SINET: Ethiop. J. Sci., 46(2): 163-175, 2023, 2023 ISSN: 0379–2897 (PRINT)

eISSN: 2520–7997

Date received: January 10, 2023; Date revised: October 30, 2023; Date accepted: November 12, 2023
DOI: https://dx.doi.org/10.4314/sinet.v46i2.5

Identifying Amharic-Tigrigna Shared Features: Towards Optimizing Implementation of Under Resourced Languages

Lemlem Hagos*, Million Meshesha, Solomon Atnafu and Solomon Teferra

Addis Ababa University, Addis Ababa. Ethiopia. E-mail: Lemlem.hagos@aau.edu.et

ABSTRACT: In this article, exploratory research is conducted to analyze statistical overlap across Amharic and Tigrigna at different level of abstraction, namely, word level, CV syllable level, and at phoneme level. Amharic and Tigrigna are among the most widely spoken Ethiosemitic languages in Ethiopia, yet under resourced to be fully integrated into TTS applications that assist oral society in their day-to-day activities. Text to speech research requires linguistic resources involving intensive text analysis and acoustic resources that involve digital signal analysis. TTS researches for Ethiosemitic languages have been explored on monolingual basis which require fragmented research activities towards the resource intensive task. Investigating the level of overlap for Amharic and Tigrigna gives an insight to reuse shared acoustic and linguistic resources across these languages and reduce duplication of effort in the process of designing higher level applications such as TTS. According to our statistical analysis, Amharic and Tigrigna share 86.36% at phonemic level, 85.93% at CV syllable level, and encouraging level of overlap at the word level. The extent to which these languages overlap at different level of abstraction implies the opportunity to reduce duplication of effort in the design and development of bilingual and multilingual TTS for Ethiosemitic polyglots.

Keywords/phrases: Amharic-Tigrigna, Shared phonemes, Shared Syllables, Shared words

INTRODUCTION

Ethiopian Semitic languages (ESL) is a sub family of South-Semitic language which in turn is a sub family of West Semitic language under the Semitic language of the Afro-Asiatic super family (Bender & Fulas, 1978). It includes Geez, Tigrigna, Tigre, Amharic and Argoba (Bender & Fulas, 1978). Amharic and Tigrigna are the second and the third most spoken Semitic languages in the world, next to Arabic. While Amharic is spoken by more than 30 million native speakers, Tigrigna has more than 10 million native speakers (Ethnologue, 2022).

Sharing in languages span from the elementary building blocks such as sound system all through syntactic and semantic structure of the symbols as well as the intentions specified in the writings or the discourse to pragmatics involving the implication and interpretation of the message be it between the lines or beyond the lines. The need to capture the shared features among languages is to facilitate a cost-effective design of bilingual and multilingual systems for the languages. Contexts shared among languages at a higher level contribute a lot towards successful collaboration among polyglot societies (Thomas J, 2013). The reason for languages to exhibit shared features is mainly attributed to their common ancestors as well as frequent interaction of people across regions. It is expected that languages of similar family do share certain features (Sengupta & Saha, 2015). Amharic and Tigrigna originate from the same Ethiosemitic languages, grouped under the two big categories of the Ethiosemitic super family, namely South and North respectively (Tekabe Legesse, 2021; Bulakh, 2019; Edzard, 2019). Amharic and Tigrigna being part of the same Ethiosemitic language family (Bender & Fulas, 1978), there is a need to investigate their relatedness, namely at phoneme, syllable, and word level, so that resource intensive researches such as text to speech synthesis would benefit from reusability of shared features. As a matter of fact, nowadays Text to Speech researches on Ethiosemitic languages focus on monolingual basis, with emphasis on Amharic or Tigrigna as a base language individually (Lemlem Hagos & Million Meshesha, 2015).

With the advancement in speech synthesis technology, researchers are aiming to achieve more natural sounding and intelligible speech output. Text to speech synthesis enables computers to convert arbitrary text into audible speech (Taylor, 2009). Text to speech synthesis undergoes the process of text analysis and speech generation (Dutoit, 1997). The text analysis is responsible for determining the underlying structure of the sentence and the phonemic composition of each word. This is because strings of phonemes form larger units such as syllables; which in turn form words, constituting phrases and sentences. These structures need to be indicated in the underlying representations for an utterance, because aspects of how a sentence is pronounced depends on the locations of these types of boundaries showing pronunciation of each word, syntactic structure for the sentence and semantic focus to resolve ambiguity (Taylor, 2009).Speech generation part of text to speech synthesizer transforms the abstract linguistic representation into speech waveform. It is responsible for phonetic realization of each phoneme (Taylor, 2009).The speech synthesis part is also concerned with the selection and concatenation of appropriate speech units given the phoneme string as well as a speech waveform (Dutoit, 1997). In this article, we focus on text analysis as a base for designing a bilingual TTS.

