

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

Acknowledgment.....	2
List of Figures	5
List of Tables	5
Apstrakt.....	6
Abstract.....	7
Introduction.....	8
I. Project Description.....	10
1. Project description	10
a) Purpose	10
b) Scope	10
c) Stakeholders.....	10
d) Naming conventions and Definitions	11
2. Overview	11
3. Problem statement.....	14
5. Literature review.....	22
6. Proposed system	23
II. Requirement Analysis.....	25
7. Business Requirement.....	25
7.1. Character Frequency analysis.....	25
8. System Requirement analysis	33
Use case	33
III. Design and implementation	35
9. Designing keyboard layout.....	35
10. Dictionary preparation.....	39
IV. Evaluation and Tests.....	40
11. Conclusion and recommendations.....	41
12. Future Work.....	43
V. References.....	45
Appendices.....	47
Appendix A: Proposed Amharic alphabet	47
Appendix B: Relative letter frequency in Amharic alphabet.....	48

Appendix C: Top 100 frequent words, an excerpt from Amharic dictionary 49

List of Figures

Figure 1: Multiling Keyboard.....	18
Figure 2: Ha le ha me based Amharic keyboard	18
Figure 3: Query based Agerigna keyboard.....	18
Figure 4: amharickeboard.com	19
Figure 5: Searching in Amharic online user experience	20
Figure 6: Sample Ethio-Telecom tele marketing message	21
Figure 7Ethiopic Unicode standard excerpt	22
Figure 8: Geez 'go' (21%) vs Amharic 'go' (79%) mixed usage share in the Amharic corpus	27
Figure 9.Relative frequency percentage of Amharic letters.	29
Figure 10: Relative frequency of letters ordered by frequency	29
Figure 11. Relative Frequency of Amharic Vowel Transformations.....	30
Figure 12: Character transformation and usage frequency.....	31
Figure 14: Usage pattern of the four "Ha's"	32
Figure 15: Usage pattern of two "A's".....	32
Figure 16: Usage pattern of two "Tse's"	33
Figure 18: Landscape default Amharic Keyboard layout.....	36
Figure 19: Portrait shifted Amharic Keyboard layout.....	38
Figure 20: Label modification handling	39
Figure 21: Usage pattern of the two "Se's"	42
Figure 22:8th Order ('wa') frequency.....	43

List of Tables

Table 1 Example of transliteration of syllables with vowels and transformation.....	13
Table 2 Jah Test: Worst case spelling error example to write "Atse Haile Selassie" where only option 135 highlighted in blue is canonically correct.....	16
Table 3: Proposed keyboard's total number of key labels.....	24
Table 4: Amharic root characters relative frequency ordered by Unicode standard and/or Ethiopic alphabet.....	28
Table 5. Amharic root characters relative frequency ordered by character frequency	28
Table 6: Character frequency table of Amharic Fidel	47
Table 7: Relative Amharic letters frequency	48
Table 8: Top 100 frequent words, an excerpt from Amharic dictionary	50

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 and Eritrean 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 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 (U-T 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 alphabetic¹ or syllabic² writing system. It has 350 characters that are used for writing many

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

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 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

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 Sabeian 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]

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 $26 \times 5 = 130$ 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	1 st	2 nd	3 rd	4 th	5 th	6 th	7 th	8 th
Ge'ez	ግዕዝ / ge'ez	ክዕብ / ka`lb	ሣልስ / sals	ራብዕ / rab`I	ኃምስ / hams	ሰድስ / sads	ሰብዕ / sab`I	ሰምን / samin
Ethiopic	በ	ቡ	ቢ	ባ	ቤ	ብ	ቦ	ቧ
Vowel	ɜ	u	i	a	e	ə	o	‘wa
Latin	Bɜ	Bu	Bi	Ba	Be	B	Bo	B’wa
Example	B ird	B ull	B ill	B at	B et	B rown	B ob	Q uality ³
Ethiopic	በ rd	ቡll	ቢll	ባ t	ቤ t	ብrown	ቦ b	ቧlity

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, 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

³ ‘wa’ is a diphthong i.e. a vowel with two different targets

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

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

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 sällasé" which will be referenced as 'Jah ⁵test' 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

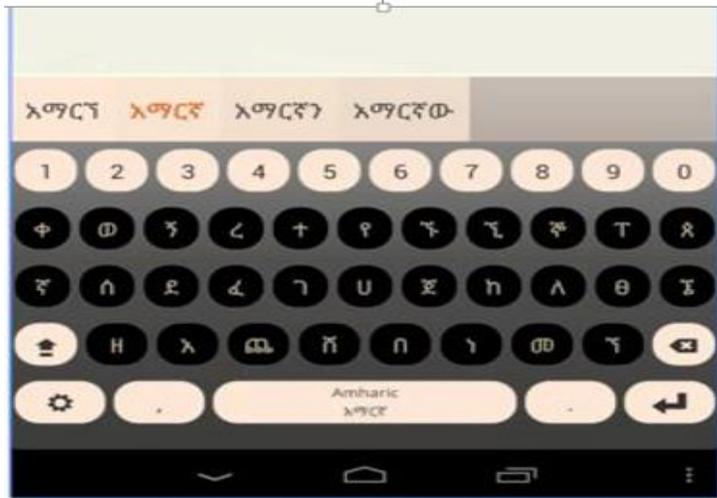


Figure 1: Multiling Keyboard

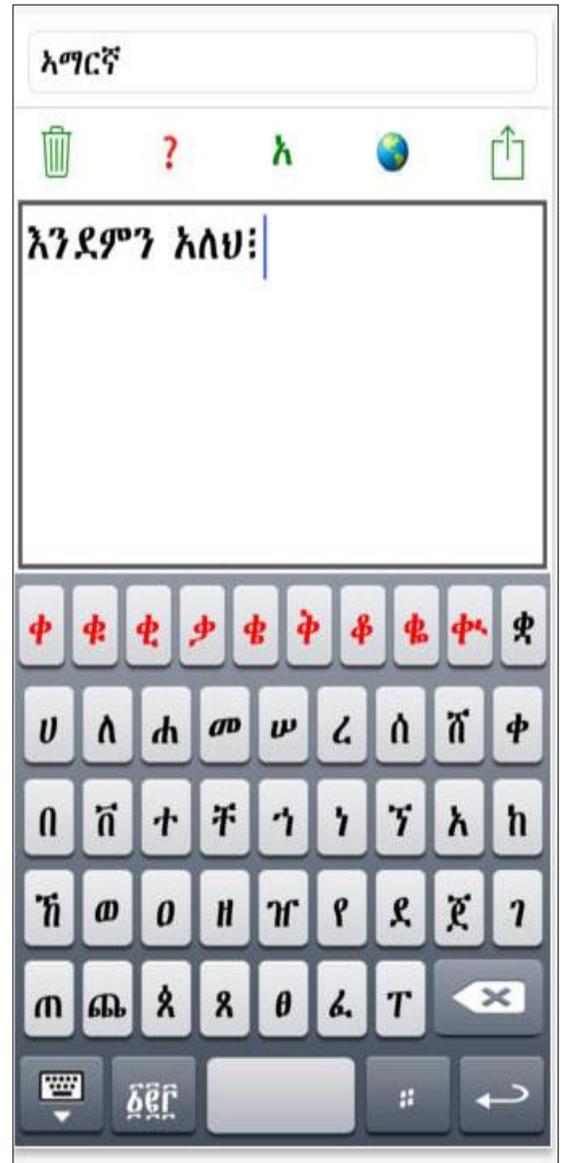


Figure 2: Ha le ha me based Amharic keyboard



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!

