀ' and 'ጸ' characters which sound absolutely the same. Authors mostly take the former as canonical and their justification behind the choice is because the first one resembles the 'sun' in shape since it is more circular than the later one. Scholars who proposed reform couldn't get the green light from conservative scholars and the problem continues up to date.

Historian and linguist Pr.Getachew arguing how the problem started believes "In the course of time some sounds disappear from the language (leaving their symbols behind)". (see [12])



1	አፄ ሀይለ ሥላሴ	33	አፄ ሃይለ ሥላሴ	65	አፄ ሐይለ ሥላሴ	97	አፄ ሓይለ ሥላሴ	129	አፄ ኀይለ ሥላሴ	161	አፄ ኃይለ ሥላሴ	193	አፄ ኻይለ ሥላሴ
2	ኣፄ ሀይለ ሥላሴ	34	ኣፄ ሃይለ ሥላሴ	66	ኣፄ ሐይለ ሥላሴ	98	ኣፄ ሓይለ ሥላሴ	130	ኣፄ ኀይለ ሥላሴ	162	ኣፄ ኃይለ ሥላሴ	194	ኣፄ ኻይለ ሥላሴ
3	ዐፄ ሀይለ ሥላሴ	35	ዐፄ ሃይለ ሥላሴ	67	ዐፄ ሐይለ ሥላሴ	99	ዐፄ ሓይለ ሥላሴ	131	ዐፄ ኀይለ ሥላሴ	163	ዐፄ ኃይለ ሥላሴ	195	ዐፄ ኻይለ ሥላሴ
4	ዓፄ ሀይለ ሥላሴ	36	ዓፄ ሃይለ ሥላሴ	68	ዓፄ ሐይለ ሥላሴ	100	ዓፄ ሓይለ ሥላሴ	132	ዓፄ ኀይለ ሥላሴ	164	ዓፄ ኃይለ ሥላሴ	196	ዓፄ ኻይለ ሥላሴ
5	አጼ ሀይለ ሥላሴ	37	አጼ ሃይለ ሥላሴ	69	አጼ ሐይለ ሥላሴ	101	አጼ ሓይለ ሥላሴ	133	አጼ ኀይለ ሥላሴ	165	አጼ ኃይለ ሥላሴ	197	አጼ ኻይለ ሥላሴ
6	ኣጼ ሀይለ ሥላሴ	38	ኣጼ ሃይለ ሥላሴ	70	ኣጼ ሐይለ ሥላሴ	102	ኣጼ ሓይለ ሥላሴ	134	ኣጼ ኀይለ ሥላሴ	166	ኣጼ ኃይለ ሥላሴ	198	ኣጼ ኻይለ ሥላሴ
7	ዐጼ ሀይለ ሥላሴ	39	ዐጼ ሃይለ ሥላሴ	71	ዐጼ ሐይለ ሥላሴ	103	ዐጼ ሓይለ ሥላሴ	135	ዐጼ ኀይለ ሥላሴ	167	ዐጼ ኃይለ ሥላሴ	199	ዐጼ ኻይለ ሥላሴ
8	ዓጼ ሀይለ ሥላሴ	40	ዓጼ ሃይለ ሥላሴ	72	ዓጼ ሐይለ ሥላሴ	104	ዓጼ ሓይለ ሥላሴ	136	ዓጼ ኀይለ ሥላሴ	168	ዓጼ ኃይለ ሥላሴ	200	ዓጼ ኻይለ ሥላሴ
9	አፄ ሀይለ ስላሴ	41	አፄ ሃይለ ስላሴ	73	አፄ ሐይለ ስላሴ	105	አፄ ሓይለ ስላሴ	137	አፄ ኀይለ ስላሴ	169	አፄ ኃይለ ስላሴ	201	አፄ ኻይለ ስላሴ
10	ኣፄ ሀይለ ስላሴ	42	ኣፄ ሃይለ ስላሴ	74	ኣፄ ሐይለ ስላሴ	106	ኣፄ ሓይለ ስላሴ	138	ኣፄ ኀይለ ስላሴ	170	ኣፄ ኃይለ ስላሴ	202	ኣፄ ኻይለ ስላሴ
11	ዐፄ ሀይለ ስላሴ	43	ዐፄ ሃይለ ስላሴ	75	ዐፄ ሐይለ ስላሴ	107	ዐፄ ሓይለ ስላሴ	139	ዐፄ ኀይለ ስላሴ	171	ዐፄ ኃይለ ስላሴ	203	ዐፄ ኻይለ ስላሴ
12	ዓፄ ሀይለ ስላሴ	44	ዓፄ ሃይለ ስላሴ	76	ዓፄ ሐይለ ስላሴ	108	ዓፄ ሓይለ ስላሴ	140	ዓፄ ኀይለ ስላሴ	172	ዓፄ ኃይለ ስላሴ	204	ዓፄ ኻይለ ስላሴ
13	አጼ ሀይለ ስላሴ	45	አጼ ሃይለ ስላሴ	77	አጼ ሐይለ ስላሴ	109	አጼ ሓይለ ስላሴ	141	አጼ ኀይለ ስላሴ	173	አጼ ኃይለ ስላሴ	205	አጼ ኻይለ ስላሴ
14	ኣጼ ሀይለ ስላሴ	46	ኣጼ ሃይለ ስላሴ	78	ኣጼ ሐይለ ስላሴ	110	ኣጼ ሓይለ ስላሴ	142	ኣጼ ኀይለ ስላሴ	174	ኣጼ ኃይለ ስላሴ	206	ኣጼ ኻይለ ስላሴ
15	ዐጼ ሀይለ ስላሴ	47	ዐጼ ሃይለ ስላሴ	79	ዐጼ ሐይለ ስላሴ	111	ዐጼ ሓይለ ስላሴ	143	ዐጼ ኀይለ ስላሴ	175	ዐጼ ኃይለ ስላሴ	207	ዐጼ ኻይለ ስላሴ
16	ዓጼ ሀይለ ስላሴ	48	ዓጼ ሃይለ ስላሴ	80	ዓጼ ሐይለ ስላሴ	112	ዓጼ ሓይለ ስላሴ	144	ዓጼ ኀይለ ስላሴ	176	ዓጼ ኃይለ ስላሴ	208	ዓጼ ኻይለ ስላሴ
17	አፄ ሀይለ ሥላሤ	49	አፄ ሃይለ ሥላሤ	81	አፄ ሐይለ ሥላሤ	113	አፄ ሓይለ ሥላሤ	145	አፄ ኀይለ ሥላሤ	177	አፄ ኃይለ ሥላሤ	209	አፄ ኻይለ ሥላሤ
18	ኣፄ ሀይለ ሥላሤ	50	ኣፄ ሃይለ ሥላሤ	82	ኣፄ ሐይለ ሥላሤ	114	ኣፄ ሓይለ ሥላሤ	146	ኣፄ ኀይለ ሥላሤ	178	ኣፄ ኃይለ ሥላሤ	210	ኣፄ ኻይለ ሥላሤ
19	ዐፄ ሀይለ ሥላሤ	51	ዐፄ ሃይለ ሥላሤ	83	ዐፄ ሐይለ ሥላሤ	115	ዐፄ ሓይለ ሥላሤ	147	ዐፄ ኀይለ ሥላሤ	179	ዐፄ ኃይለ ሥላሤ	211	ዐፄ ኻይለ ሥላሤ
20	ዓፄ ሀይለ ሥላሤ	52	ዓፄ ሃይለ ሥላሤ	84	ዓፄ ሐይለ ሥላሤ	116	ዓፄ ሓይለ ሥላሤ	148	ዓፄ ኀይለ ሥላሤ	180	ዓፄ ኃይለ ሥላሤ	212	ዓፄ ኻይለ ሥላሤ
21	አጼ ሀይለ ሥላሤ	53	አጼ ሃይለ ሥላሤ	85	አጼ ሐይለ ሥላሤ	117	አጼ ሓይለ ሥላሤ	149	አጼ ኀይለ ሥላሤ	181	አጼ ኃይለ ሥላሤ	213	አጼ ኻይለ ሥላሤ
22	ኣጼ ሀይለ ሥላሤ	54	ኣጼ ሃይለ ሥላሤ	86	ኣጼ ሐይለ ሥላሤ	118	ኣጼ ሓይለ ሥላሤ	150	ኣጼ ኀይለ ሥላሤ	182	ኣጼ ኃይለ ሥላሤ	214	ኣጼ ኻይለ ሥላሤ
23	ዐጼ ሀይለ ሥላሤ	55	ዐጼ ሃይለ ሥላሤ	87	ዐጼ ሐይለ ሥላሤ	119	ዐጼ ሓይለ ሥላሤ	151	ዐጼ ኀይለ ሥላሤ	183	ዐጼ ኃይለ ሥላሤ	215	ዐጼ ኻይለ ሥላሤ
24	ዓጼ ሀይለ ሥላሤ	56	ዓጼ ሃይለ ሥላሤ	88	ዓጼ ሐይለ ሥላሤ	120	ዓጼ ሓይለ ሥላሤ	152	ዓጼ ኀይለ ሥላሤ	184	ዓጼ ኃይለ ሥላሤ	216	ዓጼ ኻይለ ሥላሤ
25	አፄ ሀይለ ስላሤ	57	አፄ ሃይለ ስላሤ	89	አፄ ሐይለ ስላሤ	121	አፄ ሓይለ ስላሤ	153	አፄ ኀይለ ስላሤ	185	አፄ ኃይለ ስላሤ	217	አፄ ኻይለ ስላሤ
26	ኣፄ ሀይለ ስላሤ	58	ኣፄ ሃይለ ስላሤ	90	ኣፄ ሐይለ ስላሤ	122	ኣፄ ሓይለ ስላሤ	154	ኣፄ ኀይለ ስላሤ	186	ኣፄ ኃይለ ስላሤ	218	ኣፄ ኻይለ ስላሤ
27	ዐፄ ሀይለ ስላሤ	59	ዐፄ ሃይለ ስላሤ	91	ዐፄ ሐይለ ስላሤ	123	ዐፄ ሓይለ ስላሤ	155	ዐፄ ኀይለ ስላሤ	187	ዐፄ ኃይለ ስላሤ	219	ዐፄ ኻይለ ስላሤ
28	ዓፄ ሀይለ ስላሤ	60	ዓፄ ሃይለ ስላሤ	92	ዓፄ ሐይለ ስላሤ	124	ዓፄ ሓይለ ስላሤ	156	ዓፄ ኀይለ ስላሤ	188	ዓፄ ኃይለ ስላሤ	220	ዓፄ ኻይለ ስላሤ
29	አጼ ሀይለ ስላሤ	61	አጼ ሃይለ ስላሤ	93	አጼ ሐይለ ስላሤ	125	አጼ ሓይለ ስላሤ	157	አጼ ኀይለ ስላሤ	189	አጼ ኃይለ ስላሤ	221	አጼ ኻይለ ስላሤ
30	ኣጼ ሀይለ ስላሤ	62	ኣጼ ሃይለ ስላሤ	94	ኣጼ ሐይለ ስላሤ	126	ኣጼ ሓይለ ስላሤ	158	ኣጼ ኀይለ ስላሤ	190	ኣጼ ኃይለ ስላሤ	222	ኣጼ ኻይለ ስላሤ
31	ዐጼ ሀይለ ስላሤ	63	ዐጼ ሃይለ ስላሤ	95	ዐጼ ሐይለ ስላሤ	127	ዐጼ ሓይለ ስላሤ	159	ዐጼ ኀይለ ስላሤ	191	ዐጼ ኃይለ ስላሤ	223	ዐጼ ኻይለ ስላሤ
32	ዓጼ ሀይለ ስላሤ	64	ዓጼ ሃይለ ስላሤ	96	ዓጼ ሐይለ ስላሤ	128	ዓጼ ሓይለ ስላሤ	160	ዓጼ ኀይለ ስላሤ	192	ዓጼ ኃይለ ስላሤ	224	ዓጼ ኻይለ ስላሤ

P(A tse Haile Silassie)=P(A)*P(tse)*P(Ha)*P(i)*P(le)*P(Si)*P(la)*P(ssie)=4*2*7*1*1*2*1*2=**224**

Table 2 Jah Test: Worst case spelling error example to write "Atse Haile Selassie" where only option 135 highlighted in blue is canonically correct

As it is explained in the introduction above, there are many causes of poor orthography in Amharic. The major ones are:

1. Lack of consensus among scholars on the reformation of Ethiopic script: following the introduction of modern education, there have been different proposals on reformation of Ethiopic script. In 1993, a group of linguistic scholars at Ethiopian Languages Study and Research Center published a dictionary using new proposed “Fidel’ by reducing the total number of characters to 218. (28*7+22). However, this proposed system gets no acceptance due to lack of strict national implementation policy and the strong criticism it gets from traditional scholars.

2. Lack of standard design among different Amharic typing tools by different software vendors: there are lots of different proprietary or free software and online web services which allow users to input Ethiopic letters. In 2010, Microsoft ™ released localized Windows Vista operating system in Amharic; since then it has enabled users to write in Ethiopic in its own Office packages (see [13]). Amharic language package is also a handy tool, especially for older users who hardly understand English to interact in their own native language. Google ™ accepts Ethiopic through its Chrome browser extension and its Ethiopian domain Google ™ .com.et. There are also Ethiopic keyboards in Google Play ™ for Android devices like Multiling and Agergna.

Software Package	1st	2nd	3rd	4th	5th	6th	7th	8th	Jah Test[4]
Power Geez	key	u	i	a	ie	e	o	Caps+wa	19
Key man	e	u	i	a	y	key	o	Shift+w	21
Microsoft ™	e	u	i	a	ie	‘	o	ua	25
Amharic dictionary	e	u	i	a	shift+E	key	o	Shift+w	21
Google Input Tools	e	u	i	a		key	o	Can’t	19
Multiling	key	u	i	a	New key	e	o	Can’t	22

Unfortunately, it is not possible to find a simple Amharic sentence to test all Amharic symbols unlike English’s “the quick brown fox jumps over the lazy dog“ keyboard test. I used the previous phrase “Atse haile selassie~ ats’e haylä səllasé” which will be referenced as ‘Jah [5]test’ now onwards from the above figure in the previous page for testing and the result on the number of key presses is located in the last column of the above table. The difference in key layout is not restricted only in vowel makers. The main characters are also not immune to mismatch, which makes it cumbersome for users to adapt each and every keyboard layout.