Accordingly, the purpose of this article is to identify and statistically analyze shared features of Amharic and Tigrigna, which will serve as a base reusable resource in the process of shifting from monolingual TTS to bilingual and even multilingual TTS. The shared resources also enhance transfer learning among the languages even at the monolingual level. Such an effort aims at reducing unnecessary duplication of effort in linguistic as well acoustic resource preparation which are expensive, yet mandatory aspects of TTS research. Investigating shared resources helps in the optimal design of statistical models in artificial intelligence which are resource intensive. Thus investigating shared features between two languages also contributes towards design of economic statistical models.

One of the interesting aspects of these languages is that they are historically, culturally, socially, economically, and politically interrelated. Accordingly, the two languages are polyglot, meaning an individual who speaks one of the languages often understands the message of the other if not speak it fluently.

The rest of this article is organized as follows. First, we present related works, focusing on reusability of shared features of languages for TTS. Second, design architecture is depicted along with algorithm and description of respective components. Third, statistical analysis of shared features of Amharic and Tigrigna, at the phoneme, CV syllable and word level, is presented. Forth, discussion of implication of the shared features analyzed in previous section is presented with emphasis towards its contribution in the process of shifting Ethiopic TTS research from monolingual to bilingual and multilingual. Finally, concluding remark is presented, wrapping up the essence of the research endeavor in this article.

Related Works

In this section, we review literature that focus on utilization of shared features of related languages for the purpose of optimizing linguistic and acoustic resources. Our review spans from low resourced languages that share phonemes such as Catalan-Spanish to transfer learning employed across resourced languages and under resourced languages such as English-Mandarin. The purpose of the review is to show the relevance of focusing on shared aspects of languages so that linguistic and acoustic resources necessary in the development of machine enabled tasks such as TTS are achievable with optimal cost.

Catalan-Spanish (Esquerra, Bonafonte, & Vallv, 1997) as well as Urdu-Sindhi (Shah, Ansari, & Das, 2004) demonstrated the phonemic overlap across respective pair of languages leading to reduction in the required speech dataset to design bilingual TTS. The pair of languages under consideration also show the role of using shared features of related languages in building a cost-effective bilingual speech synthesis especially for low resourced languages.

Researches also illustrate the possibility of integrating English and Japanese which are resourced language to economically model low resourced languages such as Mongolian (Byambadorj, Nishimura, Ayush, Ohta, & Kitaoka, 2021), and Mandarin (Zhao, Nguyen, Wang, & Ma, 2020) to exploit shared elements at the phoneme level, irrespective of the fact that these pair of languages are genetically unrelated. The aforementioned effort is towards designing a model for multilingual text to speech synthesis for low resourced languages.

The review explored so far revealed that most of the researches concentrated on European and Asian languages. These researches proved the existence of shared phonemes across polyglot languages. However, there is limited works done for African languages in general and Ethiopian languages in particular. As a matter of fact, in Ethiopia there are Semitic, Cushitic, and Omotic language families. In this study, we focus on Ethiopian Semitic languages with specific emphasis on two of the most widely used languages, Amharic and Tigrigna from South and North EthioSemitic languages. It is imperative that there is a need to explore the level of overlap of these languages so that shared features are reusable across the languages. This is important especially for under resourced languages where producing linguistic and acoustic resources required for TTS is both expensive and time-consuming.

Design

In this section, we propose a design of Amharic-Tigrigna feature overlap analysis that enables statistical investigation of shared features of Amharic and Tigrigna text to promote reusability of shareable aspects of the languages in the process of designing TTS for Ethiopian Semitic languages. The proposed blueprint is used for exploring shared features of bilingual languages at the phoneme, syllable and word level, where it takes free text, and generates statistics of overlap of the text at the selected measure of units. The steps involved in the design are preprocessing text, segmenting text, and overlap analysis (see Figure 1). The preprocessing step takes care of data cleaning and normalization. The segmenter splits normalized text into words, syllables and phonemes as per the requirement. The overlap analyzer computes the shared words, syllables and phonemes across Amharic and Tigrigna text.

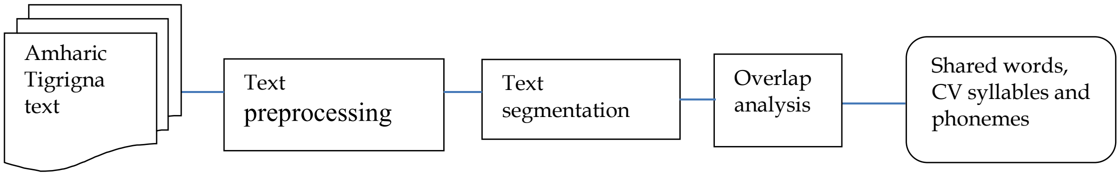

Figure 1 Proposed design for determining shared units

Preprocessing

The input Amharic and Tigrigna text taken from newspapers and novels consists of nonstandard words (NSW), and punctuation marks. NSWs are words that are not found in a dictionary and their representation leads to more than one way of readings. NSWs include numerals, abbreviations and acronyms. For the purpose of text analysis, we filter out NSWs in the preprocessing phase. As a result, free text is converted into normalized text (as shown in Figure 2).