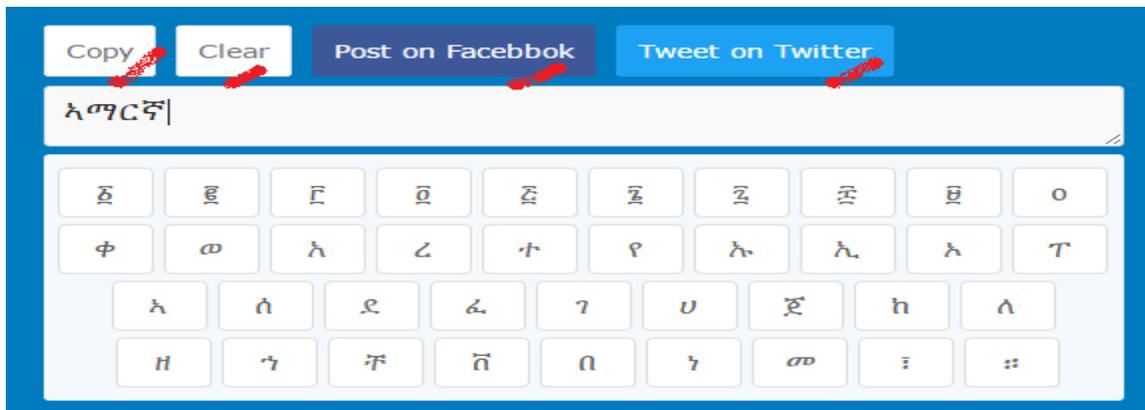


Figure 4: amharickeyboard.com

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.

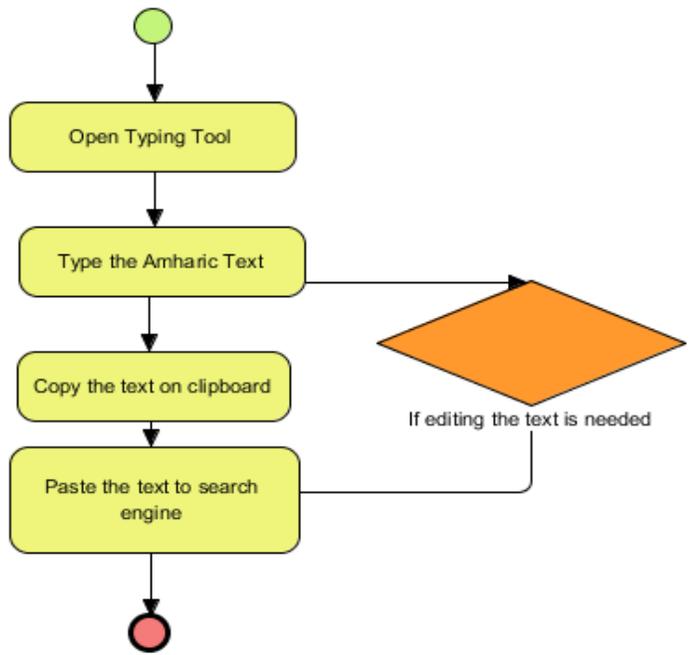


Figure 5: Searching in Amharic online user experience

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.

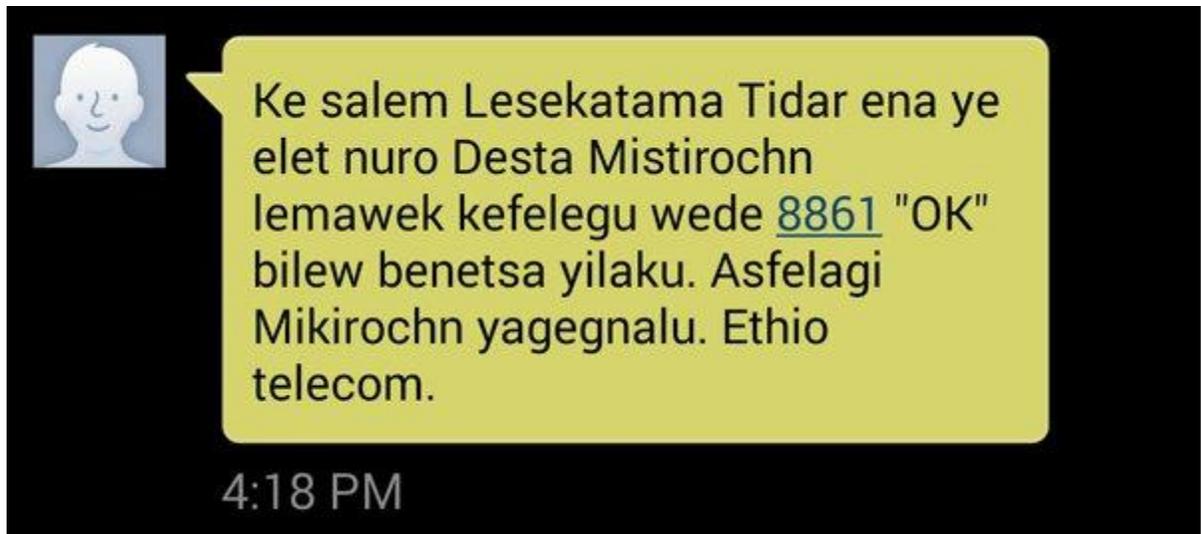


Figure 6: Sample Ethio-Telecom tele marketing message

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 ሆ (Nailo Seharan), ኘ (Ge'ez), ቆ (Agewigna) etc. ቐ (Tigirigna) and ደ (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 ሙቅ ለ/Mäqälle (Mekelle) vs ሙቐ ለ/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. ማቀሌ/Mäqälle is nationally used by Amharic native speakers except most articles of Reporter magazine that were submitted by Tigrean origin journalists who wrote ማቐለ / 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	ሐ	ሐ	ሐ	ሐ	ሐ	ሐ	ሐ	ሐ	ሐ
1210	ሐ	ሐ	ሐ	ሐ	ሐ	ሐ	ሐ	ሐ	ሐ	ሐ	ሐ	ሐ	ሐ	ሐ	ሐ	ሐ
1220	ሐ	ሐ	ሐ	ሐ	ሐ	ሐ	ሐ	ሐ	ሐ	ሐ	ሐ	ሐ	ሐ	ሐ	ሐ	ሐ
1230	ሐ	ሐ	ሐ	ሐ	ሐ	ሐ	ሐ	ሐ	ሐ	ሐ	ሐ	ሐ	ሐ	ሐ	ሐ	ሐ
1240	ሐ	ሐ	ሐ	ሐ	ሐ	ሐ	ሐ	ሐ	ሐ	ሐ	ሐ	ሐ	ሐ	ሐ	ሐ	ሐ
1250	ሐ	ሐ	ሐ	ሐ	ሐ	ሐ	ሐ	ሐ	ሐ	ሐ	ሐ	ሐ	ሐ	ሐ	ሐ	ሐ
1260	ሐ	ሐ	ሐ	ሐ	ሐ	ሐ	ሐ	ሐ	ሐ	ሐ	ሐ	ሐ	ሐ	ሐ	ሐ	ሐ

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 ለ 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 ((ሆ,ቆ,ኄ,ገ,ከ,ዎ,ዳ,ጎ)) 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 (ከጌ, ቈ, ኸ, ኸጌ) 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 “አ ኩስ ግግ” has only four Google™ pages search result, whereas አ ከ ሱግግ, 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 ሽ 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 ኃ, ኣ, ደ in their analogues Amharic namesake characters and lack of ‘wa’ sound in ወ and ዩ .