[4] Referring the king by his other popular name, Jah, among the Rastafari movement

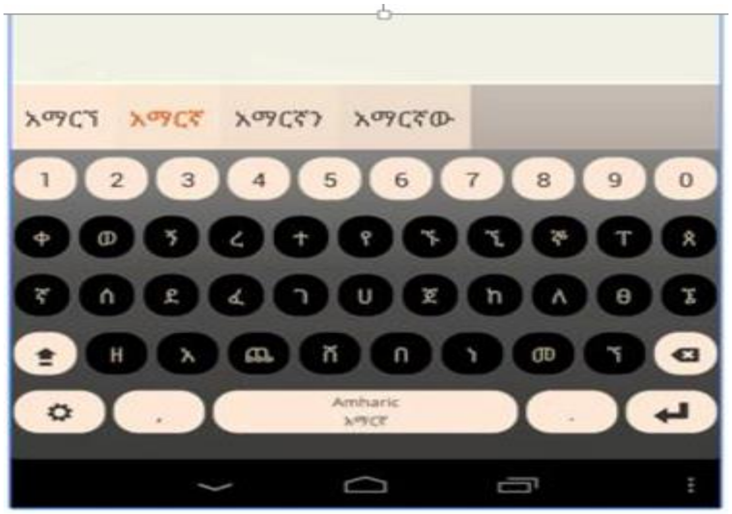
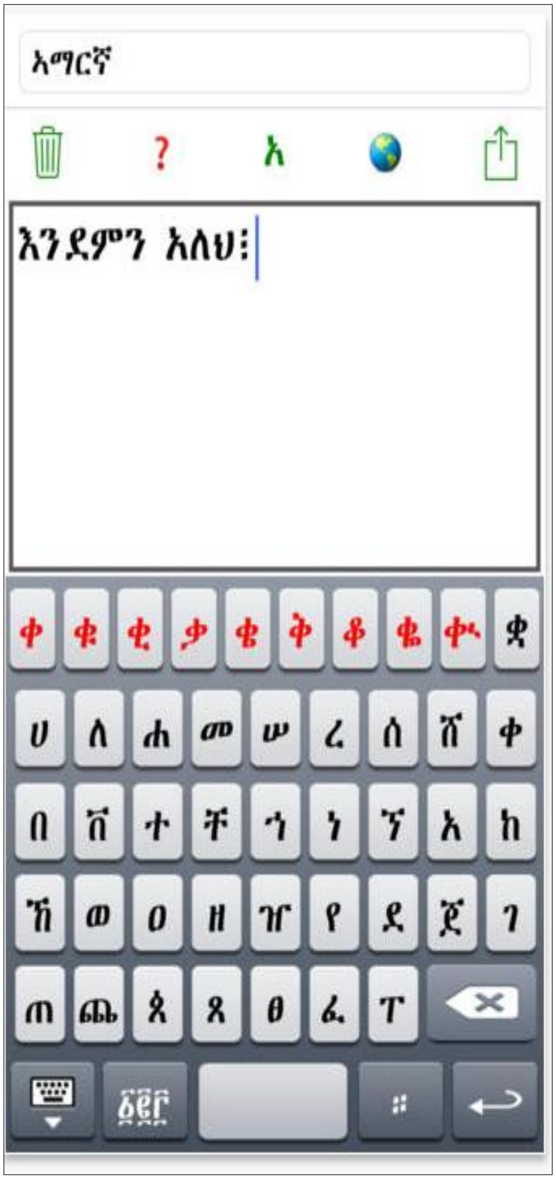
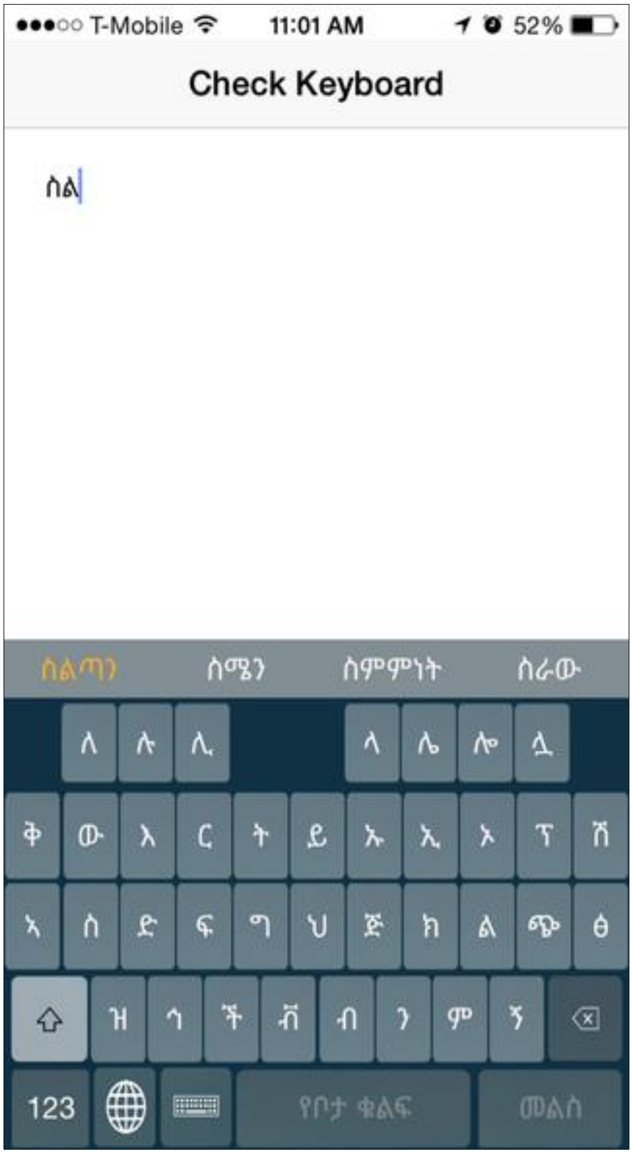


Figure 1: Multiling Keyboard


አማርኛ

እንደምን አለህ፡

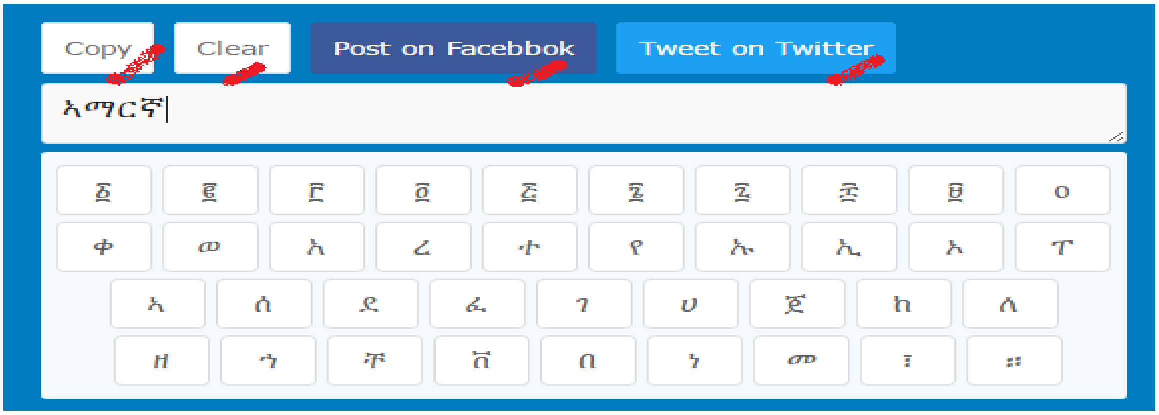
Figure 2: Ha le ha me based Amharic keyboard



ስል

Figure 3: Query based Agerigna keyboard

When it comes to online web services including social media, blogs, emails etc. most users use one of the above shown application software developed to edit text documents of personal computers and copy-paste to bring it to the cyber world. The online typing websites like *amaregna.com*, *type Amharic* ,*Amharic keyboard* etc. created for merely typing Amharic also have buttons like Facebook post, tweet, Google ™ , email to ease the copy-paste process as displayed in the screenshot linked below. However, these tools minimize the power of online experience to the mere social media activities. Imagine how dull an experience can cut and paste be during online chatting!



አማርኛ

Figure 4: amharickeboard.com

It is very common to see this in social medial interactions, YouTube video titles written in unofficial transliteration and hybrid usage due to lack of easy to use browser extensions and mobile apps. Consequently, browsing the internet becomes an irksome process due to the extreme content inconsistency among different websites.

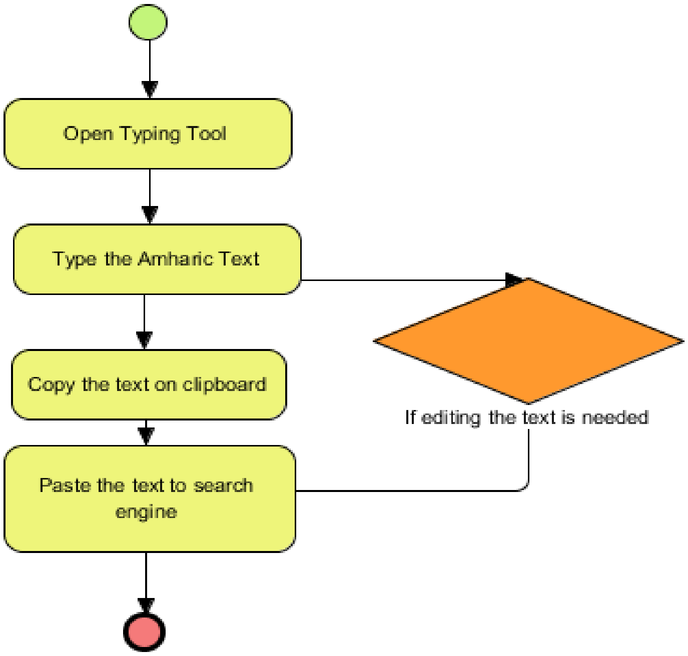
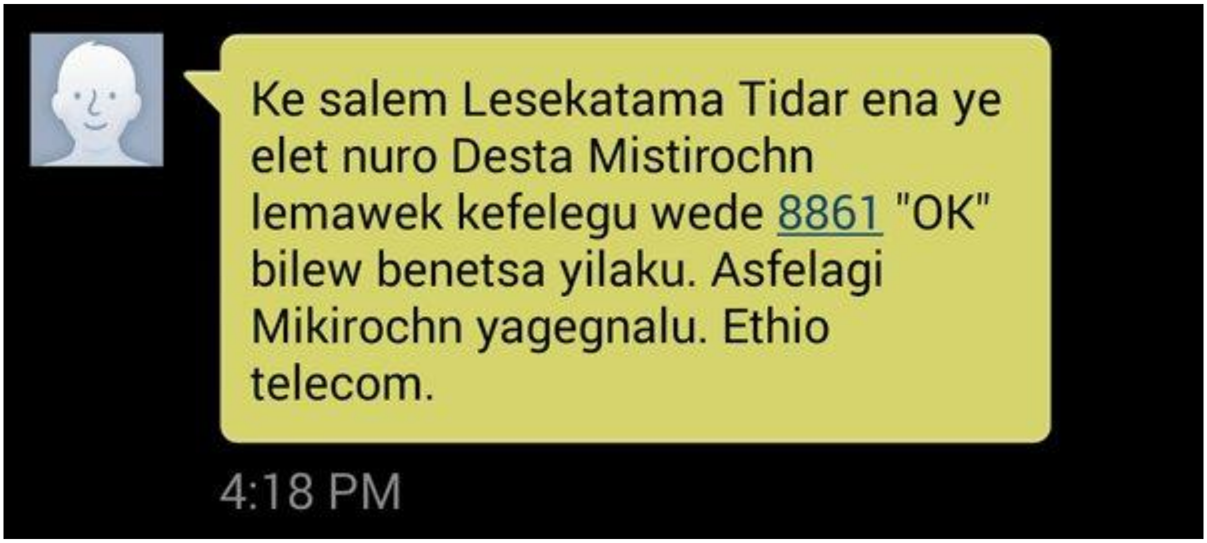

Open Typing Tool
Type the Amharic Text
Copy the text on clipboard
Paste the text to search engine
If editing the text is needed

Figure 5: Searching in Amharic online user experience

3. Absence of Amharic to Latin transliteration

   Amharic to Latin transliteration or vice versa was not a big issue until the recent rapid growth of second generation Ethiopian diaspora communities and text communication tools. When the actual Ethiopic script is unavailable due to technical difficulties, users communicate by transliterating Amharic text into Latin. There is not even an informal transliteration tool or any guide available.