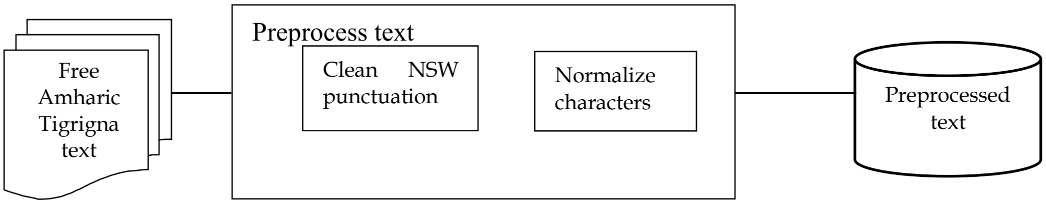

Figure 2. Block diagram for preprocessing text.

Listing 1 shows a generic algorithm to clean input Amharic and Tigrigna text from numerals and punctuation marks. Predefined set of numerals and punctuation marks are used to check if the text contains them. In addition, for text normalization algorithm shown in Listing 2, redundant characters in both languages, a dictionary that stores a pair of redundant and core characters is used as a reference.

```
Algorithm: Preprocessing (clean numeral, punctuation and other symbols)
    Input: text _Amharic,  text_ Tigrigna//unprocessed text file;
    Output: text_ Amharic, text_ Tigrigna//preprocessed text file
    Character Variant_ Amharic = set of duplicate Amharic_ characters
    Character Variant_ Tigrigna = set of duplicate Amharic_ characters
    numerals = set of numerals
    punct = set of punctuation and other symbols
    clean Text_ Amhric = {}
    clean Text_ Tigrigna = {}

    while not eof(text_Amharic) do
      for each item in text_Amharic
         if item in numerals or punct then
            replace item with space
        append item to cleanText_Amharic

    while not eof(text_Tigrigna) do
      for each item in text_Tigrigna
        if item in numerals or punct then
          replace item with space
       append item to cleanText_Tigrigna
end of algorithm
```

Listing 1 Algorithm: text cleaning

```
Algorithm: Preprocessing (normalize text)
    Input: cleanText_Amharic, cleanText_Tigrigna//from file
    Output: normalizedText_Amharic, normalizedText_Tigrigna

    characterVariant_Amahric = set of redundant Amharic characters
    characterVariant_Tigrigna = set of redundant Tigrigna characters
    variantCommonPair_Amharic ={}
    variantCommonPair_Tigrigna ={}

    normalizedText_Amharic = {}
    normalizedText_Tigrigna ={}

    while not eof(cleanText_Amharic) do
        for item in cleanText_Amharic //cleaned of numbers and punctuations
          if item exists in characterVariant_Amharic then
            replace item with variantCommonPair_Amharic(value)
          append item to normalizedText_Amharic
    while not eof(cleanText_Tigrigna) do
      for item in cleanText_Tigrigna
          if item exists in characterVariant_Tigrigna then
            replace item with variantCommonPair_Tigrigna
          append item to normalizedText_Tigirigna
end of algorithm
```

Listing 2 Algorithm: Text normalization

Segmentation

Segmentation is the process of splitting text into its constituent parts, such as word, characters, and phonemes. Figure 3 depicts block diagram for segmenting text, where normalized text is chopped down into words, syllables and phonemes, and stored as bag of words (BOW), bag of syllables (BOS), and bag of phonemes (BOP).