Characters	NO of Main Sounds	Vowels	Exception	Total Syllables
Main [U -ፒ]	28	8	2(ወ,ዩ)	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 Bible⁶, Quran⁷, Daniel Kibret views blog⁸, Zone 9 blog⁹, 50 copies of Reporter Magazine from November 2013-June 2014¹⁰, Amharic Wikipedia August articles dump and Ethio media website¹¹. The full content of the first two religious books is chosen for their rich content and their direct influence on orthography of Amharic by their respective relationship with Hebrew, Greek

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.ethiomeia.com

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 environment used to develop the proposed keyboard.

```
iconv -f utf-8 -t utf-16 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 ከ and ገ characters. For example, መኮንን VS መኮንን / mekönnin and ገንዳር VS ጉንዳር / Gondär.

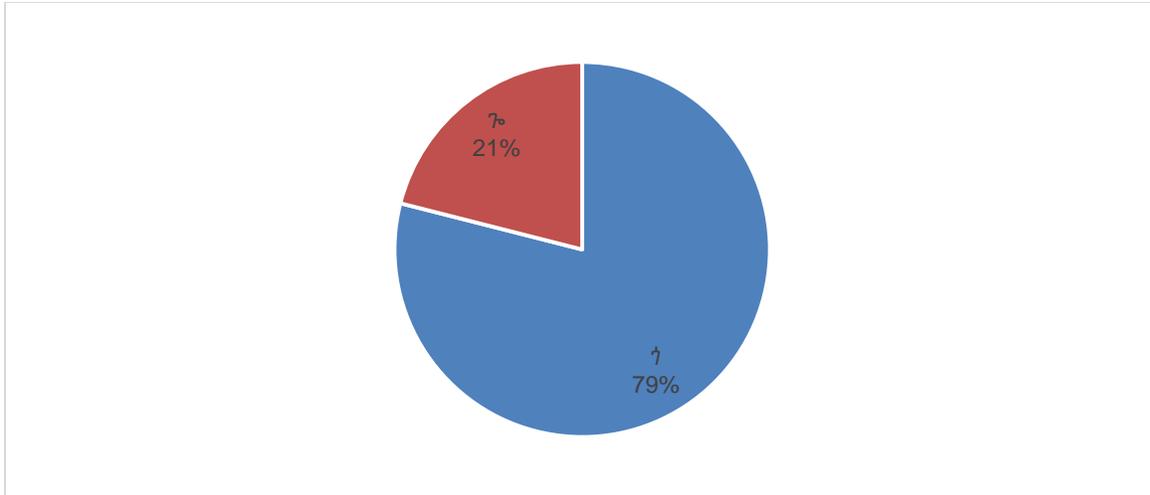


Figure 8: Geez 'go' (21%) vs Amharic 'go' (79%) mixed usage share in the Amharic corpus

In addition, very few non Amharic characters like ቐ, ቒ 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.html> , 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	ኸ	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	ጫ	57,871	0.47
30	አ	22,029	0.18
31	ጸ	48,271	0.39
32	ፀ	24,929	0.20
33	ረ	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	ረ	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	ጫ	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	አ	22,029	0.18
32	ኸ	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

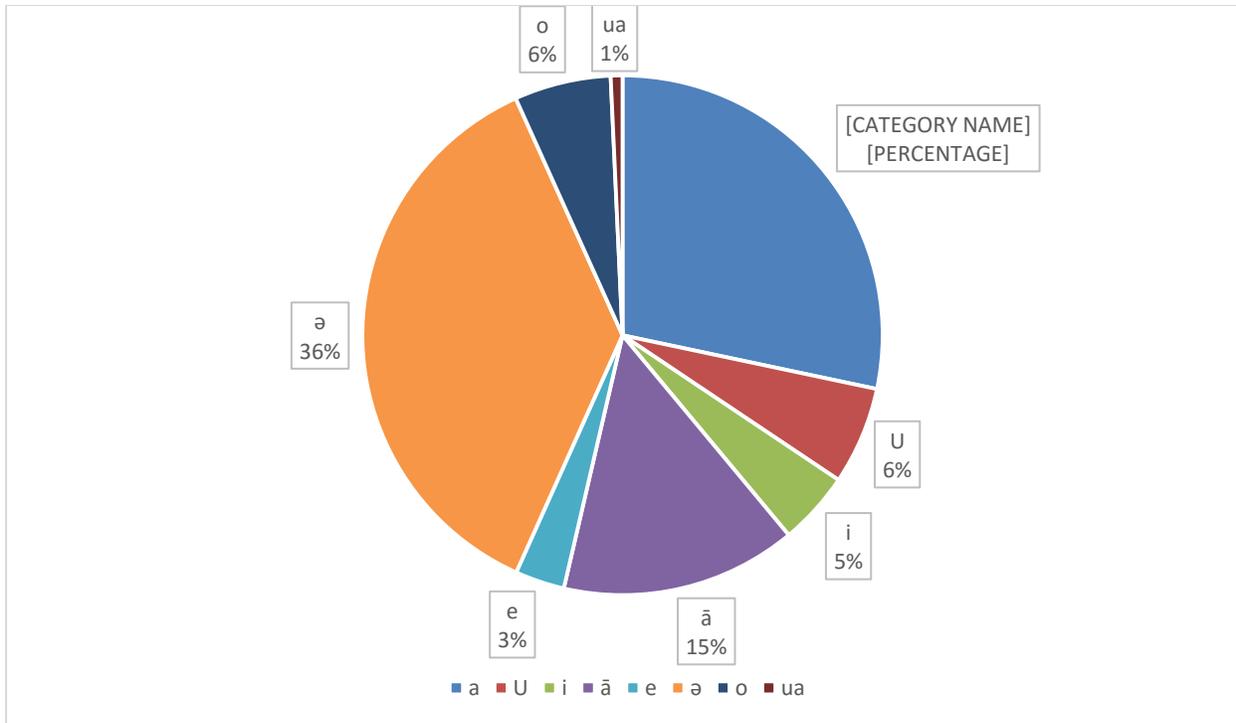


Figure 11. Relative Frequency of Amharic Vowel Transformations

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 with 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.