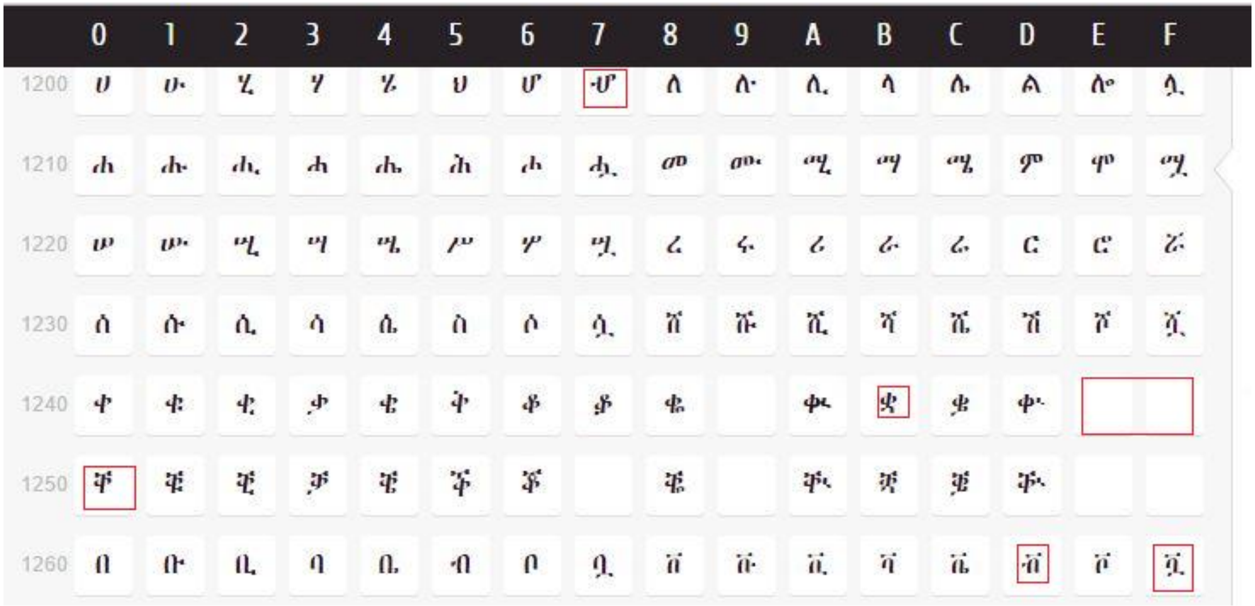
Ke salem Lesekatama Tidar ena ye elet nuro Desta Mistirochn lemawek kefelegu wede 8861 "OK" bilew benetsa yilaku. Asfelagi Mikirochn yagegnalu. Ethio telecom.

4:18 PM

Figure 6: Sample Ethio-Telecom tele marketing message

4. Inconsistency at the 8th column of Unicode table:

   In Unicode the whole Ethiopic script is represented in the range between 1200 to 137c as a matrix of 43 consonants crossed with 8 vowels which has 384 characters out of 512 potential spaces. The Unicode table can hold a two pair of main consonant with its own 8 transformations for each. However, there is reserved space for some characters that do not have 8 transformations. Moreover, extended characters of different languages like ሇ (Nailo Seharan), ኈ (Ge'ez), ቇ (Agewigna) etc. ቐ (Tigirigna) and ዸ(Sidamgna, Gedeogna, Oromiffa) are another case in point that were found at the data cleaning phase of Amharic dictionary corpus preparation that is going to be discussed in detail under the methodology chapter. The infamous መቀሌ/Mäqälle (Mekelle) vs መቐለ/Mäq'älä(Makale) social media debate following the Mekelle vessel misspelling controversy (see [14]) is an indication where the current ethno-sensitive socio-political structure could also strengthen the hurdles that contributed to poor Amharic spelling. መቀሌ/Mäqälle is nationally used by Amharic native speakers except most articles of Reporter magazine that were submitted by Tigrean origin journalists who wrote መቐለ/ Mäq'älä instead.



	0	1	2	3	4	5	6	7	8	9	A	B	C	D	E	F
1200	ሀ	ሁ	ሂ	ሃ	ሄ	ህ	ሆ	ሇ	ለ	ሉ	ሊ	ላ	ሌ	ል	ሎ	ሏ
1210	ሐ	ሑ	ሒ	ሓ	ሔ	ሕ	ሖ	ሗ	መ	ሙ	ሚ	ማ	ሜ	ም	ሞ	ሟ
1220	ሠ	ሡ	ሢ	ሣ	ሤ	ሥ	ሦ	ሧ	ረ	ሩ	ሪ	ራ	ሬ	ር	ሮ	ሯ
1230	ሰ	ሱ	ሲ	ሳ	ሴ	ስ	ሶ	ሷ	ሸ	ሹ	ሺ	ሻ	ሼ	ሽ	ሾ	ሿ
1240	ቀ	ቁ	ቂ	ቃ	ቄ	ቅ	ቆ	ቇ	ቈ		ቊ	ቋ	ቌ	ቍ		
1250	ቐ	ቑ	ቒ	ቓ	ቔ	ቕ	ቖ		ቘ		ቚ	ቛ	ቜ	ቝ		
1260	በ	ቡ	ቢ	ባ	ቤ	ብ	ቦ	ቧ	ቨ	ቩ	ቪ	ቫ	ቬ	ቭ	ቮ	ቯ

Figure 7Ethiopic Unicode standard excerpt

5. Literature review

A Morse code of the Amharic language, described in [15], was the first of its kind in character frequency analysis proposed to speed-up transmission of telegraphic messages in Amharic. Fast forward, An Amharic Speech Corpus for Large Vocabulary Continuous Speech Recognition is the most recent text analysis done on the archive EthioZena website [16].

The 1993 Amharic dictionary developed by Ethiopian languages study and research center, described in [11], tried to tackle the case of namesake characters though it lacks acceptance.

Daniel Yakob’s work on Amharic misspelling [10] is pioneering in most of its forms. In that paper, author divided the accepted level of Amharic spelling into basic, intermediate and advanced or hypercorrect levels. Moreover, he listed out the main causes as symbol redundancy, phonology-orthography disconnect, foreign language transcription and errors inherent from typing systems (Ibid, [10, p. 4]), which is the subject of this project

Ethiopic collation standard first edition document [17] also sets a standard for sorting Ethiopic text and amends “character Set Information encoding” sorting order.

6. Proposed system

This project does not have any intention of joining the debate whether non-mandatory characters should be removed or not. Rather, it would abide by the status quo and propose a virtual keyboard that can be used to input all Amharic characters. However, it will not shy away from recommending linguists in favor of the reformation process. The reformation is usually misjudged as a mere attempt to minimize the number of characters. However, Amharic should get independence from both parent languages like Ge'ez and other sister Semitic and Cushitic sister languages. The notion that the four 'Ha's, two 'Se's, two 'A's and two 'Tse's have different canonical meaning is very lightweight unless it is for sacred Ge'ez language. The reformation camp also should clear misconceptions and push to sustain the characters as a subset of the wider Ethiopic script (Fidel), however exclude from usage in Amharic orthography. For example, ቐ is equally unimportant as ዸ is for Amharic. These characters should be used only in their own language, Tigrigna and Oromiffa respectively, not in Amharic.

The modified version of "Fidel" proposal for Amharic as a matrix of 34 consonants crossed with 8 consonants done as in the Unicode table can be found in Appendix 1.

Unicode has to leave the 8th column exclusively for Amharic so that the 16 character space of a Unicode table row can hold two Ethiopic characters with their transformation. The current mixing of extended characters of different languages ((ሇ,ቇ,ኟ,ፇ,ኯ,ፗ,ዯ,ኇ)) make for an additional task for the programmers, and when exceptions are not handled well, the problem escalates into contributing to misspelling. The standardization update in Unicode table eases the development process which will indirectly improve Amharic orthography.

The extended 4 by 5 characters (ኰ, ቈ, ጐ, ኈ) are not necessary in modern Amharic literature anymore, except for their fourth column which is included in 8th column of the above proposed table's source character. Therefore, the design will be entirely based on the above 34 main characters and punctuation marks (፡, ።, ፣, ፤, ፦).

As English requires no diacritics and uses only 26 letters of ISO basic Latin alphabet unlike other European languages, Amharic should also get independence from its parent Ge'ez and other Cushitic and Semitic languages in both syntax and semantics. Though in this project I decided to follow the status quo, I recommend reformation camp to push for not using specially extended characters in Amharic orthography. For example, the less known official Ge'ez "አኰሰም" has only four Google ™ pages search result, whereas አክሱም, the Amharic equivalent, has more than 14 results. I also believe software vendors, as a problem solver, should push forward and disable those non Amharic keys as they have added character ⶠ by their own will with or without the consent of conservative linguists. Designing a keyboard exclusively for Amharic language will significantly reduce extended and dialectic non-Amharic characters from Tigire, Sebatbiet, Mursi, Gumuz, Afan Oromo etc, which improves the spelling.

Therefore the proposed Amharic keyboard has only 272 character subsets of Ethiopic distributed according to the table below. The exceptions are repetition of ኋ, ኧ, ጿ in their analogues Amharic namesake characters and lack of 'wa' sound in ወ and የ.

Characters	NO of Main Sounds	Vowels	Exception	Total Syllables
Main [ሀ-ፐ]	28	8	2(ወ,የ)	222
Name Sake [ሐ-ጸ]	6	8	3	45
Punctuation[፡]	5	0	0	5
Total Number of Symbols				272

Table 3: Proposed keyboard's total number of key labels

II. Requirement Analysis

7. Business Requirement

The implementation phases of the project are:

1. Character frequency analysis
2. Preparation of cleaned Amharic corpus of the most frequently used words
3. Design and implementation of Amharic keyboard layout
4. Evaluating the product

7.1. Character Frequency analysis

The character frequency analysis of Amharic is the most fundamental requirement for understanding the pattern of letters transformation which in turn helps in designing a very effective keyboard. Unfortunately, the only Amharic news corpus collected from news archives of Walta information center is no longer available for public usage, (Argaw and Asker, 2005). Therefore, I am forced to prepare a 50MB .txt file from Bible[6], Quran[7], Daniel Kibret views blog[8], Zone 9 blog[9], 50 copies of Reporter Magazine from November 2013-June 2014[10] , Amharic Wikipedia August articles dump and Ethio media website[11]. The full content of the first two religious books is chosen for their rich content and their direct

[6] Interlitt, Lapsley/Brooks Foundation,1992-1993, amharic.bible.pdf@gmail.com

[7] Nejashi Publishing , 1997 , nejashi@ethionet.et

[8] Daniel Kibret views Blog,, http://danielkibret.com,

[9] Zone Nine Blog, http://zone9ethio.blogspot.com/

[10] Reporter Magazine , http://www.ethiopianreporter.com/

[11] www.ethiomedia.com

influence on orthography of Amharic by their respective relationship with Hebrew, Greek and Arabic languages. The other sources are the most popular blogs and magazine which have relatively better repository of the current socio-political activities going on both inside and outside of Ethiopia and written by plenty of distinguished authors.

Title	Affiliation	Fraction used	Size(KB)
Bible	Religion	Full	8,438
Quran	Religion	Full	1,351
Reporter Magazine	Politics, Economics, Social	50 copies	22,172
Zone 9 Blog 2005	Politics, Economics, Social	2005 e.c compilation	1,844
Daniel Kibret Views	Social	80 articles	934
Ethiopian Review	Diaspora Ethiopians view	900 pages random articles	3,449
Amharic Wikipedia	General	August 2014 Dump	6,458

Offline explorer and Internet Download Manager (IDM) have been used as web crawler to collect the articles from the web. The following line of Linux commands show the merging and conversion of articles to txt file to make the text analytics easier.

```
pdfunite *.pdf Amharic.pdf
pdftotext Amharic.pdf Amharic.txt
```

Unlike most European scripts where working in memory efficient UTF-8 is recommended, It was necessary to convert UTF-16 considering the windows and java enviorment used to develop the proposed keyboard.

```
        iconv -f utf-8 -t uff-le input.txt > output.txt
```

After data collection, I have done data preprocessing by removing non-Ethiopic letters, punctuation marks, numbers found in the data. As Daniel Yacob of the Ge’ez Frontier foundation noted in [10], I found confusion in using the 7th and 8th vowel maker syllables especially in ኮ and ጎ characters. For example, መኮንን VS መኰንን/ mekōnnin and ጎንዳር VS ጐንዳር/ Gondär.

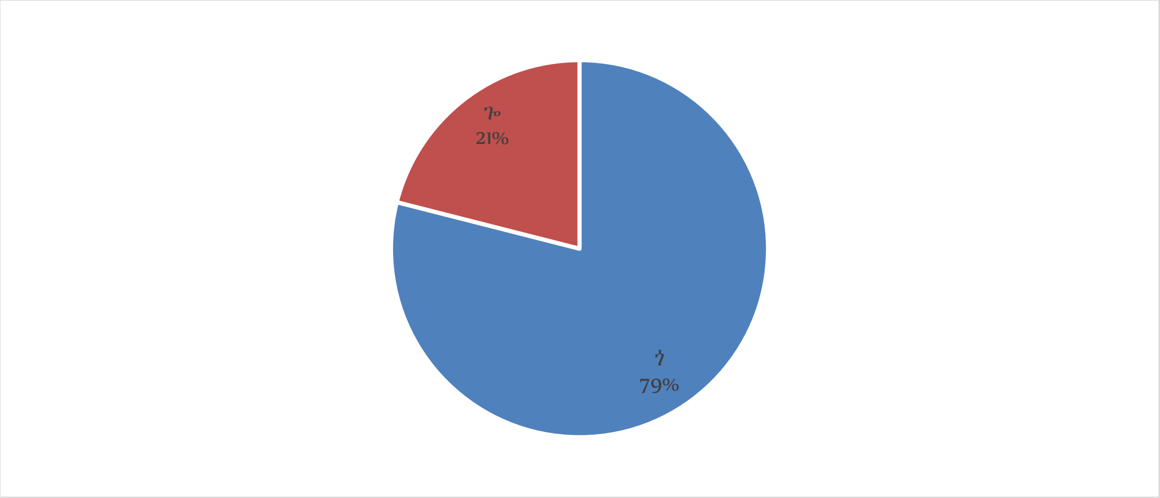


Figure 8: Geez 'go' (21%) vs Amharic 'go' (79%) mixed usage share in the Amharic corpus

In addition, very few non Amharic characters like ቐ, ዸ which are crafted to incorporate borrowed words of other sister languages are removed at this phase.

I used an online character frequency analyzer of Tampere University of technology [12] which accepts a text file and displays the Unicode value, character symbol and the frequency of the character in the document in tabular format.

So it wasn’t necessary to transliterate the Amharic content to Latin equivalent unlike previous attempts that are described in [18] for the mere frequency analysis purpose as previous researches on Ethiopic used to be done. I have collected the frequency of the above documents and found the following results.

The limitation of the corpus preparation was the inability to extend the cleaning process to remove borrowed foreign words, and correct spelling and grammar mistakes. Although the main sources of data are well known official publications, there was room for improvement which needs future work of linguists.