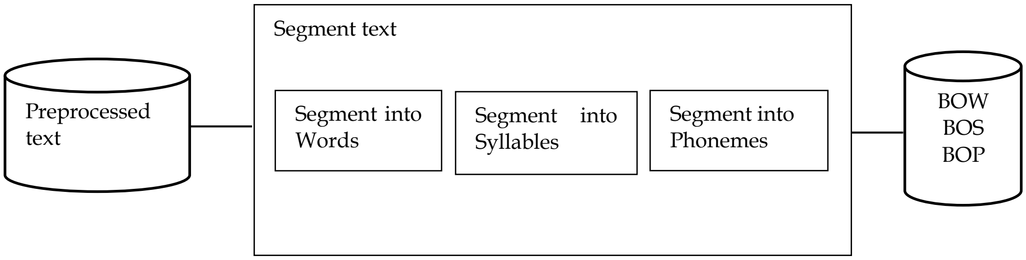

Figure 3 Block diagram for segmentation

Listing 3 shows the algorithm to step by step segment Amharic and Tigrigna text into the desired measure of units, namely words, syllables and phonemes. As per the design in Figure 3, the algorithm accepts normalized Amharic and Tigrigna text and generates bag of words, syllables and phonemes for each language. The process involves iterative segmentation. First, it segments text into words and stores the result as BOW for each language. Then BOW of each language is further segmented into BOS, which in turn is further segmented into BOP. Segmentation of words into CV syllables can be done without dealing with grapheme to phoneme conversion as each character in both Amharic and Tigrigna is a CV syllable by its own nature. When we need to segment the syllables into phonemes, there is a need for grapheme to phoneme conversion which we implemented using a look-up table.

```
Algorithm: Segmentation
    Input: normalizedText_Amharic, normalizedText_Tigrigna
    Output: BOW, BOS, BOP for each language
    g2p_dict_Amharic =set of grapheme-phoneme pair_Amharic
    g2p_dict_Tigrigna =set of grapheme-phoneme pair_Tigrigna
    BOW_Amharic = {}
    BOS_Amharic = {}
    BOP_Amharic = {}
    BOW_Tigrigna = {}
    BOS_Tigrigna = {}
    BOP_Tigrigna = {}
    while not eof (normalizedText_Amharic) do
        BOW_Amahric = split into words (normalizedText_Amharic)
         for each item in BOW_Amharic
                If item matches g2p_dict_Amharic(key) then
                    replace item with g2p_dict_Amharic (value)
                append item to BOS_Amharic
         BOP_Amharic = split into phonemes (BOS_Amharic)

    //follow the same logic for Tigrigna
    while not eof (normalizedText_Tigrigna)do
        BOW_Tigrigna = split into words (normalizedText_Tigrigna)
         for each item in BOW_Tigrigna
              If item matches g2p_dict_Tigrigna(key) then
                   replace item with g2p_dict_Tigrigna (value)
             append item to BOS_Tigrigna
        BOP_Tigrigna = split into phonemes (BOS_Tigrigna)
end of algorithm
```

Listing 3 Algorithm: Segmentation

Compute shared units

This paper aims to present the extent to which Amharic and Tigrigna share linguistic units. To investigate the overlap statistically, an attempt is made to identify the word, syllables, and phonemes in common across the two languages, as shown in Figure 4.

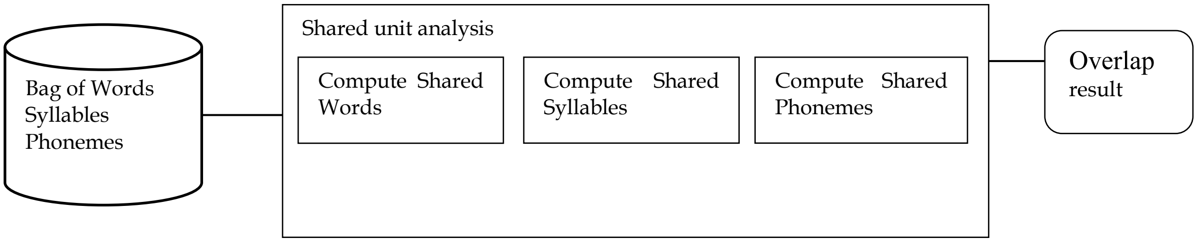

Figure 4 Block diagram for shared unit analysis.

Listing 4 shows algorithm to compute level of overlap in words, syllables and phonemes across Amharic and Tigrigna text.

```
Algorithm: ComputeOverlapPercentage ()
Input: Amharic and Tigrigna BOW, BOS, BOP (in general BOW)
Output: percentage of overlap
word_shared = { } // container for shared words, syllables, phoneme
BOW_Amharic_unique = { } // container for unique Amharic words/syllables/phoneme
BOW_Tigrigna_unique = { } // container for unique Amharic words/syllables/phoneme
count_shared = 0 // container for number of shared words, syllable, phoneme
count_BOW_Amharic = 0 //container for number of Amharic words, syllable, phoneme
count_BOW_Tigrigna = 0 //container for number of Tigrigna words, syllable, phoneme
while not eof(BOW_Amharic)
  for word in BOW_Amharic
    if word not in BOW_Amharic_unique
      then append word to BOW_Amharic_Unique
      count_BOW_Amharic+=1
while not eof(BOW_Tigrigna)
  for word in BOW_Tigrigna
    if word not in BOW_Tigrigna_Unique
      then append word to BOW_Tigrigna_Unique
      count_BOW_Tigrigna+=1
while not eof (BOW_Amharic_Unique)
  While not eof(BOW_Tigrigna_Unique)
    if word in BOW_Amharic_Unique is same as word in BOW_Tigrigna_Unique
      if word not in word_shared
        append word to word_shared
      count_shared+=1

$shared\_percentage = \frac{count\_shared * 100}{count\_BOW\_Amharic + count\_BOW\_Tigrigna - count\_shared}$
end of algorithm
```

Listing 4 Algorithm: Compute Overlap Percentage

The algorithm in Listing 4 shows how to compute shared words, syllables and phonemes, considering bag of words (BOW) as a generic input. The same procedure is followed for analyzing syllable and phoneme level overlap for the languages under study, extending the input to bag of syllables (BOS) and bag of phonemes (BOP). In the process, the algorithm identifies unique words, syllables and phonemes in the respective input BOW, BOS, and BOP for each language. It then identifies iteratively the shared words, syllables and phonemes across the pair of input data and finally computes the percentage of shared elements.