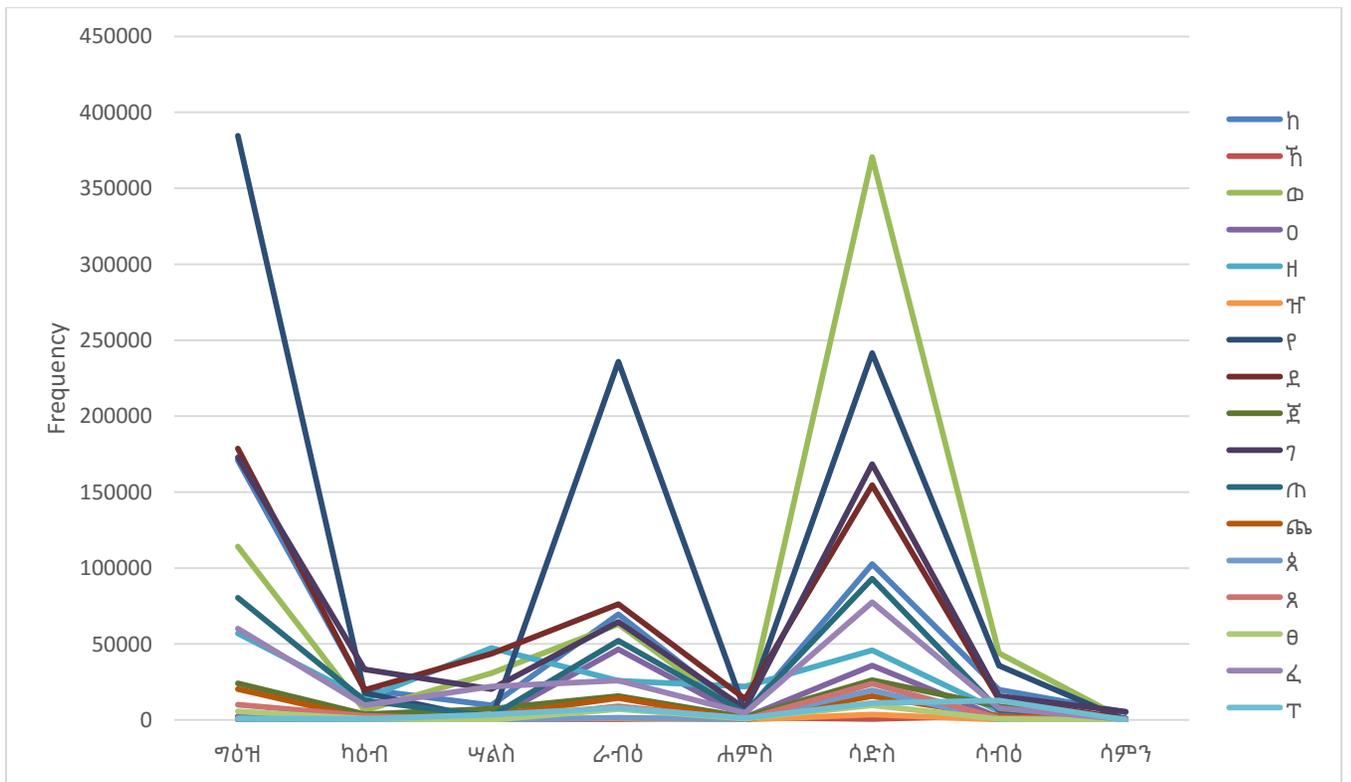
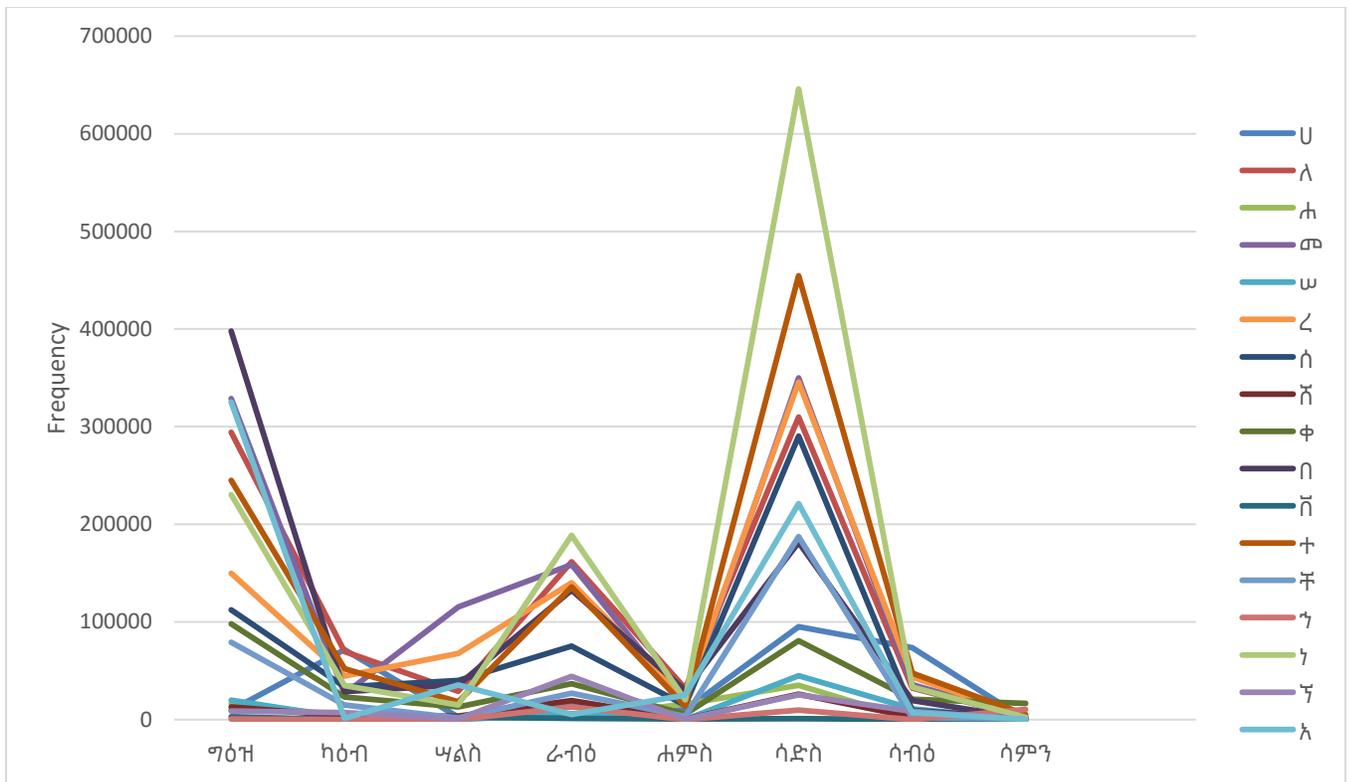


Figure 12: Character transformation and usage frequency

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 ኡ is used 90% of the time. Second, due to poor Unicode standardization, there is, for example, one ኣ (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.

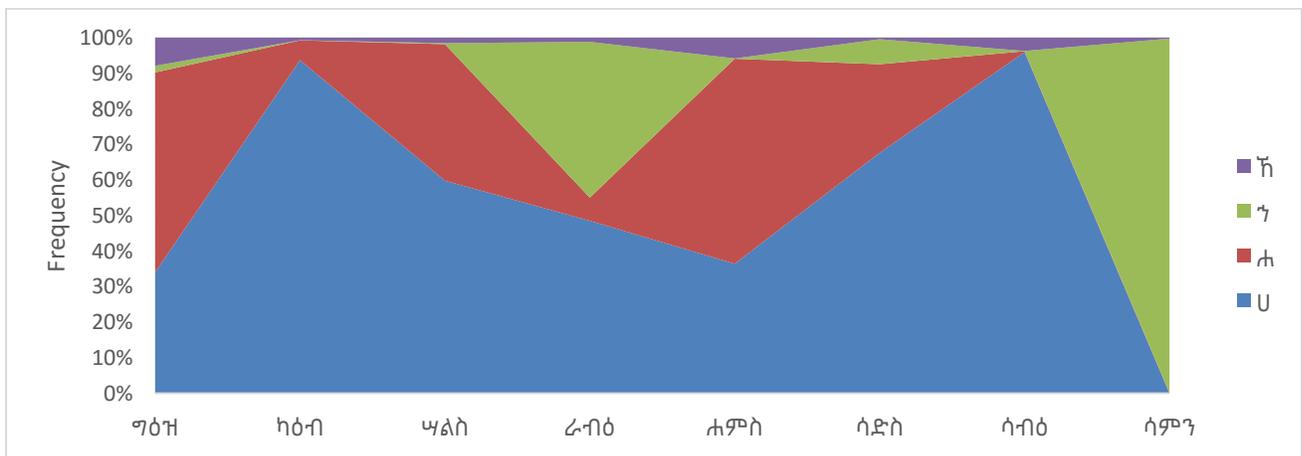


Figure 13: Usage pattern of the four "Ha's"

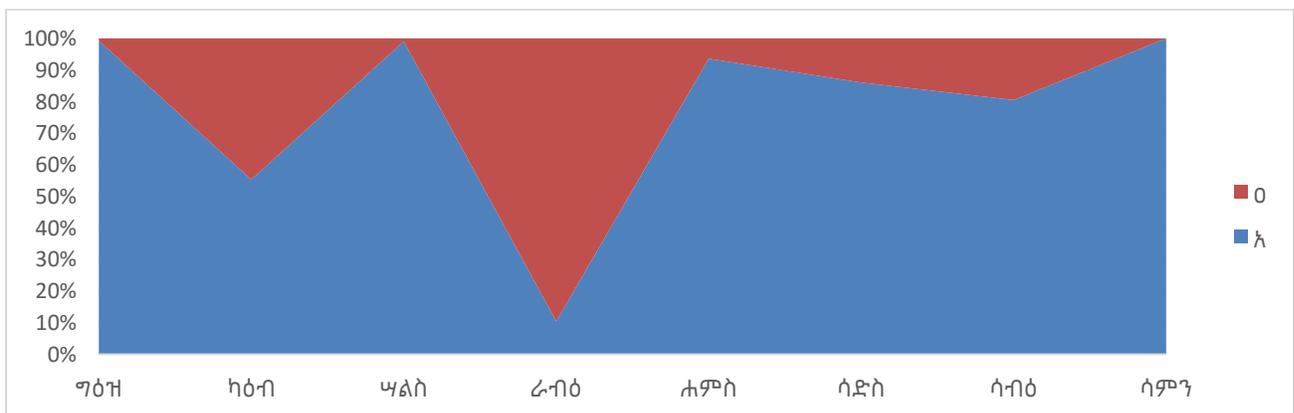


Figure 14: Usage pattern of two "A's"

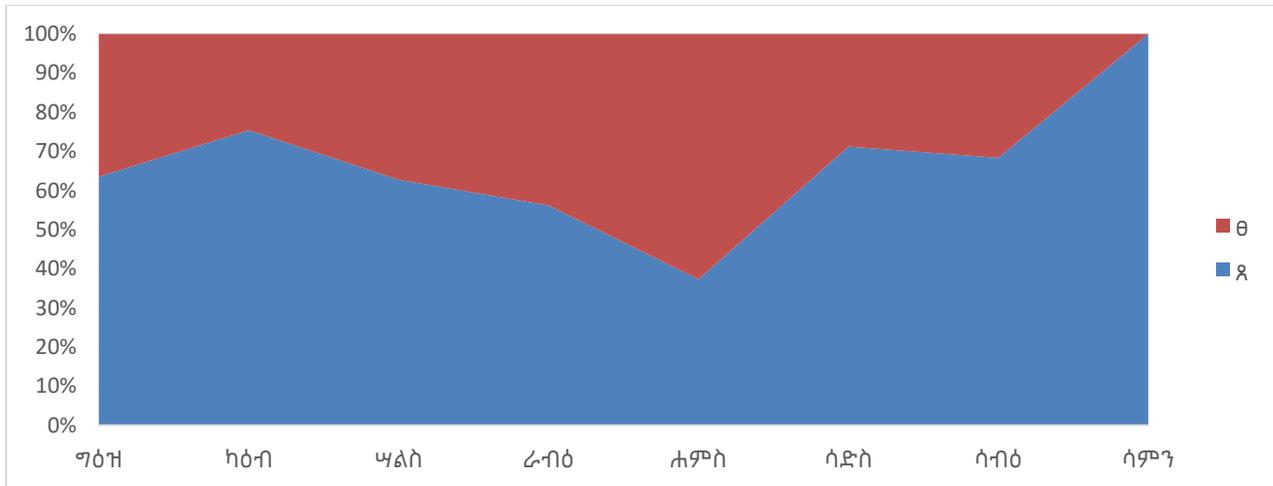


Figure 15: Usage pattern of two "Tse's"

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 keyboard 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. User clicks the Amharic tab from the right corner of the keyboard
3. User clicks the 6th main character 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 'Н'. 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 “ቅ ወ እ ር ት ይ ኡ ኢ ኦ ፕ አ ስ ድ ፍ ግ ሀ ጅ ክ ል ዝ ፀ ች ሽ ብ ን ም” and only an additional key to hold the ‘ua’ transformation of all letters. Shift/ Caps lock will be “ቅ ወ ፊ ር ጥ ይ ዑ ዲ ዖ ጵ ዐ ስ ድ ፍ ግ ሀ ሽ ክ ል ዝ ፀ ች ሽ ብ ን ም. 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="ፍ" />
    <Key android:codes="4829" android:keyLabel="" android:popupCharacters="ኸ" />
    <Key android:codes="4769" android:keyLabel="" android:popupCharacters="ዑ" />
    <Key android:codes="4770" android:keyLabel="" android:popupCharacters="ዲ" />
    <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 ቆ, ቇ, ቈ, ቉, ቊ, ቋ, ቌ, ቍ, 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.is-
Shifted())?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-
Code() <=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)+"";
    }
    }
}