[12] https://www.cs.tut.fi/~jkorpela/charcount.html8 , Jukka “Yucca” Korpela

No	Letter	Frequency	Percentage
1	ሀ	278,320	2.28
2	ለ	931,894	7.62
3	ሐ	75,670	0.62
4	መ	1,033,210	8.45
5	ሠ	84,723	0.69
6	ረ	816,393	6.68
7	ሰ	578,071	4.73
8	ሸ	69,471	0.57
9	ቀ	295,591	2.42
10	በ	827,312	6.77
11	ቨ	9,325	0.08
12	ተ	972,474	7.96
13	ቸ	316,563	2.59
14	ኀ	34,134	0.28
15	ነ	1,170,558	9.58
16	ኘ	95,431	0.78
17	አ	621,632	5.09
18	ከ	402,114	3.29
19	ኸ	8,292	0.07
20	ወ	630,351	5.16
21	ዐ	89,638	0.73
22	ዘ	218,288	1.79
23	ዠ	6,008	0.05
24	የ	921,663	7.54
25	ደ	501,299	4.10
26	ጀ	88,338	0.72
27	ገ	492,357	4.03
28	ጠ	253,805	2.08
29	ጨ	57,871	0.47
30	ጰ	22,029	0.18
31	ጸ	48,271	0.39
32	ፀ	24,929	0.20
33	ፈ	208,625	1.71
34	ፐ	38,226	0.31
Total		12,222,876	100

Table 4: Amharic root characters relative frequency ordered by Unicode standard and/or Ethiopic alphabet

Rank	Letter	Frequency	Percentage
1	ነ	1,170,558	9.58
2	መ	1,033,210	8.45
3	ተ	972,474	7.96
4	ለ	931,894	7.62
5	የ	921,663	7.54
6	በ	827,312	6.77
7	ረ	816,393	6.68
8	ወ	630,351	5.16
9	አ	621,632	5.09
10	ሰ	578,071	4.73
11	ደ	501,299	4.10
12	ገ	492,357	4.03
13	ከ	402,114	3.29
14	ቸ	316,563	2.59
15	ቀ	295,591	2.42
16	ሀ	278,320	2.28
17	ጠ	253,805	2.08
18	ዘ	218,288	1.79
19	ፈ	208,625	1.71
20	ኘ	95,431	0.78
21	ዐ	89,638	0.73
22	ጀ	88,338	0.72
23	ሠ	84,723	0.69
24	ሐ	75,670	0.62
25	ሸ	69,471	0.57
26	ጨ	57,871	0.47
27	ጸ	48,271	0.39
28	ፐ	38,226	0.31
29	ኀ	34,134	0.28
30	ፀ	24,929	0.20
31	ጰ	22,029	0.18
32	ቨ	9,325	0.08
33	ኸ	8,292	0.07
34	ዠ	6,008	0.05
Total		12,222,876	100

Table 5. Amharic root characters relative frequency ordered by character frequency

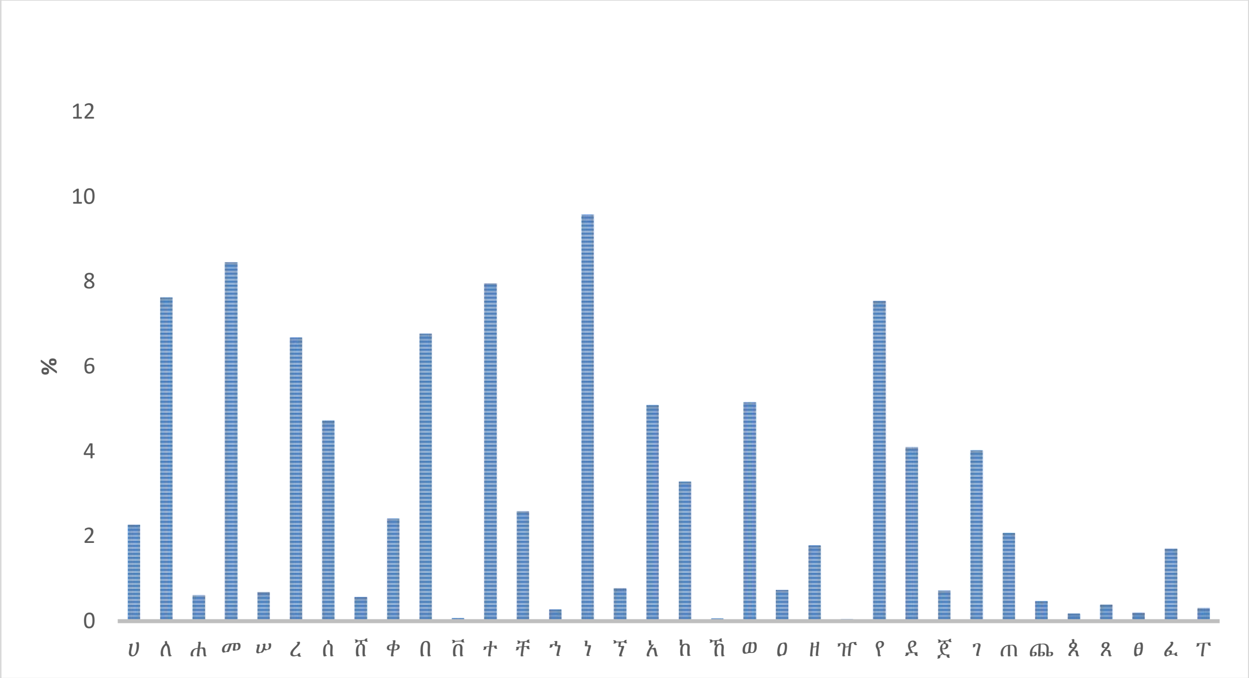


Figure 9.Relative frequency percentage of Amharic letters.

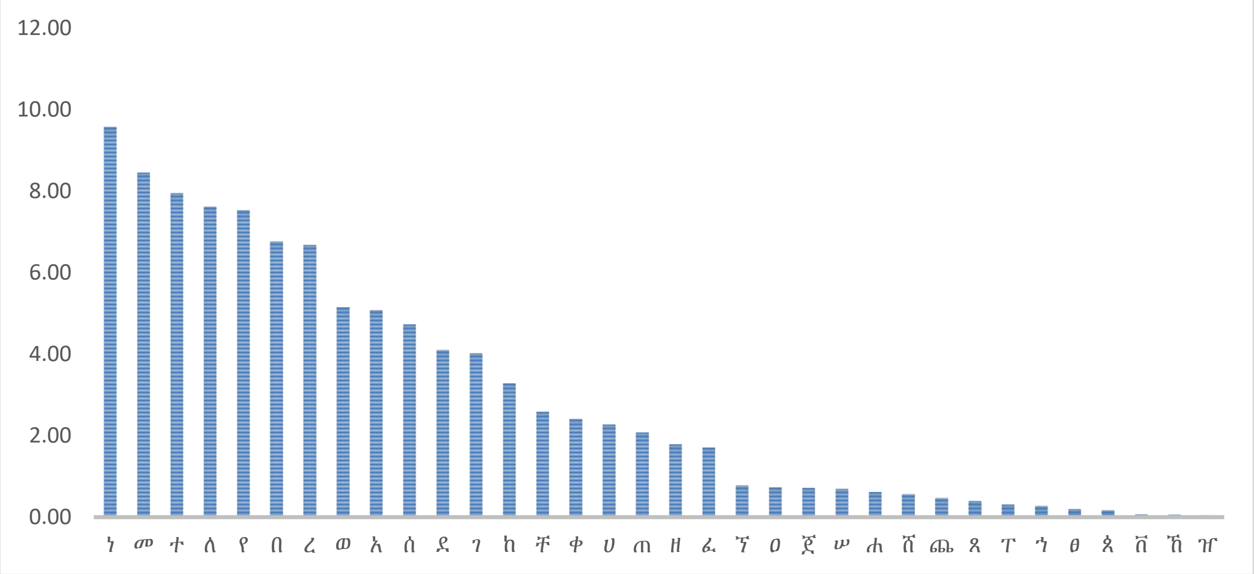


Figure 10: Relative frequency of letters ordered by frequency

In addition to main characters frequency, which will be used to decide in choosing the characters, there are the default characters i.e. letters that are displayed before the shift/caps lock key. Vowel letters transformation is also crucial.

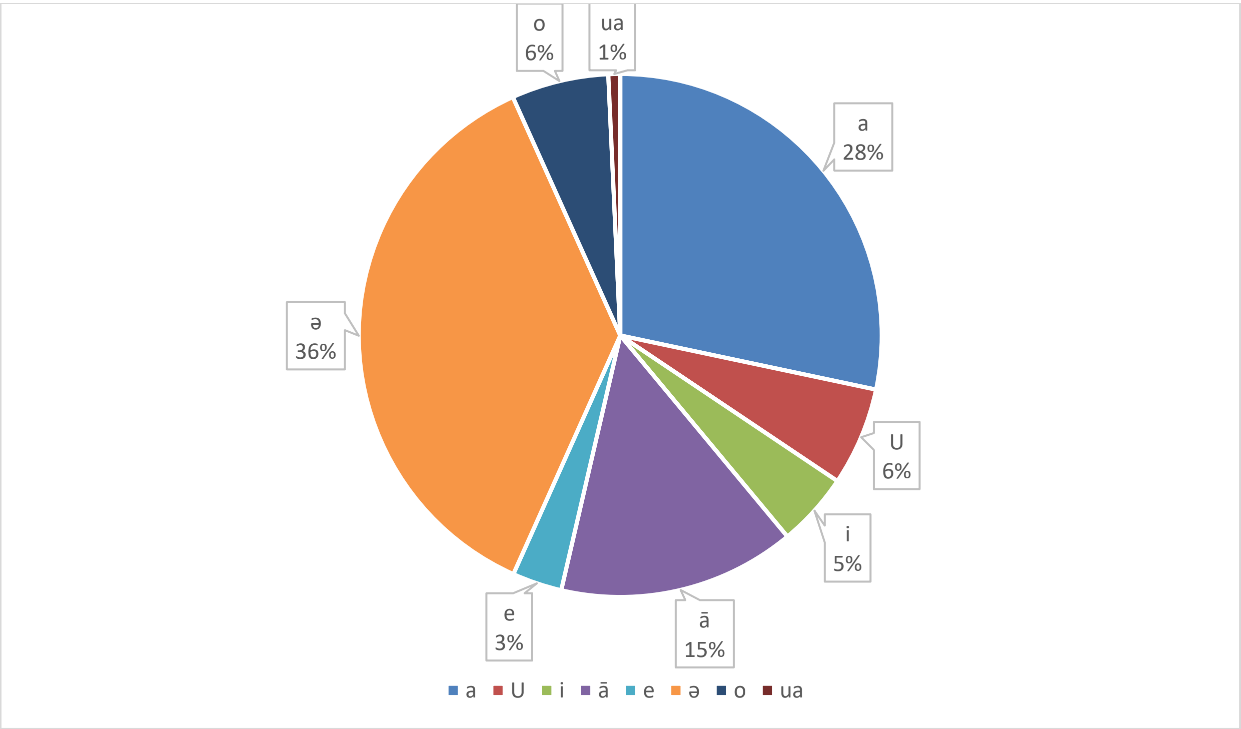


Figure 11. Relative Frequency of Amharic Vowel Transformations

The sixth vowel position (36%) is the most frequent as ‘e’ is most often used in English [19], which in turn entails that the default position of main characters should be at 6th position. By doing so, the users will be saved from writing two letters to create the most frequent character position since by default keyboard stands for 6th position wish single character. There are few keyboards in the market that have ‘a’ default position, which creates confusion among the users, although they raise 28% second most frequent position and main character position in the Amharic Fidel. However, any keyboard configuration that does not put the 6th position as a default can be considered unscientifically based and very far away from facts as displayed in charts and tables above. Let’s take a look at how the line graph spikes in the next page at the 6th position in most of the Amharic letters.