Statistical Analysis of Shared Features and Discussion of Result

Amharic and Tigrigna share linguistic and acoustic features at different level of abstraction. Here we investigate these overlaps at phoneme, syllable and word level. As per the aim of the paper, an attempt is made to identify shared features of Amharic and Tigrigna at word, syllable and phoneme levels. Understanding shared features of Ethiopian languages in general and Amharic and Tigrigna in particular helps to develop multilingual and polyglot applications such as TTS.

Phoneme level overlap

Even though there are 35 consonantal segments in the integrated character set of Amharic and Tigrigna, the existence of phonetically redundant consonants in both languages reduce the number of unique sound characters. Accordingly, Amharic is composed of 211 characters with unique sound, that is 196 (28 by 7) core characters plus 15 (3 by 5) labialized characters. Similarly, Tigrigna consists of 249 uniquely pronounceable characters, that is 224 (32 by 7) core characters along with 25 (5 by 5) labialized characters. Accordingly, a total of seven of the 35 consonantal segments, (ሐ, ሠ, ቐ, ኀ, ኸ, ጸ and ዐ) are subtracted, due to redundancy in sound or absence in the character set of Amharic. The consonantal phonemes ሐ, ኀ and ኸ are redundant with the consonantal phoneme ሀ, so are the consonantal phonemes ሠ, ጸ and ዐ with ሰ, ፀ and አ, respectively. In addition, the consonantal segment ቐ is nonexistent in the character set of Amharic.

Similarly, Tigrigna has got phonetically redundant consonantal phonemes. Hence, the consonantal phoneme ኀ is phonetically redundant with the consonantal phoneme ሀ, so are the consonantal phonemes ሠ and ጸ with ሰ and ፀ, respectively. Thus, out of the 35 consonantal segments, there are 32 unique sound core consonantal segments providing 32 by 7 characters, as well as 5 by 5 labialized characters in Tigrigna.

Both Amharic and Tigrigna are known to be phonemic languages (Baye Yimam, 2007) (Daniel Teklu, 2008), where the phonemes are either consonants or vowels. Even though there are seven vowels namely, ə, u, i, a, e, ɨ, o, in both Amharic and Tigrigna languages, there is controversial number of consonantal phonemes in these languages.

According to (Girmay Berhane, 1983), Tigrigna has 29 consonantal phonemes and seven vowels. The plosive labiovelars, ጐ[gw], ኰ[kw], ቈ[qw], as well as the fricative labiovelars, ዀ[xw] and ቘ[ɣw] are derivable from their respective core consonantal segments. Furthermore, Girmay notes that the fricative velars, ኸ[x] and ቐ [ɣ], are allophones of ከ [k] and ቀ [q] respectively. Thus, these are not included in the consonantal chart of Tigrigna (Girmay Berhane, 1983). According to (Tsehaye Tefera, 1979) and (Daniel Teklu, 2008), however, aforementioned derivable and allophone segments as well as the phoneme ቨ [V] are included in the consonant chart of Tigrigna. As a result, the number of consonants in Tigrigna would be 37, which is composed of 32 core consonants and 5 labialized composite segments.

Similarly, Amharic contains debatable number of consonants. (Baye Yimam, 2007), argues that

Amharic has 30 consonants, including the core and labialized consonants (Mulugeta Syoum, 2001), however, reduces the number of core consonants to 21 and 6 derivable palatal consonants. Mulugeta also argues the possibility of recovering the voiceless glottal ኧ [ʔ] as a consonant. As indicated in (Baye Yimam, 2007), the three labialized velars, ኰ[kʷ], ጐ[gʷ] and ቈ[qʷ], are considered as part of the 30 consonants in the consonant chart of Amharic. Adding Mulugeta's recoverable glottal ኧ [ʔ] phoneme, the number of Amharic consonants increases to 31, out of which 28 are core consonants, and 3 are labialized ones.

Phonemic analysis of Amharic and Tigrigna shows 31 out of 37 consonantal phonemes are shared between the languages. In other words, the entire set of consonantal phonemes in Amharic is contained in that of Tigrigna. Considering the additional seven vowel phonemes, common across these languages, total phonemic overlap between Amharic and Tigrigna becomes 86.36%. This will create a great advantage to go for designing a cost-effective bilingual TTS for the two languages.

Character (CV Syllable) level overlap

Total character set of Amharic and Tigrigna is composed of 245 core characters which is a result of the 35 core consonants by seven vowels matrix. Here, it is noted that the redundant characters are not removed. In addition to the core (CV based) characters, there are 25 labialized characters which are composed of 5 labialized velars (ኰ[kʷ], ጐ[gʷ], ቈ[qʷ], ዀ[xʷ], ቘ [ɣʷ]) integrated with 5 labialized vowels (wə, wi, wa, we, wɨ). There are also few characters locally called incomplete (ጎደሎ ፊደላት) characters because they are composed of a consonant together with a short 'wa' sound. Thus, there are more than 270 characters in both Amharic and Tigrigna because of the availability of the incomplete characters (ጎደሎ ፊደላት), which are partially considered in this analysis.