```

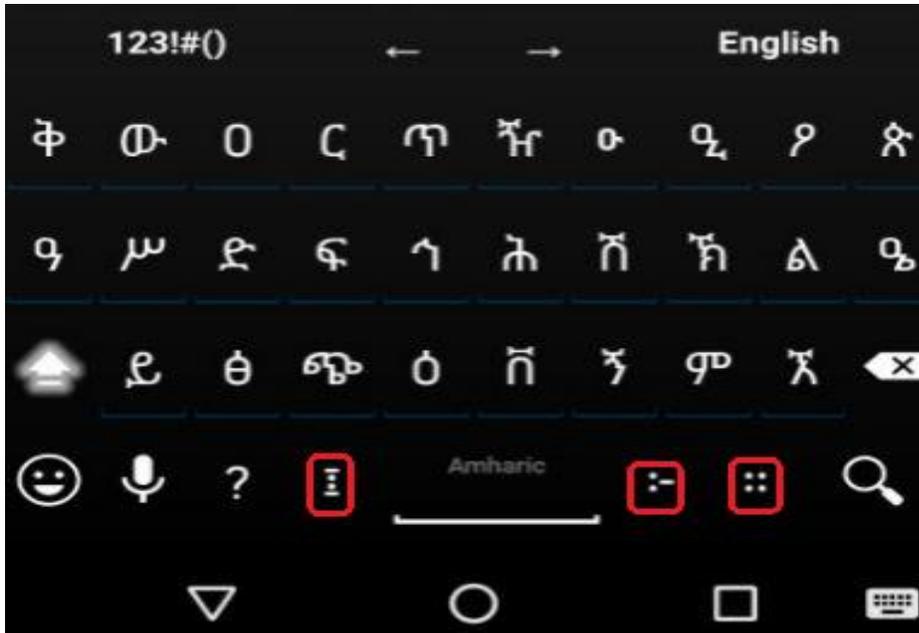


Figure 17: Portrait shifted Amharic Keyboard layout

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.

```

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
}