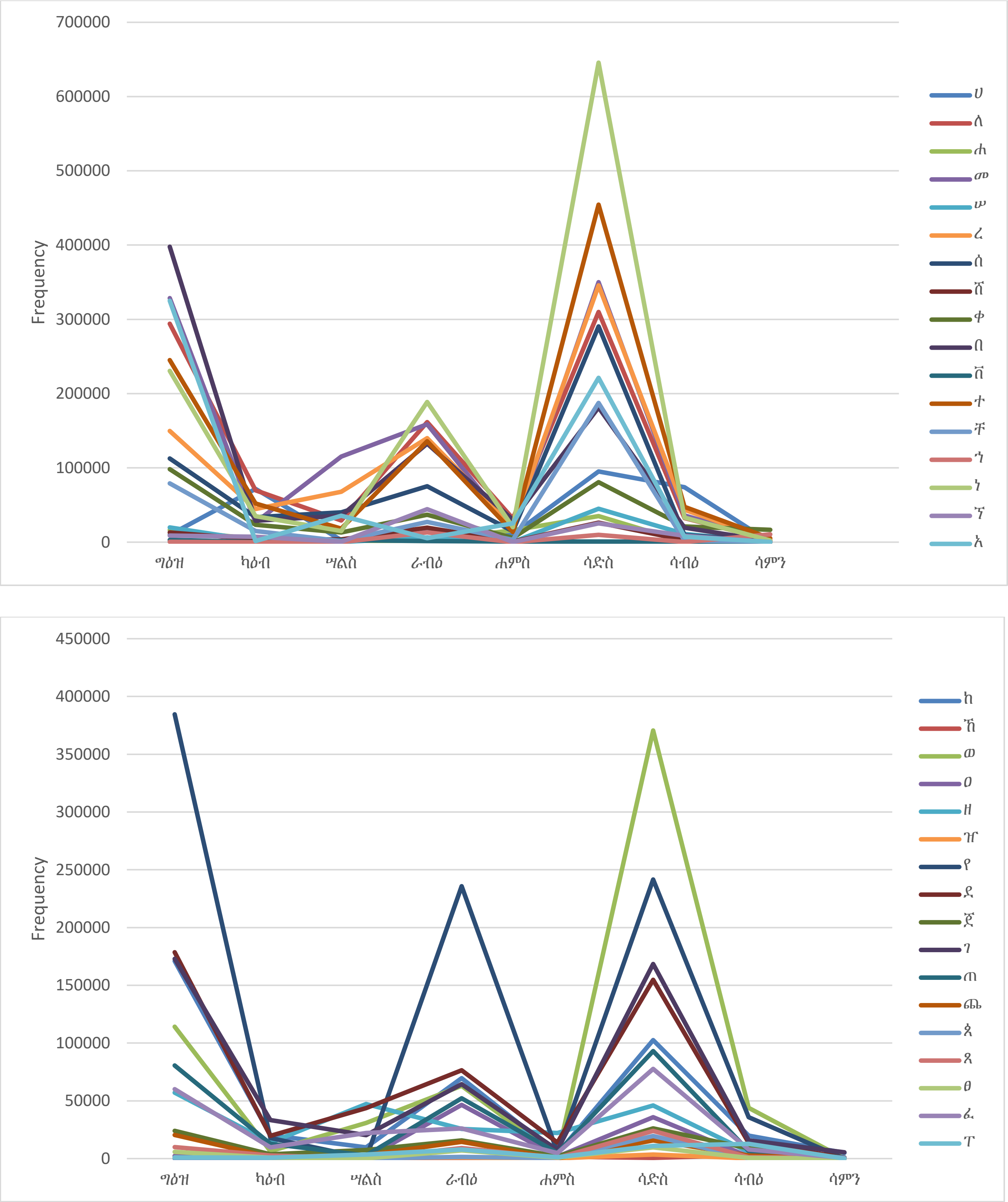


Figure 12: Character transformation and usage frequency

The case of namesake characters is much more complicated than the usually misquoted “Let us minimize the number of namesake characters” academic debate. First, there is sometimes no explicit usage pattern that will allow us to drop one of them. For example the fourth ‘ኣ’ has hardly ever been used while ዓ is used 90% of the time. Second, due to poor Unicode standardization, there is, for example, one ኋ (hua) which has almost exclusive usage share due to lack of other ‘hua’s in the font design process. Therefore, since it needs the will of many stakeholders and the case is beyond the scope of this project, they are handled at the programming level by exceptions and nothing has been changed out of the status quo beyond recommendation.

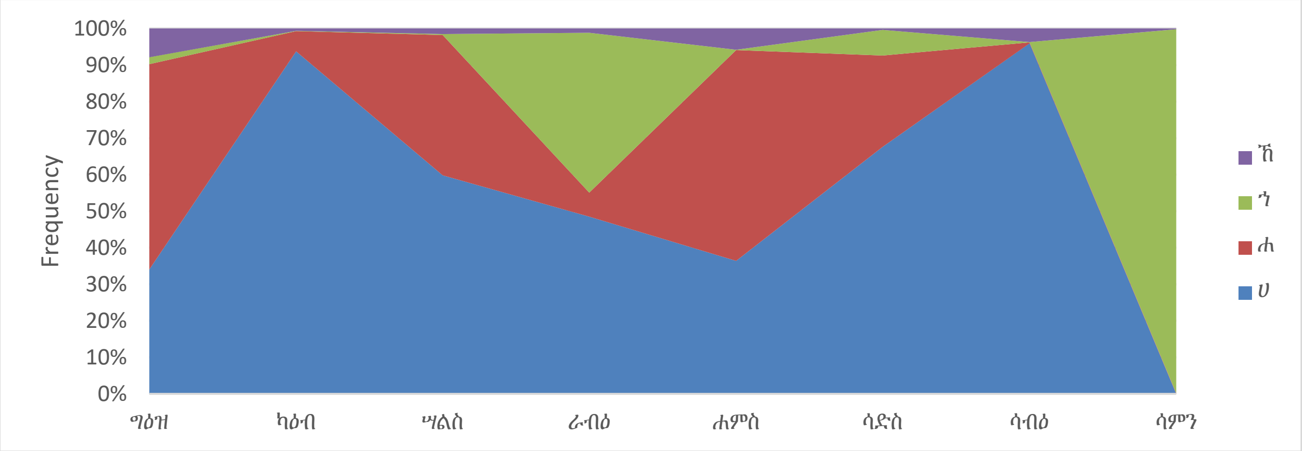


Figure 13: Usage pattern of the four "Ha's"

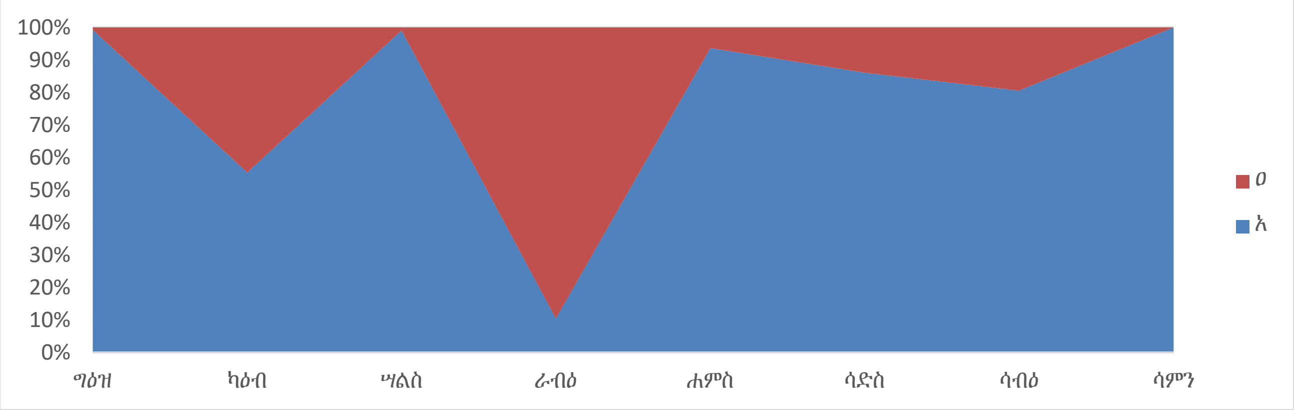


Figure 14: Usage pattern of two "A's"

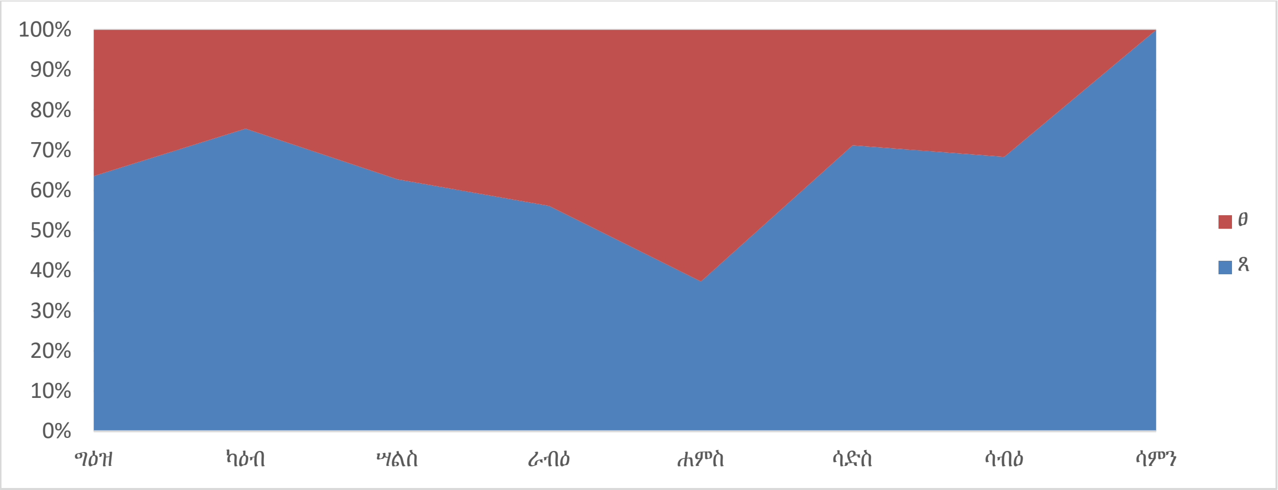


Figure 15: Usage pattern of two "Tse's"

The last, but not least, fundamental thing is regarding punctuation marks. From the Latin keyboard '?' is the most frequent while ፣, ።, ፤ are repeatedly used from Ethiopic in respective order. Contrary to the wide belief, '፡', the punctuation mark used to separate words and currently replaced by space, has an avoidable share in the widely used punctuation marks frequency. Moreover, the Ethiopic numerals have only share 0.005% in the corpus and they are not practically used except in some religious and linguistic Wikipedia articles.

8. System Requirement analysis

The Amharic language pack is a module of open source project called "AnySoft keyboard". Rather than reinventing the wheel, it embeds basic functionalities of this project and builds Amharic keyboard as it is usually the case in most localization efforts. So Amharic keyboard is a keyboard that allows to type 272 characters of Ethiopic, with its own dictionary which is going to be chosen by any interested user which is the only actor. The actors are the 'User' and 'Anysoft kyboard API'.

Use case

Any user starts by installing Amharic keyboard (U1). Then from settings the user will select to use Amharic in addition to the default English keyboard (U2). Hereafter, the user can start typing any Amharic character (U3) and use the auto complete (U4) to speed-up the typing process. Anysoft Keyboard will serve as a support architecture system.

U1: Install Amharic keyboard

Preconditions: The typist must own Android KitKat (4.4+) or latest versions

1. User installs Anysoft keyboard
2. Open Anysoft keyboard settings
3. Click "Get more keyboards" at Google™ play store
4. Select Amharic Language pack and click install.

Postconditions: Amharic keyboard is successfully installed on the device.

U2: Select Amharic Keyboard (Includes U1)

1. Open Anysoft keyboard Settings
2. Click keyboards
3. Check the Amharic keyboard checkbox

U3: Types using Amharic keyboard (Includes U2 and U4)

Preconditions: Amharic keyboard is selected

1. User opens any typing application from the device
2. Use clicks the Amharic tab from the right corner of the keyboard
3. User clicks the 6th main chapter of the word going to be typed
4. If the character is 6th position, move on to the next character; otherwise select one of the transformed values displayed over vowel transformation keys.
5. Include U4

U4: Use autocomplete suggestion (Include U3)

1. Include U3
2. System displays auto-complete suggestions
3. User selects one of the auto-complete suggestions

III. Design and implementation

9. Designing keyboard layout

Touch typists usually don’t look at the keyboard and search each key where it resides. Habit plays a major role for a rapid and accurate typing while the user’s mind is focused on the idea. Since habit comes mostly from the common English keyboard, I tried to correlate close sound mapping with Ethiopic sounds and place it in respective place. The best example for transformation of two symbols into one is the relationship of Latin ‘nj’ to Cyrillic ‘Њ’. So when the user enters a letter, its fate depends upon whether the letter next to it will modify it or not.

Therefore the “qwerty” keyboard will be taken as benchmark and changed into “ቅውእርትይኡኢኦፕአስድፍግህጅክልዝፅችሽብንም” and only an additional key to hold the ‘ua’ transformation of all letters. Shift/ Caps lock will be “ቅውዕርጥይዑዒዖጵዐስድፍኅህኽክልዝፅችሽብንም. Note that I used the English vs Amharic key analogy like q for ቅ and character frequency displayed at Figure 9. All the vowel locations temporarily hold the transformation of Amharic root Alphabet letter during each key press. Therefore we need to place only the root characters in the Amharic XML QWERTY file.

```
<Keyboard xmlns:android="http://schemas.android.com/apk/res/android"
   android:keyWidth="10%p"
    android:horizontalGap="0px"
    android:verticalGap="0px"
    >

<Row>
    <!--Display the most frequent characters on default and popup characters on shift
press-->
    <Key android:codes="4677" android:keyLabel="" android:keyEdgeFlags="left"  />
    <Key android:codes="4813" android:keyLabel=""/>
    <Key android:codes="4768" android:keyLabel="" android:popupCharacters="ዐ" />
    <Key android:codes="4653" android:keyLabel="" />
    <Key android:codes="4725" android:keyLabel=""  android:popupCharacters="ጥ" />
    <Key android:codes="4829" android:keyLabel="" android:popupCharacters="ኽ"/>
    <Key android:codes="4769" android:keyLabel="" android:popupCharacters="ዑ "/>
    <Key android:codes="4770" android:keyLabel="" android:popupCharacters="ዒ"/>
    <Key android:codes="4774" android:keyLabel="" android:popupCharacters="ዖ"/>
    <Key android:codes="4949" android:keyLabel="" android:popupCharacters="ጵ"
android:keyEdgeFlags="right" />
</Row>
……
</Keyboard>
```

Note that the Amharic alphabet key code range is between 4608 and 49666.

```
public boolean isAmharicAlphabetRange(Keyboard.Key key)
{
    return ((key.label.hashCode()>=4608 && key.label.hashCode()<=4951)?true:false);
}
```

Once the keyboard is loaded, during each alphabet key press the vowel locations marked red in the screenshot below need to display the transformation value of root key pressed.