Table 1 summarizes character level overlap between Amharic and Tigrigna writing system. In the analysis, we consider core characters and labialized characters which exhibit tremendous CV overlap across the languages.

Table 1. Character (CV Syllable) level overlap in Amharic and Tigrigna.

	Amharic	Tigrigna	Shared row	Shared %
Core characters	217	245	217	88.57%
Labialized characters	15	25	15	60%
Total	232	270	232	85.93%

Amharic and Tigrigna share a total of 232 characters which is composed of 217 (31 by 7) core (CV) characters plus 15 (3 by 5) labialized characters. There is a total of 270 characters in both Amharic and Tigrigna, which consist of 245 (35 by 7) core characters plus 25 (5 by 5) labialized characters. Thus, the character level overlap between Amharic and Tigrigna is therefore 85.93%, which again justifies the feasibility of developing a bilingual TTS for Amharic and Tigrigna.

The character level overlap shown in Table 1(85.93%) refers to the CV and CWv[1] grapheme level overlap. This overlap serves as a basis for text production in both Amharic and Tigrigna. In this analysis, the incomplete characters are partially included as they are considered to be derivable. The grapheme level of overlap is slightly different from the phonemic level overlap (86.36%). This is because consonantal phonemes of same sound that exist in the character set of these languages are presented in different level of distribution, yet duplicate sounds are normalized in the consonantal phoneme consideration. Referring to the character set (የፊደል ገበታ) of each language there

[1] CW$_v$ refers to consonant-short w-vowel combination (^{w}ə, ^{w}i, ^{w}a, ^{w}e, ^{w} i).

are 35 and 21 same sound characters in Amharic[2] and Tigrigna[3]respectively. Another reason why the phonemic overlap is slightly different from that of character level overlap is that in the phonemic analysis, the velar labialized consonants are considered to have the same weight as the core consonants while in the character level analysis the velar labialized consonants are incomplete as they generate five variants rather than seven unlike the other core consonants.

The results of phonemic overlap and character level overlap across the two languages are not far from each other. This is because of the counter balance between consideration velar labialized consonants in consonantal charts irrespective of their incompleteness and inclusion of redundant sound characters in the character set of both Amharic and Tigrigna.

Word Level Overlap

In addition to phoneme level and character level overlap, Amharic and Tigrigna exhibit word level overlap, where the shared words could reflect one of the following;

i) Same spelling, same pronunciation and same semantics (see Table 2).

ii) Same spelling, slightly different pronunciation, and same semantics (see Table 3).

iii) Same semantics different in one or two phonemes. (See Table 4).

Words used for word level analysis are taken from Tigrigna-English dictionary (Efrem Zecarias, 2007). Sample of shared words across Amharic and Tigrigna that are spelt and pronounced the same way in both languages besides convoying the same semantics are shown in Table 2.

Table 2. Amharic and Tigrigna words with the same spelling, same pronunciation and same meaning.

Same meaning word in Amharic & Tigrigna	Pronunciation (IPA) Tigrigna/Amharic	Tigrigna/ Amharic semantics
ሃብታም	Habtam	Wealthy
ሃይማኖት	Haymanot	Religion
ህንጻ	hɨns’a	Building
ለገሰ	ləggəsə	Grant
ልምምድ	lɨmɨmɨd	practice
ላም	Lam	Cow
ለጠፈ	lət’t’əfə	Glue
መስመር	məsmər	Line
መስከረም	məskərəm	September
መስጊድ	məsgid	mosque
መኪና	məkina	Car
ረብሻ	rəβʃa	agitate
ሬሳ	resa	Dead body
ሩዝ	ruz	Rice
ሰላም	səlam	peace
ሰማይ	səmaj	sky
ሰራዊት	sərawit	army
ሳሙና	samuna	Soap
ሸመተ	ʃəmmətə	purchase
ሸጠ	ʃət’ə	Sell
ሸፈነ	ʃəffənə	cover

In addition to the shared words with the same spelling, pronunciation, and semantics (Table 2), there are words pronounced different in either Amharic or Tigrigna, yet share the same semantics as depicted in Table 3, and Table 4. Here, we present sample words that look the same in the grapheme and convey the same meaning in both Amharic and Tigrigna. The only difference we observed is the way the words are pronounced. Most of the pronunciation difference is reflected with geminates in Amharic and insertion of epenthesis vowel in Tigrigna. This may affect the naturalness of TTS but not intelligibility of TTS.

[2] ሀኸሐኀ→ሀ፡ሰሠ→ሰ፡ፀጸ→ፀ Each consonant assume seven orders to generate the character 4*7+2*7+2*7-(3*7)=35 extra same sound characters. This didn’t consider the Aa issue

[3] ሀኀ→ሀ፡ሰሠ→ሰ፡ፀጸ→ፀ 2*7+2*7+2*7-(3*7)=21 extra same sound characters

Table 3. Amharic and Tigrigna word with same spelling, different pronunciation and same gloss.