```

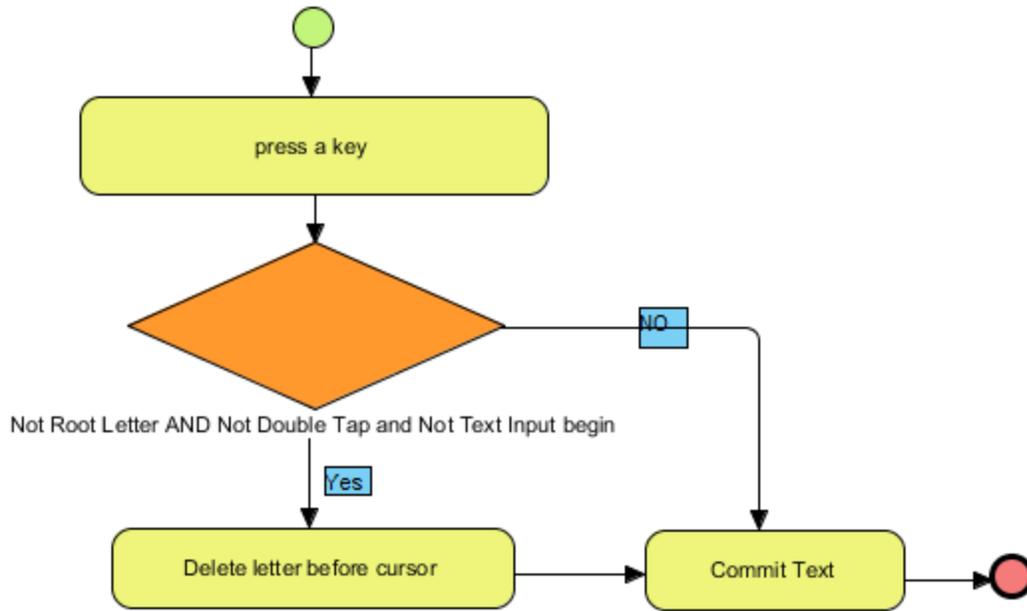


Figure 18: Label modification handling

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.

```

1  <?xml version="1.0" encoding="UTF-8" ?>
2  <wordlist>
3  <w f="42918">ነው</w>
4  <w f="28265">ለይ</w>
5  <w f="17030">ወደ</w>
6  <w f="14485">ውስጥ</w>
7  <w f="13187">ግን</w>
8  <w f="12652">ቀን</w>
9  <w f="11380">ነበር</w>
10 <w f="11145">ጋር</w>
11 ... f="10050">ወደ</w>
...
10418 <w f="35">ጥላይት</w>
10419 <w f="35">ጥረዘዳንቶች</w>
10420 <w f="35">ጥረዳዎች</w>
10421 <w f="35">ጥርታጋል</w>
10422
10423 </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 እ ና), 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]

$$\text{KSPC} = \frac{|I|}{|T|}$$

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 ግ
 - C. Type 'ó'
 - D. Go back and delete the space hence the input will be fixed from 'ግ ó' to 'ግó' 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 hailesillasié' 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.

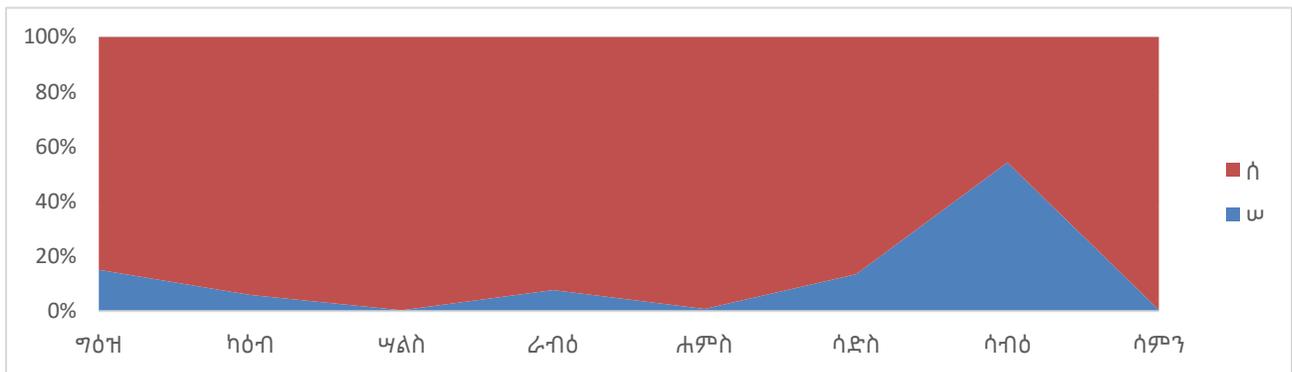


Figure 19: Usage pattern of the two "Se's"

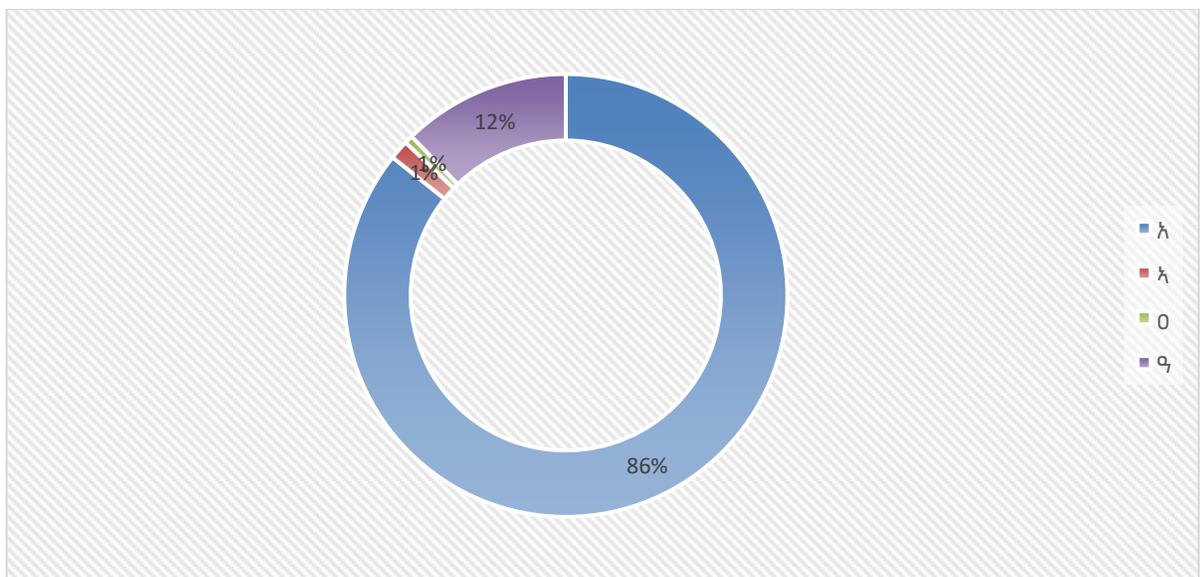


Figure 20: Usage patter of the four 'a's

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 ኣ and አ are included in the extended forms while they absolutely resemble in shape with the root characters and recurrently used as displayed in the following

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

Supporting different operating systems, apple TM's IOS and Microsoft's Windows TM 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 internet 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																									
1200	ሀ	4608	ሀ	1228	ሀ	4648	ሀ	1269	ሀ	4712	ሀ	1298	ሀ	4760	ሀ	12D0	ሀ	4816	ሀ	1300	ሀ	4864	ሀ	1338	ሀ	4920
1201	ሁ	4609	ሁ	1229	ሁ	4649	ሁ	1268	ሁ	4713	ሁ	1299	ሁ	4761	ሁ	12D1	ሁ	4817	ሁ	1301	ሁ	4865	ሁ	1339	ሁ	4921
1202	ሂ	4610	ሂ	122A	ሂ	4650	ሂ	126A	ሂ	4714	ሂ	129A	ሂ	4762	ሂ	12D2	ሂ	4818	ሂ	1302	ሂ	4866	ሂ	133A	ሂ	4922
1203	ሃ	4611	ሃ	122B	ሃ	4651	ሃ	126B	ሃ	4715	ሃ	129B	ሃ	4763	ሃ	12D3	ሃ	4819	ሃ	1303	ሃ	4867	ሃ	133B	ሃ	4923
1204	ሄ	4612	ሄ	122C	ሄ	4652	ሄ	126C	ሄ	4716	ሄ	129C	ሄ	4764	ሄ	12D4	ሄ	4820	ሄ	1304	ሄ	4868	ሄ	133C	ሄ	4924
1205	ህ	4613	ህ	122D	ህ	4653	ህ	126D	ህ	4717	ህ	129D	ህ	4765	ህ	12D5	ህ	4821	ህ	1305	ህ	4869	ህ	133D	ህ	4925
1206	ሆ	4614	ሆ	122E	ሆ	4654	ሆ	126E	ሆ	4719	ሆ	129E	ሆ	4766	ሆ	12D6	ሆ	4822	ሆ	1306	ሆ	4870	ሆ	133E	ሆ	4926