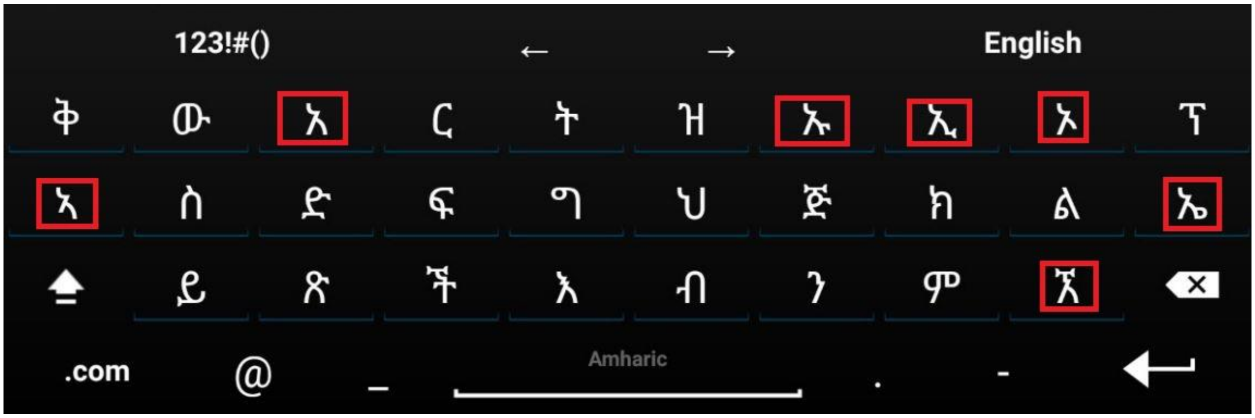


Figure 16: Landscape default Amharic Keyboard layout

Unlike English and other Latin languages where the shifted label is the upper case of the default key or vice versa, if the Amharic keyboard is shifted, the popup character assigned at the XML file is the shifted character label. The root character is the most frequent 6th position of transformation, hence the others are reached by plus or minus values from -2 up to 6. The most frequent problem, i.e. inclusion of non-Amharic characters like ቇ, ኇ, ፇ, ኯ, ዏ, ዯ, ጏ, in existing Amharic keyboards is skipping of the exception that is handled at the switch case part of the code in the next page. The exception allows to filter out non Amharic character located at the 8th column of latest Ethiopic Unicode table. The punctuation marks located at the last row of the keyboard (key 37-40) are populated by simple invalidate function displayed below.

```
invalidateKey(keys.get(37));
keys.get(37).label=(char)(4964)+""; //፤
```

```
if( key.pressed )
            {
//Tranform the vowel key locations to the transformed value of pressed root key
                label=( key.popupCharacters != null&&mKeyboard.isShifted())?key.popupCharacters:key.label;
                key.label=(char)(label.charAt(0))+"";
                invalidateKey(keys.get(6));
                keys.get(6).label=(char)(label.charAt(0)-5)+"";
                invalidateKey(keys.get(10));
                keys.get(10).label=(char)(label.charAt(0)-4)+"";
                invalidateKey(keys.get(11));
                keys.get(11).label=(char)(label.charAt(0)-3)+"";
                invalidateKey(keys.get(12));
                keys.get(12).label=(char)(label.charAt(0)+1)+"";
                invalidateKey(keys.get(14));
                keys.get(14).label=(char)(label.charAt(0)-2)+"";
                invalidateKey(keys.get(23));
                keys.get(23).label=(char)(label.charAt(0)-1)+"";
                invalidateKey(keys.get(32));

                switch ((mKeyboard.isShifted()&&key.label.hashCode()<=4951)?key.popupCharacters.hashCode():key.label.hashCode())
                {

                    case 4613:
                        keys.get(32).label=(char)(label.charAt(0)+134)+"";
                        break;
                    case 4677:
                    case 4741:
                    case 4781:
                    case 4797:
                    case 4877:
                        keys.get(32).label=(char)(label.charAt(0)+6)+"";
                        break;
                    case 4813:
                    case 4845:
                        keys.get(32).label=(char)(label.charAt(0)-2)+"";
                        break;
                    case 4933:
                        keys.get(32).label=(char)(label.charAt(0)-6)+"";
                        break;
                    default:
                        keys.get(32).label=(char)(label.charAt(0)+2)+"";
                }

            }
```

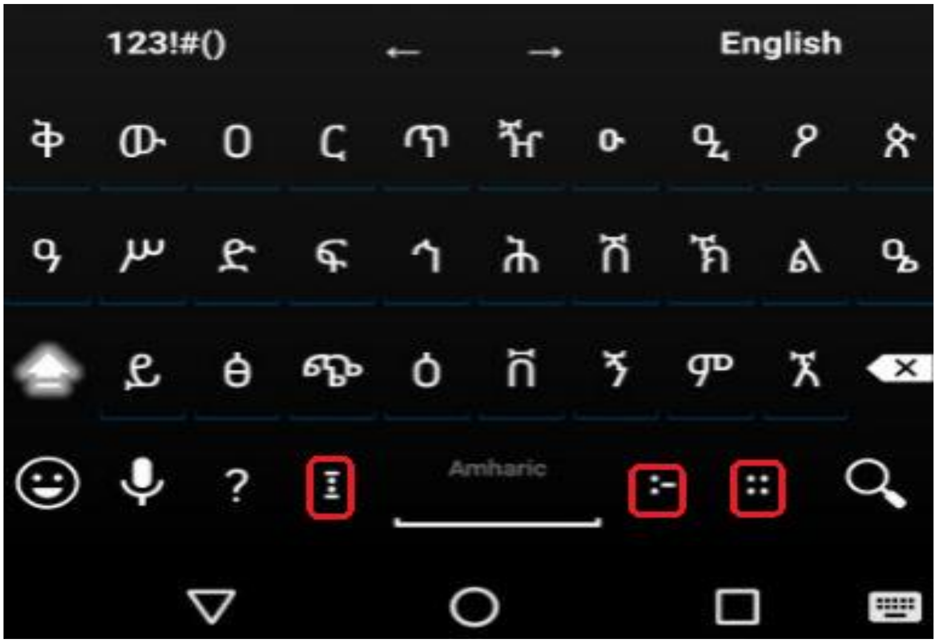


Figure 17: Portrait shifted Amharic Keyboard layout

The next fundamental thing to the display is – what to commit? For example the word "Fidel" is written by three letters in Amharic using Fi-de-l (ፊደል).If the next letter is the modified version of root letter committed, then delete the character before the cursor and commit the modified label, otherwise commit the label as shown in the excerpt code displayed below. The micro seconds difference between the two java statements creates an illusion in users mind that the root letter is transformed into its vowel form in a very fluid way due to human brain's speedy perception lack.

```java
if(key.label.hashCode()%8!=5 &&
ic.getTextBeforeCursor(1,0).hashCode()!=key.label.hashCode())
{
    ic.deleteSurroundingText(1, 0); //Delete the root letter
    ic.commitText(label, 0); //Commit the modified letter

}
else {


    ic.commitText(label, 1); // commit the root letter
}
```

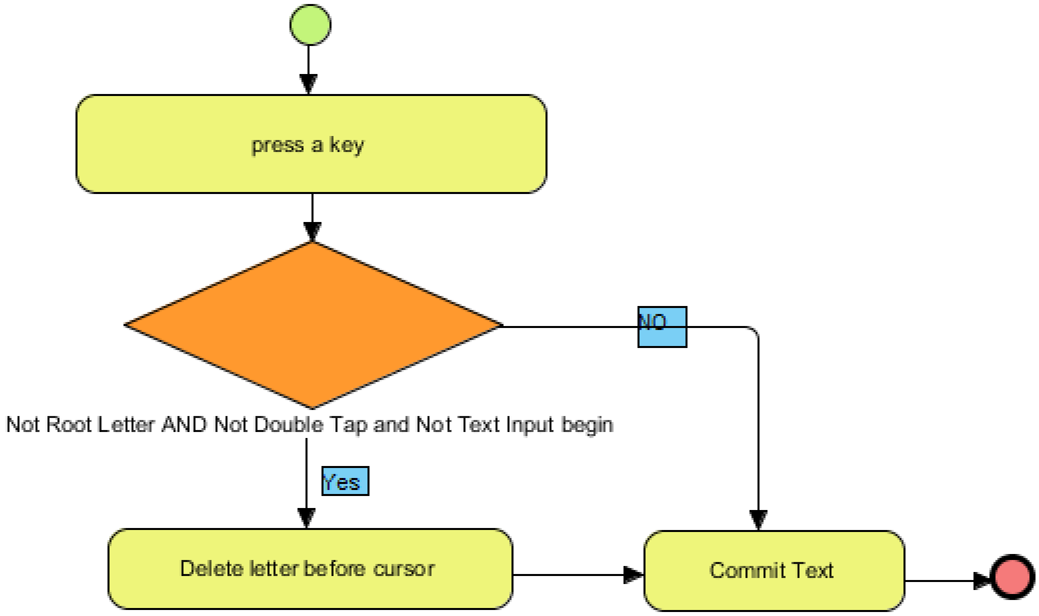
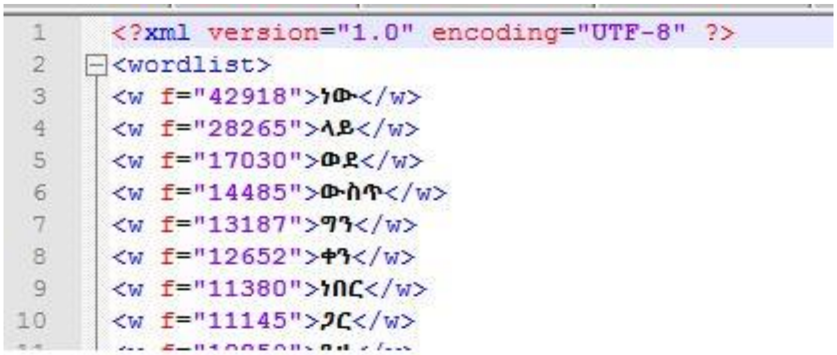

press a key
NO
Not Root Letter AND Not Double Tap and Not Text Input begin
Yes
Delete letter before cursor
Commit Text

Figure 18: Label modification handling

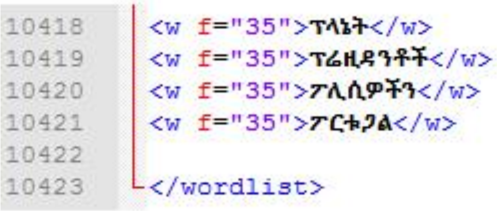
10. Dictionary preparation

The word frequency analysis is done using AntConc freeware on the above documents and got the rank of 350,000 Amharic words. The dictionary contains keywords and its frequency value as shown in the following xml except.



```
<?xml version="1.0" encoding="UTF-8" ?>
<wordlist>
<w f="42918">ነው</w>
<w f="28265">ላይ</w>
<w f="17030">ወደ</w>
<w f="14485">ውስጥ</w>
<w f="13187">ግን</w>
<w f="12652">ቀን</w>
<w f="11380">ነበር</w>
<w f="11145">ጋር</w>
```

....



```
<w f="35">ፕላኔት</w>
<w f="35">ፕሬዚዳንቶች</w>
<w f="35">ፖሊሲዎችን</w>
<w f="35">ፖርቱጋል</w>

</wordlist>
```

The dictionary gives suggestion and auto complete services, but no auto correction, which is beyond the scope of this project. The table at appendix C is an excerpt which shows the top most frequent words in Amharic literature. The single letters considered as most frequent words are mistakes due to unofficial reformations (using ና instead of እና), definite article and case makers.

IV. Evaluation and Tests

There are different types of metrics proposed by researchers to evaluate performance of typing tools. Key Stroke per Character (KSPC) is"Simply the ratio of length of input Stream to the length of transcribed text". [20]

$$KSPC=\frac{|IS|}{|T|}$$

The maximum number of keystrokes per character (KPC) is 3(Press shift+ press root+ press transformed vowel order), while the minimum and frequently used is obviously 1 (root letter). The main idea of bringing the most frequent keys to the default position (20 of the 34 root characters set ~58%) is to bring the number of keystrokes as minimum as possible.

The proposed keyboard tackles the shortcoming of existing keyboards and speeds up the typing process. Specifically

1. Double tap in proposed keyboard is exclusively used to type double letter words like it usually is supposed to be except the 16 key keypads. This avoids the unnecessary hustle and time spent during double tap.
2. Even if it is not as much as Ge'ez, there are some Amharic words have one of the vowel makers following the 6th order root letter. For example in existing keyboards in order to write 'ግዕዝ/ge'ez' the user have to go through the following steps.
    A. Type 'ግ'
    B. Press 'Space' otherwise the following vowel maker is going to transform ግ to ገ
    C. Type 'ዕ'

D. Go back and delete the space hence the input will be fixed from 'ግ ዕ' to 'ግዕ' and continue typing. This challenge of fixing an input stream that incorporates navigation keys, and edit functions (delete/back space) and/or cursor movement [20] is solved in the new keyboard.

3. The new keyboard is challenged to the 'Jah test' that was discussed at the requirement analysis section and able to write 'Atse hailesillasie' with 19 key strokes. In spite of Words per minute (WPM) , the most frequently used empirical measure of text entry performance in English text in addition to gestures per second (GPS) and Key strokes per second(KSPC) Ibid, [20, p. 106] could not be used due to lack of previous work on Amharic.

11. Conclusion and recommendations

In conclusion, in this project we recommend:

A. It is operationally impractical to have a single Ethiopic keyboard for all Ethiopian languages. Rather, there needs to be different localized keyboard layout for each and every Ethiopic-based language to improve orthography in each language specifically in Ethiopic generally. Since current days Amharic uses only 70.8% of the 384 total Ethiopic characters, adding those unnecessary keys in keyboard layout design will contribute nothing, but misspelling.

B. Linguists need to move forward in finding out mandatory characters needed for Amharic orthography from the Ethiopic Fidel. As any language in the world, evolution of spelling, pronunciation and letter representation could potentially vary over the centuries. So, the so called 'name-sake characters' are not name-sake when we see it at Ethiopic-level i.e. they have different sound in other languages unlike Amharic. However, users already displayed their own preference and usage bias as displayed in the graph below in the practical world.