Amharic/ Tigrigna word	Tigrigna Pronunciation IPA	Amharic Pronunciation IPA	Semantics
ምርጫ	mɨrrɨca	mɨrca	election
ረገመ	rəgəmə	rəggemə	curse
ሸተተ	ʃətətə	ʃəttətə	smell
ቀረጸ	qərəs’ə	qərrəs’ə	Shape/engrave
ቀደመ	qədəme	qəddəmə	Exceed
በረረ	bərərə	bərrərə	Fly
ተመለሰ	təmələsə	təməlləsə	return
ተጋደመ	təgadəmə	təgaddəmə	Lie
ነደደ	nədədə	nəddədə	Burn
ነገረ	nəgərə	nəggərə	tell/inform
ነጠረ	nət’ərə	nət’t’ərə	jump/leap
ነፈሰ	nəfəsə	nəffəsə	Blow
ኣመነ	ʔamənə	ʔammənə	Believe
ከሰሰ	kəsəsə	kəssəsə	Accuse
ከፈለ	kəfələ	kəffələ	Pay
ገለጸ	gələs’ə	gəlləs’ə	express
ጸጥታ	s’ət’ta	s’ət’t’ɨta	silence

Apart from geminates and insertion of epenthesis vowel, pronunciation difference between Amharic and Tigrigna is observed in relation with the bilabial consonant /b/ which is realized as plosive stop in Amharic and fricative in Tigrigna.

Table 4. Amharic and Tigrigna words with same spelling, same gloss and different pronunciation.

Amharic/ Tigrigna Word	Tigrigna Pronunciation IPA	Amharic Pronunciation IPA	Semantics
ወደብ	wədəβ	wədəb	port
ዘነበ	zənəβə	zənnəbə	rain
ደረበ	dərrəβə	dərrəbə	double
ደቡብ	dəβuβ	dəbub	south
ጥበብ	t’ɨβəβ	t’ɨbəb	art

Table 4 shows sample shared words that are pronounced differently because of the bilabial consonant /b/, yet do not bring difference in the meaning of the word.

Amharic and Tigrigna also share part of a word which carries shared semantics where there is a difference in one or two phonemes in between. Table 5 below depicts sample words that reflect the same meaning but spelt slightly different. Such sharing can be a basis for translation among local languages. Shared semantics as in Table 5, indicates that there is a root word level overlap across Amharic and Tigrigna that can contribute towards optimization of resources through reuse of common features.

Table 5. Shared semantics (seemingly at root word level).

Tigrigna spelling/IPA	Amharic spelling/IPA	Semantics
ቡን/bun	ቡና /buna	Coffee
ዕርቂ /ʕɪrqi	እርቅ/ʔɪrq	Reconciliation
በርበረ/bərbərə	በርበሬ /bərbəre	red pepper
ቀመም /qəməm	ቅመም/qɪməm	Spice
ጥቅምቲ/ t’ɪqɪmti	ጥቅምት t’ɪqɪmt	October
ነብሪ /nəbɪri	ነብር/nəbɪr	Tiger
እግሪ/ʔɪgɪri	እግር/ʔɪgɪ	Foot

Most of the common words shared between Amharic and Tigrigna are pronounced the same way; there are a few geminates in Amharic though. In addition, there is variation in pronouncing the bilabial consonant /b/across Amharic and Tigrigna, where most of the time the plosive/b/ becomes fricative /β/ in Tigrigna.

DISCUSSION OF RESULT AND IMPLICATIONS

Overlap analysis is done in a layer of contexts. First, the phoneme set as well as the character set of Amharic and Tigrigna languages are explored as a result of which there is 86.36% overlap at the phoneme level and 85.93% overlap at character level. Secondly, we explored the existence of overlap in Amharic and Tigrigna at the word level where the shared words exhibit similarity in spelling, pronunciation and meaning. Shared words that are spelled and meaning the same but with slight difference in pronunciation are also explored. The words of similar spelling and pronunciation along with those shared words that differ slightly in pronunciation across the languages will have a great contribution in building cost-effective bilingual TTS applications

As shown in (Esquerra, Bonafonte, & Vallv, 1997); (Shah, Ansari, & Das, 2004) shared phonemes are explored in building bilingual Catalan-Spanish and Urdu-Sindhi. Our research thus extends the exploration of shared features of related languages from the level of phonemes to syllable and word level overlaps.

The benefit of exploring shared features between Amharic and Tigrigna is multifold. In addition to alleviating the problem of tedious tasks involved in text analysis for speech synthesis of Ethiopian languages, the existence of shared features across Amharic and Tigrigna contributes towards creating a paradigm shift in the design of TTS for Ethiopian languages from monolingual to bilingual and multilingual harnessing the shared features of related local languages. Even at the monolingual level, it benefits transfer learning models where a model trained with one languages can serve for the other language with minimal modification to capture the differences.

There is a natural correlation among phonemes, CV syllables and words, where words are contain CV syllables, which in turn are composed of phonemes. The finding of this paper shows Amharic and Tigrigna exhibit overlaps at the level of phonemes, CV syllables and words. Phonemes being the basis for both CV syllables and words, they can be used to design a cost-effective bilingual TTS with acceptable level of intelligibility for polyglot speakers.