128B	ሐ	4747	ሐ	122F	ሐ	4655	ሐ	126F	ሐ	4718	ሐ	129F	ሐ	4767	ሐ	12D7	ሐ	4775	ሐ	12DF	ሐ	4871	ሐ	133F	ሐ	4927
1208	ለ	4616	ለ	1230	ለ	4656	ለ	1270	ለ	4720	ለ	12A0	ለ	4768	ለ	12D8	ለ	4824	ለ	1308	ለ	4872	ለ	1340	ለ	4928
1209	ሉ	4617	ሉ	1231	ሉ	4657	ሉ	1271	ሉ	4721	ሉ	12A1	ሉ	4769	ሉ	12D9	ሉ	4825	ሉ	1309	ሉ	4873	ሉ	1341	ሉ	4929
120A	ሊ	4618	ሊ	1232	ሊ	4658	ሊ	1272	ሊ	4722	ሊ	12A2	ሊ	4770	ሊ	12DA	ሊ	4826	ሊ	130A	ሊ	4874	ሊ	1342	ሊ	4930
120B	ላ	4619	ላ	1233	ላ	4659	ላ	1273	ላ	4723	ላ	12A3	ላ	4771	ላ	12DB	ላ	4827	ላ	130B	ላ	4875	ላ	1343	ላ	4931
120C	ሌ	4620	ሌ	1234	ሌ	4660	ሌ	1274	ሌ	4724	ሌ	12A4	ሌ	4772	ሌ	12DC	ሌ	4828	ሌ	130C	ሌ	4876	ሌ	1344	ሌ	4932
120D	ል	4621	ል	1235	ል	4661	ል	1275	ል	4725	ል	12A5	ል	4773	ል	12DD	ል	4829	ል	130D	ል	4877	ል	1345	ል	4933
120E	ሎ	4622	ሎ	1236	ሎ	4662	ሎ	1276	ሎ	4726	ሎ	12A6	ሎ	4774	ሎ	12DE	ሎ	4830	ሎ	130E	ሎ	4878	ሎ	1346	ሎ	4934
120F	ሊ	4623	ሊ	1237	ሊ	4663	ሊ	1277	ሊ	4727	ሊ	12A7	ሊ	4775	ሊ	12DF	ሊ	4831	ሊ	1313	ሊ	4883	ሊ	133F	ሊ	4927
1210	ሐ	4624	ሐ	1238	ሐ	4664	ሐ	1278	ሐ	4728	ሐ	12A8	ሐ	4776	ሐ	12E1	ሐ	4832	ሐ	1320	ሐ	4896	ሐ	1348	ሐ	4936
1211	ሐ	4625	ሐ	1239	ሐ	4665	ሐ	1279	ሐ	4729	ሐ	12A9	ሐ	4777	ሐ	12E2	ሐ	4833	ሐ	1321	ሐ	4897	ሐ	1349	ሐ	4937
1212	ሐ	4626	ሐ	123A	ሐ	4666	ሐ	127A	ሐ	4730	ሐ	12AA	ሐ	4778	ሐ	12E3	ሐ	4834	ሐ	1322	ሐ	4898	ሐ	134A	ሐ	4938
1213	ሐ	4627	ሐ	123B	ሐ	4667	ሐ	127B	ሐ	4731	ሐ	12AB	ሐ	4779	ሐ	12E4	ሐ	4835	ሐ	1323	ሐ	4899	ሐ	134B	ሐ	4939
1214	ሐ	4628	ሐ	123C	ሐ	4668	ሐ	127C	ሐ	4732	ሐ	12AC	ሐ	4780	ሐ	12E5	ሐ	4836	ሐ	1324	ሐ	4900	ሐ	134C	ሐ	4940
1215	ሐ	4629	ሐ	123D	ሐ	4669	ሐ	127D	ሐ	4733	ሐ	12AD	ሐ	4781	ሐ	12E6	ሐ	4837	ሐ	1325	ሐ	4901	ሐ	134D	ሐ	4941
1216	ሐ	4630	ሐ	123E	ሐ	4670	ሐ	127E	ሐ	4734	ሐ	12AE	ሐ	4782	ሐ	12E7	ሐ	4838	ሐ	1326	ሐ	4902	ሐ	134E	ሐ	4942
	ሐ	4631	ሐ	123F	ሐ	4671	ሐ	127F	ሐ	4735	ሐ	12B3	ሐ	4787	ሐ	12E8	ሐ	4839	ሐ	1327	ሐ	4903	ሐ	134F	ሐ	4943
1218	ሐ	4632	ሐ	1240	ሐ	4672	ሐ	1280	ሐ	4736	ሐ	12B8	ሐ	4792	ሐ	12E9	ሐ	4840	ሐ	1328	ሐ	4904	ሐ	1350	ሐ	4944
1219	ሐ	4633	ሐ	1241	ሐ	4673	ሐ	1281	ሐ	4737	ሐ	12B9	ሐ	4793	ሐ	12E10	ሐ	4841	ሐ	1329	ሐ	4905	ሐ	1351	ሐ	4945
121A	ሐ	4634	ሐ	1242	ሐ	4674	ሐ	1282	ሐ	4738	ሐ	12BA	ሐ	4794	ሐ	12EA	ሐ	4842	ሐ	132A	ሐ	4906	ሐ	1352	ሐ	4946
121B	ሐ	4635	ሐ	1243	ሐ	4675	ሐ	1283	ሐ	4739	ሐ	12BB	ሐ	4795	ሐ	12EB	ሐ	4843	ሐ	132B	ሐ	4907	ሐ	1353	ሐ	4947
121C	ሐ	4636	ሐ	1244	ሐ	4676	ሐ	1284	ሐ	4740	ሐ	12BC	ሐ	4796	ሐ	12EC	ሐ	4844	ሐ	132C	ሐ	4908	ሐ	1354	ሐ	4948
121D	ሐ	4637	ሐ	1245	ሐ	4677	ሐ	1285	ሐ	4741	ሐ	12BD	ሐ	4797	ሐ	12ED	ሐ	4845	ሐ	132D	ሐ	4909	ሐ	1355	ሐ	4949
121E	ሐ	4638	ሐ	1246	ሐ	4678	ሐ	1286	ሐ	4742	ሐ	12BE	ሐ	4798	ሐ	12EE	ሐ	4846	ሐ	132E	ሐ	4910	ሐ	1356	ሐ	4950
121F	ሐ	4639	ሐ	124B	ሐ	4683	ሐ	128B	ሐ	4747	ሐ	12BF	ሐ	4799	ሐ	12EB	ሐ	4843	ሐ	132F	ሐ	4911	ሐ	1357	ሐ	4951