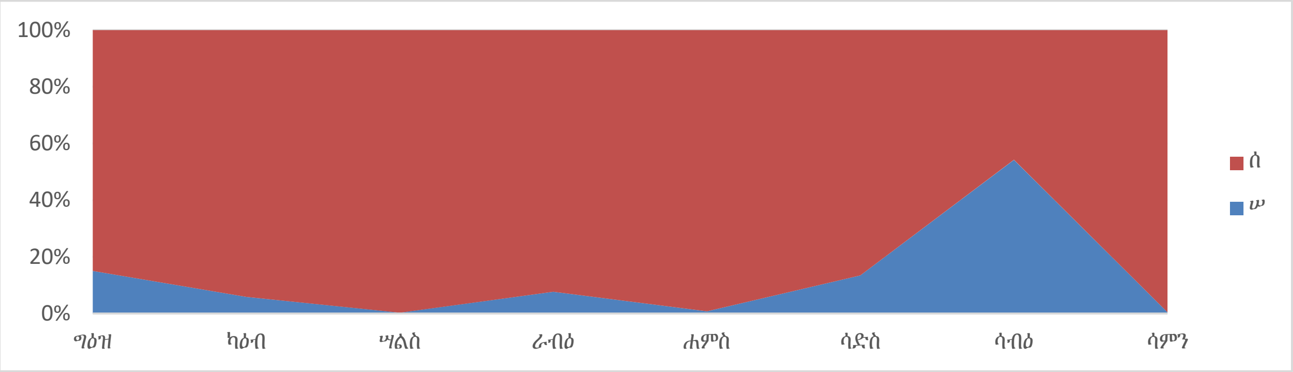


Figure 19: Usage pattern of the two "Se's"

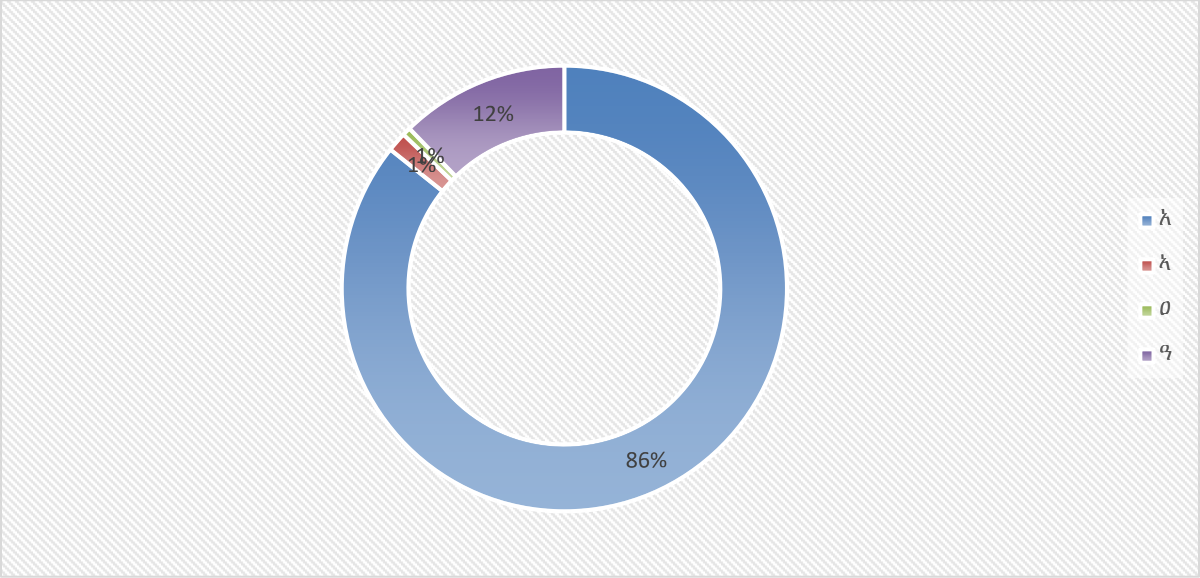


Figure 20: Usage patter of the four 'a's

C. Unicode reform on the 8th order of the Ethiopic table to be exclusively used for Amharic is very advisable. It is very ironic that the most frequent 8th order letters like ኋ and ቋ are included in the extended forms while they absolutely resemble in shape with the root characters and recurrently used as displayed in the following graph.

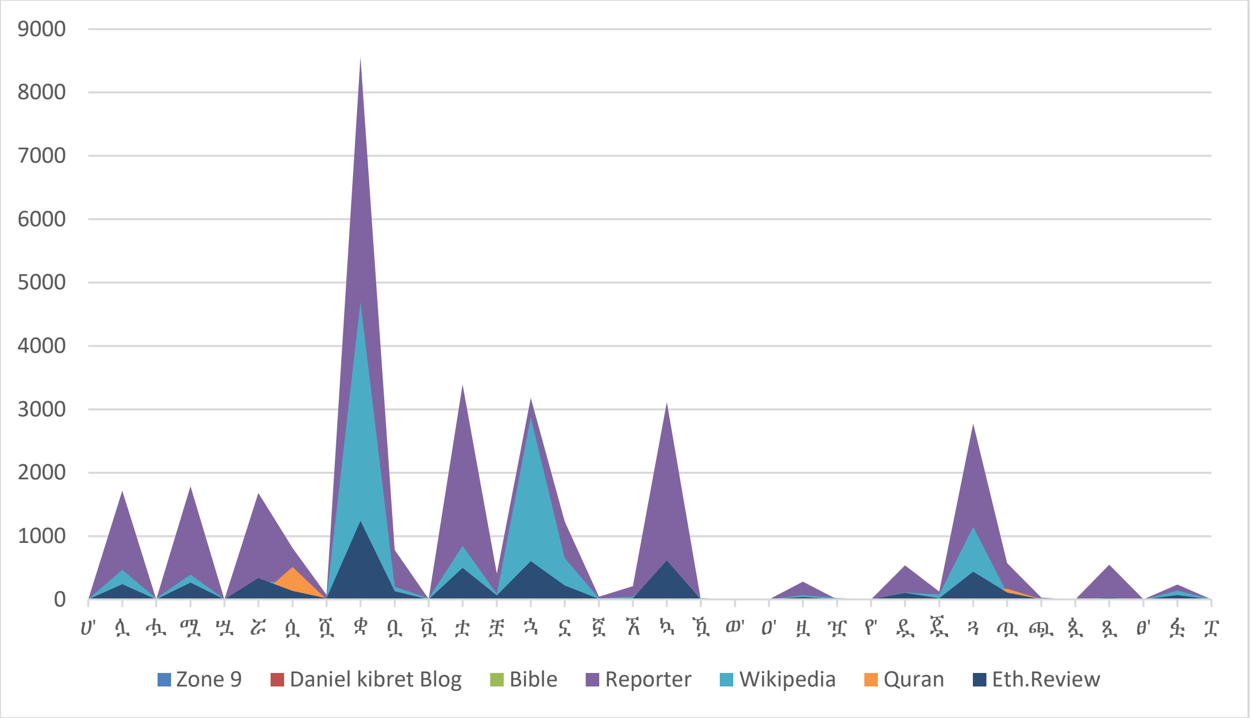


Figure 21:8th Order ('wa') frequency

D. A keyboard layout that considers the user typing experience in the widely used QWERTY keyboard [20] and character frequency improves spelling and accelerates typing speed. All keyboards that are not designed with this inconsideration are less user friendly and could potentially exacerbate misspelling.

E. The maximum keystroke per character needs to be 3. (3KSPC)

F. Input typing tools should minimize their contribution for cross-language transfer.

12. Future Work

Little research has already been done regarding the standardization of soft keyboards and their role in improving Amharic Orthography. More work is expected in Auto-correct, grammar and punctuation checks. Amharic voice recognition and voice typing should be done in the future as well.

Both fundamental ideas raised in this project i.e.

1) Standardization of soft keyboards used for inputting Ethiopic

2) Improving Orthography should be supported with additional scientific research and reforms.

Supporting different operating systems, apple ™'s IOS and Microsoft's Windows ™ is one of the immediate possible extension of this project.

Preparation of nationally used Amharic corpus and part of speech tagging will make the research process and the evaluation metrics effective.

V. References

[1] "The 2007 Population and Housing Census of Ethiopia," Population Census Commission, Addis Ababa, 2007.

[2] Internet society, "Global inernet report," Internet society, 2015.

[3] Buddle.com, "Ethiopia - Telecoms, Mobile and Broadband - Statistics and Analyses," 2015. [Online]. Available: https://www.budde.com.au/Research/Ethiopia-Telecoms-Mobile-and-Broadband-Statistics-and-Analyses.

[4] Pew Research, "Smartphone ownership rates skyrocket in many emerging economies, but digital divide remains," 22 February 2016. [Online]. Available: http://www.pewglobal.org/2016/02/22/smartphone-ownership-rates-skyrocket-in-many-emerging-economies-but-digital-divide-remains/.

[5] P. B. Yimam, "Ethiopic Writng System," *Ethiopian Studies,* p. 16, 1992.

[6] P. T. Daniels and W. Bright, The world's writing systems, New York Orxord: Oxford University Press, 1996.

[7] "Romanisation systems," November 1967. [Online]. Available: https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/499633/ROMANIZATION_SYSTEM_FOR_AMHARIC.PDF.

[8] Encyclopædia Britannica, "Amharic Language," 2016. [Online]. Available: http://www.britannica.com/topic/Amharic-language.

[9] B. Yimam, "(Ethiopian) Writing System," *journal of AAU Teacher's Association,* 1992).

[10] D. Yacob, "Application of the Double Metaphone Algorithm to Amharic Orthography," in *International Conference of Ethiopian Studies XV*, Hamburg, 2003.

[11] የኢትዮጵያ ቋንቋዎች ጥናትና ምርምር ማዕከል , አማርኛ መዝገበ ቃላት, አዲስ አበባ: አርቲስቲክ ማተሚያ ቤት, 1993.

[12] D. Getatchew, Interviewee, *writing systems.* [Interview].

[13] Microsoft, "Amharic Style Guide," Microsoft, 2011.

[14] D. Berhane, 9 January 2014. [Online]. Available: http://hornaffairs.com/am/2014/01/09/ethiopia-gebrekidan-desta-comment-misnaming-mekele/.

[15] I. Paz and H. M. Derso, "A morse code for the Amharic language," *Zede Journal,* pp. 47-51, 1965.

[16] S. T. Abate, W. Menzel and B. Tafila, "An Amharic Speech Corpus for Large Vocabulary Continuous Speech," 2005.

[17] Technical committee for Information Technology, "Ethiopic Keyboard Layout," in *Draft Ethiopian Standard (DES 13337:2007)*, Addis Ababa, 2007.

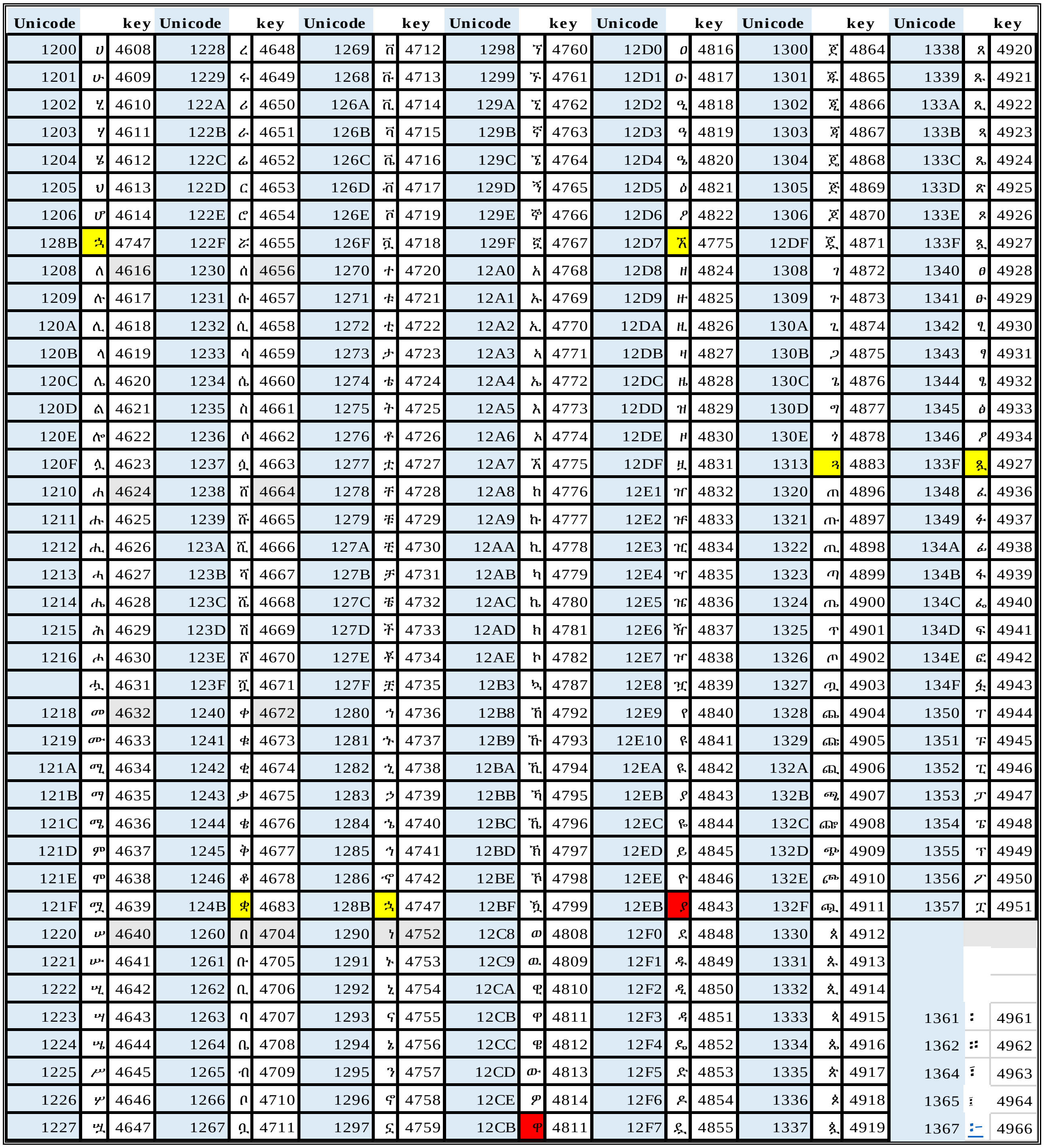
[18] A. A. Argaw and L. Asker, "Amharic-English Information Retrieval," Stockholm University/KTH, Stockholm , 2006.