Furthermore, Statistical models are resource intensive. Investigating shared resources can lead to optimal resource utilization as the shared featured are reusable. Thus, this research contributes towards designing cost-effective statistical models in general.

CONCLUDING REMARKS

In this article, an attempt is made to examine the extent to which Amharic and Tigrigna languages share linguistic and acoustic features at phoneme level, CV syllable level and word level. Even though Amharic and Tigrigna languages are classified as South and North under the EthioSemitic language family, our investigation shows that there is significant overlap between Amharic and Tigrigna. Harnessing the shared features can save both duplication of effort, and associated cost for developing TTS systems for polyglot speakers of Amharic and Tigrigna which are under resourced languages.

As a result of investigating Ethiopian Semitic languages, particularly Amharic and Tigrigna, share common features at different levels of abstraction: phoneme level, syllable level, and word level, the objective of the solution is to create a blue print that makes use of shared features of Ethiopian Semitic languages.

The emphasis of the article is on figuring out the shared features of the languages. It is observed that all phonemes as well as characters in Amharic are included in Tigrigna. However, there is language specific usage variation such as geminates and epenthesis vowel. It can be generalized that at the phoneme and character level, Tigrigna is more inclusive.

The word level overlap learned from English-Tigrigna dictionary serves as an indication of the practice in the languages to use shared words in their text. The data we analyzed is not sufficient to generalize statistically on the word level overlap. To apply the word level overlap algorithm designed on huge bilingual corpora is part of our future work. Furthermore, such investigation can be extended to other related Ethiopian languages.

Thus, intelligible TTS can be designed for the two languages which is part of our future work.

REFERENCE

1. Baye Yimam. (2007). *Amharic Grammar.* Addis Ababa: Elleni Press.
2. Bender, M. L., & Fulas, H. (1978). *Amharic Verb Morphology: A Generative Approach.* Michiga: Michiga State University.
3. Bulakh, M. (2019). Tigrinya. In *The Semitic Languages* (pp. 174-202). New York, USA: Routledge.
4. Byambadorj, Z., Nishimura, R., Ayush, A., Ohta, K., & Kitaoka, N. (2021). Text to Speech Synthesis for low resource languages using cross-lingual transfer learning and data agumentation. *EURASIP Journal on Audio, Speech, and Music Processing, 42*(1), 1-20.
5. Daniel Teklu. (2008). *Modern Tigrigna Grammar.* Addis Ababa: Biranna Press.
6. Dutoit, T. (1997). *Introduction to Text to Speech Synthesis.* Mons, Belgium: Kluwer Academic Publishers.
7. Edzard, L. (2019). Amharic. In *The Semitic Languages* (pp. 202-206). New York, USA: Routledge.
8. Efrem Zecarias. (2007, 12 18). *Tigrigna-English English-Tigrigna Dictionary.* Retrieved 08 12, 2018, from http://www.memhr.org/dic/ebook/tigeng dictionary.pdf
9. Esquerra, I., Bonafonte, A., & Vallv, F. (1997). *A Bilingual Spanish-Catalan Database of Units for Concatenative Synthesis.* Barcelona, Spain: Universitat Politècnica de Catalunya.
10. Ethnologue. (2022). Retrieved June 20, 2022, from https://www.ethnologue.com
11. Girmay Berhane. (1983). *The Phonology of Tigrigna: Generative Approach.* Addis Ababa: Addis Ababa University.
12. Lemlem Hagos, & Million Meshesha. (2015). Text To Speech Synthesis for Ethiopian Semitic Languages: Issues and the Way Forward. *12th IEEE Africon International Conference.* Addis Ababa, Ethiopia.
13. Mulugeta Syoum. (2001). *The Syllable Structure and Syllabification in Amharic.* Norwegian University of Science and Technology.
14. Sengupta, D., & Saha, G. (2015). Study on Similarity among Indian Languages Using Language Verification Framework. *Advances in Artificial Intelligence*, 1-24.
15. Shah, A. A., Ansari, A. W., & Das, L. (2004). *Bi-Lingual Text to Speech Synthesis System for Urdu and Sindhi.* Jamshoro, Pakistan: Institute of IT, University of Sindh.
16. Taylor, P. (2009). *Text to Speech Synthesis.* New York: Cambridge University Press.
17. Tekabe Legesse. (2021). Ethiosemitic languages: Classification and classification determinants. *Ampersand, 8*, 1-15.
18. Thomas J, M. D. (2013). Shared language: Towards more effective communication. *AMJ, 6*(1), 45-54.
19. Tsehaye Tefera. (1979). *Reference Grammar of Tigrigna.* Washignton DC: Georgetown University.

20. Zhao, S., Nguyen, T. H., Wang, H., & Ma, B. (2020). *Towards Natural Bilingual and Code-Switched Speech Synthesis Based on Mix of Monolingual Recordings and Cross-Lingual Voice Conversion.* arXiv:2010.08136v1 [cs. SD].