1220	ሐ	4640	ሐ	1260	ሐ	4704	ሐ	1290	ሐ	4752	ሐ	12C8	ሐ	4808	ሐ	12F0	ሐ	4848	ሐ	1330	ሐ	4912				
1221	ሐ	4641	ሐ	1261	ሐ	4705	ሐ	1291	ሐ	4753	ሐ	12C9	ሐ	4809	ሐ	12F1	ሐ	4849	ሐ	1331	ሐ	4913				
1222	ሐ	4642	ሐ	1262	ሐ	4706	ሐ	1292	ሐ	4754	ሐ	12CA	ሐ	4810	ሐ	12F2	ሐ	4850	ሐ	1332	ሐ	4914				
1223	ሐ	4643	ሐ	1263	ሐ	4707	ሐ	1293	ሐ	4755	ሐ	12CB	ሐ	4811	ሐ	12F3	ሐ	4851	ሐ	1333	ሐ	4915	1361	:	4961	
1224	ሐ	4644	ሐ	1264	ሐ	4708	ሐ	1294	ሐ	4756	ሐ	12CC	ሐ	4812	ሐ	12F4	ሐ	4852	ሐ	1334	ሐ	4916	1362	::	4962	
1225	ሐ	4645	ሐ	1265	ሐ	4709	ሐ	1295	ሐ	4757	ሐ	12CD	ሐ	4813	ሐ	12F5	ሐ	4853	ሐ	1335	ሐ	4917	1364	:	4963	
1226	ሐ	4646	ሐ	1266	ሐ	4710	ሐ	1296	ሐ	4758	ሐ	12CE	ሐ	4814	ሐ	12F6	ሐ	4854	ሐ	1336	ሐ	4918	1365	:	4964	
1227	ሐ	4647	ሐ	1267	ሐ	4711	ሐ	1297	ሐ	4759	ሐ	12CB	ሐ	4811	ሐ	12F7	ሐ	4855	ሐ	1337	ሐ	4919	1367	:	4966	

Table 6: Character frequency table of Amharic Fidel

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

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

Ran	Freq	word	Ran	Freq	word	Ran	Freq	word	Ran	Freq	word
k			k			k			k		
1	4292	ነው	29	533	ከ	57	378	ማለት	85	283	በኢትዮጵያ
2	2827	ለይ	30	530	መንግሥት	58	375	እግዚአብሔር	86	281	መረጃ
3	1703	ወደ	31	528	ዓመት	59	372	ብዙ	87	280	በፊት
4	1448	ውስጥ	32	502	ድረስ	60	367	ብሎ	88	274	ሕዝብ
5	1319	ግን	33	500	አቶ	61	357	ዘመን	89	264	አገሮች
6	1265	ቀን	34	488	በዚህ	62	343	አለ	90	263	አይደለም
7	1138	ነበር	35	478	እኔ	63	341	መካከል	91	261	መሠረት
8	1114	ጋር	36	466	ሥራ	64	331	ሁኔታ	92	261	በኩል
9	1086	ጊዜ	37	463	የኢትዮጵያ	65	322	ልጆች	93	260	ወቅት
10	1000	ሁሉ	38	459	ልጅ	66	321	እንጂ	94	260	በታ
11	9612	እንደ	39	447	ሰዎች	67	320	እንዲህ	95	259	መካታወቂያ
12	9198	ነገር	40	440	ምን	68	315	ምክንያት	96	258	ችግር
13	8703	ወይም	41	437	ኢትዮጵያ	69	314	ሁለት	97	257	የመጸኛ
14	8336	ይህ	42	434	ሆነ	70	314	ብር	98	253	ስዕል

15	8005	ደግ ሞ	43	433 5	ገጽ	71	313 8	ጉዳይ	99	249 2	ታሪክ
16	7733	ቤት	44	428 3	ስም	72	310 2	አዲስ	100	248 7	ዓለም
17	7596	ቁጥር	45	428 0	ም	73	309 5	ደረጃ	101	247 3	ምንም
18	7402	በ	46	424 7	ክፍል	74	307 9	ሆኖ	102	245 6	ሌላ
19	7027	እና	47	422 8	ከተማ	75	304 4	ዋና	103	245 5	ይህን
20	6911	አንድ	48	422 2	ያለው	76	303 1	እኔ	104	242 4	ሳይሆን
21	6519	በኋላ	49	419 9	መንገድ	77	299 0	የሥራ	105	242 2	መዋቅር
22	6481	ሰው	50	413 5	በላይ	78	296 8	ንጉሥ	106	241 7	አለው
23	6306	ዘንድ	51	402 8	ዓ	79	296 8	ጀምሮ	107	239 9	ቃል
24	6241	ዓም	52	390 6	ከዚህ	80	295 8	ከፍተኛ	108	239 5	ያለ
25	6142	ናቸው	53	383 4	አገር	81	290 1	የ	109	234 3	አሉ
26	5911	ብቻ	54	383 1	ስለ	82	289 8	አሁን	110	234 3	እንዲሁም
27	5907	እስከ	55	382 3	መሆኑን	83	287 6	ቅፅ	111	231 5	ሰዓት
28	5579	ሲሆን	56	381 1	አበባ	84	287 5	ዓመታት	112	230 4	ዋጋ

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