[19] Oxford University , The Concise Oxford Dictionary, Oxford : Oxford University Press, 2004.

[20] A. S. Arif and W. Stuerzlinger, "Analysis of Text Entry Performance Metrics," in *Science and Technology for Humanity (TIC-STH), 2009 IEEE Toronto International Conference*, Toronto, 2009.

Appendices

Appendix A: Proposed Amharic alphabet



Unicode		key	Unicode		key	Unicode		key	Unicode		key	Unicode		key	Unicode		key	Unicode		key
1200	ሀ	4608	1228	ረ	4648	1269	ቨ	4712	1298	ኘ	4760	12D0	ዐ	4816	1300	ጀ	4864	1338	ጸ	4920
1201	ሁ	4609	1229	ሩ	4649	1268	ቩ	4713	1299	ኙ	4761	12D1	ዑ	4817	1301	ጁ	4865	1339	ጹ	4921
1202	ሂ	4610	122A	ሪ	4650	126A	ቪ	4714	129A	ኚ	4762	12D2	ዒ	4818	1302	ጂ	4866	133A	ጺ	4922
1203	ሃ	4611	122B	ራ	4651	126B	ቫ	4715	129B	ኛ	4763	12D3	ዓ	4819	1303	ጃ	4867	133B	ጻ	4923
1204	ሄ	4612	122C	ሬ	4652	126C	ቬ	4716	129C	ኜ	4764	12D4	ዔ	4820	1304	ጄ	4868	133C	ጼ	4924
1205	ህ	4613	122D	ር	4653	126D	ቭ	4717	129D	ኝ	4765	12D5	ዕ	4821	1305	ጅ	4869	133D	ጽ	4925
1206	ሆ	4614	122E	ሮ	4654	126E	ቮ	4719	129E	ኞ	4766	12D6	ዖ	4822	1306	ጆ	4870	133E	ጾ	4926
128B	ኋ	4747	122F	ሯ	4655	126F	ቯ	4718	129F	ኟ	4767	12D7	ኧ	4775	12DF	ጇ	4871	133F	ጿ	4927
1208	ለ	4616	1230	ሰ	4656	1270	ተ	4720	12A0	አ	4768	12D8	ዘ	4824	1308	ገ	4872	1340	ፀ	4928
1209	ሉ	4617	1231	ሱ	4657	1271	ቱ	4721	12A1	ኡ	4769	12D9	ዙ	4825	1309	ጉ	4873	1341	ፁ	4929
120A	ሊ	4618	1232	ሲ	4658	1272	ቲ	4722	12A2	ኢ	4770	12DA	ዚ	4826	130A	ጊ	4874	1342	ፂ	4930
120B	ላ	4619	1233	ሳ	4659	1273	ታ	4723	12A3	ኣ	4771	12DB	ዛ	4827	130B	ጋ	4875	1343	ፃ	4931
120C	ሌ	4620	1234	ሴ	4660	1274	ቴ	4724	12A4	ኤ	4772	12DC	ዜ	4828	130C	ጌ	4876	1344	ፄ	4932
120D	ል	4621	1235	ስ	4661	1275	ት	4725	12A5	እ	4773	12DD	ዝ	4829	130D	ግ	4877	1345	ፅ	4933
120E	ሎ	4622	1236	ሶ	4662	1276	ቶ	4726	12A6	ኦ	4774	12DE	ዞ	4830	130E	ጎ	4878	1346	ፆ	4934
120F	ሏ	4623	1237	ሷ	4663	1277	ቷ	4727	12A7	ኧ	4775	12DF	ዟ	4831	1313	ጓ	4883	133F	ጿ	4927
1210	ሐ	4624	1238	ሸ	4664	1278	ቸ	4728	12A8	ከ	4776	12E1	ዠ	4832	1320	ጠ	4896	1348	ፈ	4936
1211	ሑ	4625	1239	ሹ	4665	1279	ቹ	4729	12A9	ኩ	4777	12E2	ዡ	4833	1321	ጡ	4897	1349	ፉ	4937
1212	ሒ	4626	123A	ሺ	4666	127A	ቺ	4730	12AA	ኪ	4778	12E3	ዢ	4834	1322	ጢ	4898	134A	ፊ	4938
1213	ሓ	4627	123B	ሻ	4667	127B	ቻ	4731	12AB	ካ	4779	12E4	ዣ	4835	1323	ጣ	4899	134B	ፋ	4939
1214	ሔ	4628	123C	ሼ	4668	127C	ቼ	4732	12AC	ኬ	4780	12E5	ዤ	4836	1324	ጤ	4900	134C	ፌ	4940
1215	ሕ	4629	123D	ሽ	4669	127D	ች	4733	12AD	ክ	4781	12E6	ዥ	4837	1325	ጥ	4901	134D	ፍ	4941
1216	ሖ	4630	123E	ሾ	4670	127E	ቾ	4734	12AE	ኮ	4782	12E7	ዦ	4838	1326	ጦ	4902	134E	ፎ	4942
	ሗ	4631	123F	ሿ	4671	127F	ቿ	4735	12B3	ኳ	4787	12E8	ዧ	4839	1327	ጧ	4903	134F	ፏ	4943
1218	መ	4632	1240	ቀ	4672	1280	ኀ	4736	12B8	ኸ	4792	12E9	የ	4840	1328	ጨ	4904	1350	ፐ	4944
1219	ሙ	4633	1241	ቁ	4673	1281	ኁ	4737	12B9	ኹ	4793	12E10	ዩ	4841	1329	ጩ	4905	1351	ፑ	4945
121A	ሚ	4634	1242	ቂ	4674	1282	ኂ	4738	12BA	ኺ	4794	12EA	ዪ	4842	132A	ጪ	4906	1352	ፒ	4946
121B	ማ	4635	1243	ቃ	4675	1283	ኃ	4739	12BB	ኻ	4795	12EB	ያ	4843	132B	ጫ	4907	1353	ፓ	4947
121C	ሜ	4636	1244	ቄ	4676	1284	ኄ	4740	12BC	ኼ	4796	12EC	ዬ	4844	132C	ጬ	4908	1354	ፔ	4948
121D	ም	4637	1245	ቅ	4677	1285	ኅ	4741	12BD	ኽ	4797	12ED	ይ	4845	132D	ጭ	4909	1355	ፕ	4949
121E	ሞ	4638	1246	ቆ	4678	1286	ኆ	4742	12BE	ኾ	4798	12EE	ዮ	4846	132E	ጮ	4910	1356	ፖ	4950
121F	ሟ	4639	124B	ቋ	4683	128B	ኋ	4747	12BF	ዃ	4799	12EB	ያ	4843	132F	ጯ	4911	1357	ፗ	4951
1220	ሠ	4640	1260	በ	4704	1290	ነ	4752	12C8	ወ	4808	12F0	ደ	4848	1330	ጰ	4912			
1221	ሡ	4641	1261	ቡ	4705	1291	ኑ	4753	12C9	ዉ	4809	12F1	ዱ	4849	1331	ጱ	4913			
1222	ሢ	4642	1262	ቢ	4706	1292	ኒ	4754	12CA	ዊ	4810	12F2	ዲ	4850	1332	ጲ	4914			
1223	ሣ	4643	1263	ባ	4707	1293	ና	4755	12CB	ዋ	4811	12F3	ዳ	4851	1333	ጳ	4915	1361	፡	4961
1224	ሤ	4644	1264	ቤ	4708	1294	ኔ	4756	12CC	ዌ	4812	12F4	ዴ	4852	1334	ጴ	4916	1362	።	4962
1225	ሥ	4645	1265	ብ	4709	1295	ን	4757	12CD	ው	4813	12F5	ድ	4853	1335	ጵ	4917	1364	፣	4963
1226	ሦ	4646	1266	ቦ	4710	1296	ኖ	4758	12CE	ዎ	4814	12F6	ዶ	4854	1336	ጶ	4918	1365	፤	4964
1227	ሧ	4647	1267	ቧ	4711	1297	ኗ	4759	12CB	ዋ	4811	12F7	ዷ	4855	1337	ጷ	4919	1367	፦	4966

Table 6: Character frequency table of Amharic Fidel

Yellow shaded are repetitions from extended Ethiopic while red are for missing.

Appendix B: Relative letter frequency in Amharic alphabet

	ሀ	ለ	ሐ	መ	ሠ		ረ	ሰ	ሸ	ቀ	በ	ቨ	ተ	ቸ	ኀ	ነ	ኘ	አ
ግዕዝ	9883	294185	16353	328641	19931		149864	112608	13005	98278	397784	3053	245145	79099	544	230517	8890	325075
ካዕብ	71870	69848	4281	26480	2098		44577	33573	3577	23292	28772	80	52190	15151	27	35043	7373	1496
ሣልስ	2788	29103	1791	115291	119		67911	40189	3647	12955	35698	2656	18264	1521	13	15063	556	35711
ራብዕ	14636	161792	1965	158670	6252		140074	75288	19435	37000	132765	1346	135951	27235	13217	188693	44418	5399
ሐምስ	10081	31925	15968	15370	113		21823	14598	706	5709	30050	824	13589	3238	21	20127	274	25251
ሳድስ	95222	309967	35111	350075	45026		345764	290497	26264	80740	181445	947	454468	187485	9908	645664	25304	221226
ሳብዕ	73840	32318	201	35963	11184		43843	9444	2721	21057	19529	419	47642	2072	58	33167	8535	7176
ሳምን	0	2756	19	2720	8		2537	1874	116	16560	1269	24	5225	762	10346	2284	81	298

	ከ	ኸ	ወ	ዐ	ዘ	ዠ	የ	ደ	ጀ	ገ	ጠ	ጨ	ጰ	ጸ	ፀ	ፈ	ፐ
ግዕዝ	170941	2294	114177	2293	57060	327	384530	178591	24062	172873	80493	20425	59	9982	5725	60121	1002
ካዕብ	20226	533	6386	1208	13960	117	17853	19633	3796	33271	12803	1370	30	2670	870	9616	707
ሣልስ	9611	71	30853	335	47306	784	254	43614	7341	20352	2622	2991	255	140	83	22126	3536
ራብዕ	69513	348	63061	46468	25734	904	235826	76331	15804	64497	52240	14490	1517	9072	7069	26097	7935
ሐምስ	4224	1609	1398	1725	22045	138	5648	13930	2015	7525	4933	153	470	539	905	4703	1247
ሳድስ	102637	578	370551	35872	45908	3487	241630	154700	26136	168414	93061	15824	19544	23963	9680	77611	11041
ሳብዕ	19945	2859	43925	1737	5804	251	35922	13621	8949	15792	6619	2568	154	1290	597	7802	12758
ሳምን	5017	24	0	0	471	20	0	879	235	5427	1034	50	4	615	0	549	9

Table 7: Relative Amharic letters frequency

Appendix C: Top 100 frequent words, an excerpt from Amharic dictionary

Rank	Freq	word	Rank	Freq	word	Rank	Freq	word	Rank	Freq	word
1	42925	ነው	29	5338	ከ	57	3787	ማለት	85	2830	በኢትዮጵያ
2	28274	ላይ	30	5307	መንግሥት	58	3758	እግዚአብሔር	86	2813	መረጃ
3	17031	ወደ	31	5286	ዓመት	59	3725	ብዙ	87	2809	በፊት
4	14486	ውስጥ	32	5026	ድረስ	60	3677	ብሎ	88	2741	ሕዝብ
5	13193	ግን	33	5002	አቶ	61	3578	ዘመን	89	2649	አገሮች
6	12654	ቀን	34	4887	በዚህ	62	3431	አለ	90	2634	አይደለም
7	11380	ነበር	35	4783	እኤአ	63	3412	መካከል	91	2617	መሠረት
8	11147	ጋር	36	4665	ሥራ	64	3314	ሁኔታ	92	2611	በኩል
9	10860	ጊዜ	37	4638	የኢትዮጵያ	65	3222	ልጆች	93	2609	ወቅት
10	10009	ሁሉ	38	4598	ልጅ	66	3215	እንጂ	94	2602	ቦታ
11	9612	እንደ	39	4470	ሰዎች	67	3208	እንዲህ	95	2596	ማስታወቂያ
12	9198	ነገር	40	4401	ምን	68	3159	ምክንያት	96	2586	ችግር
13	8703	ወይም	41	4371	ኢትዮጵያ	69	3147	ሁለት	97	2578	የሚል
14	8336	ይህ	42	4343	ሆነ	70	3145	ብር	98	2534	ስዕል
15	8005	ደግሞ	43	4335	ገጽ	71	3138	ጉዳይ	99	2492	ታሪክ
16	7733	ቤት	44	4283	ስም	72	3102	አዲስ	100	2487	ዓለም
17	7596	ቁጥር	45	4280	ም	73	3095	ደረጃ	101	2473	ምንም
18	7402	በ	46	4247	ክፍል	74	3079	ሆኖ	102	2456	ሌላ
19	7027	እና	47	4228	ከተማ	75	3044	ዋና	103	2455	ይህን
20	6911	አንድ	48	4222	ያለው	76	3031	እኔ	104	2424	ሳይሆን
21	6519	በኋላ	49	4199	መንገድ	77	2990	የሥራ	105	2422	መዋቅር
22	6481	ሰው	50	4135	በላይ	78	2968	ንጉሥ	106	2417	አለው
23	6306	ዘንድ	51	4028	ዓ	79	2968	ጀምሮ	107	2399	ቃል
24	6241	ዓም	52	3906	ከዚህ	80	2958	ከፍተኛ	108	2395	ያለ
25	6142	ናቸው	53	3834	አገር	81	2901	የ	109	2343	አሉ
26	5911	ብቻ	54	3831	ስለ	82	2898	አሁን	110	2343	እንዲሁም
27	5907	እስከ	55	3823	መሆኑን	83	2876	ቅፅ	111	2315	ሰዓት
28	5579	ሲሆን	56	3811	አበባ	84	2875	ዓመታት	112	2304	ዋጋ

Table 8: Top 100 frequent words, an excerpt from Amharic dictionary