Building and Evaluating Somali Language Corpora

Nimaan Abdillahi
Institut des Sciences et des Nouvelles Technologies
Centre d'Etudes et de Recherche de Djibouti
B.P 486 Djibouti
Nimaan.abdillahi@gmail.com

Abstract

In this paper we outline our work to build Somali language Corpora. A read-speech corpus named ***Asaas*** and containing about 10 hours and 26 minutes of good quality signal fully transcribed and well corrected with a well-balanced phonetic distribution is presented. Secondly we outline a Web-based Somali textual corpus named ***Wargeys*** and containing about 3 million of words and more than 120 000 different words. This corpus is formatted and the spelling fluctuation is standardized.

1 Introduction

Transcribed speech corpora and huge text corpora are the core of systems used to construct acoustic and language models (Jelinek, F. 1976). Constructing of large transcribed corpora is time consuming and expensive, even if some researchers (Hughes et al, 2010; Badenhorst et al, 2009; Schlippe et al, 2012; De Pauw et al, 2009) are working on how to create automatically and quickly speech corpora by using different systems including phone applications.

If large transcribed speech corpora, more than 100 hours, exist for European languages like English, French or Spanish, the situation is quite different for African languages. About 2000 languages are spoken in Africa. Large part of them is not yet written and is today threatened of disappearing. Building speech Corpora and speech processing tools are crucial for each African language.

In this paper we present in section 2 the Somali language. Section 3 will focus on the first Somali read-speech corpus called ***Asaas*** (Beginning in Somali) and also the first Web-Based Somali Language Model and text Corpus called ***Wargeys*** (Newspaper in Somali) in Section 4.

2 Somali language

Four languages are spoken in Djibouti. French and Arabic are official languages, Somali and Afar are native and widely spoken. Somali and Afar are Cushitic languages within the Afro- asiatic family. Somali language is spoken in several countries in East of Africa (Djibouti, Ethiopia, Somalia and Kenya) by a population estimated between 11 to 13 million of inhabitants. The different variants are Somali-somali, Somali-maay, Somali-dabarre, Somali-garre, Somali-jiiddu and Somali-tunni. Somali-somali and Somali-maay are the most widely spread variants (80% and 17%). We only process the Somali-somali variant, commonly known as Somali language and spoken in Djibouti. The phonetic structure of this language has 22 consonants and 5 basic vowels which all occur in front and back versions (+ATR or -ATR). These 10 vowels occur in long and short pairs, giving 20 in total (Saeed, J. 1999). There are also 5 diphthongs which occur in front and back, long and short versions. Somali is also a tone accent language with 2 to 3 lexical tones (Hyman, L. 2010; Saeed, J. 1987; Gac, D. 2002). The written system was adopted in 1972, and there are no textual archives before this date. It uses Roman letters and doesn't consider the tonal accent in the current form. Somali words are composed by the concatenation of syllable structures (Bendjaballah, S. 1998. Saeed, J. 1999).

3 Somali Read-Speech corpus

3.1 Prompts Selection

A series of documents was selected from Somali online newspapers that use variant Somali-

Proceedings of the 2014 Workshop on the Use of Computational Methods in the Study of Endangered Languages, pages 73–76,
Baltimore, Maryland, USA, 26 June 2014. ©2014 Association for Computational Linguistics

Somali presented in Section 2. These texts are used to prepare the prompts. Particular attention was given to the quality of the selected texts (diversity of topics, phoneme distribution, readability, number of errors, etc.). Table 1 shows the distribution of selected text. It consists of 72,407 words (representing 2,335 sentences) with 12,807 different words.

Component	Amount
Sentences	2 335
Words	72 407
Different Words	12 807

Table 1: Distribution of selected text

3.2 Speakers selection

French and Arabic languages are official languages in Djibouti. The transcription of Somali language in Roman letters facilitates its reading. But it is difficult to find persons who can read it fluently. Fifteen Somali-speaking men living in Djibouti and without any problem of pronunciation were preselected. At last 10 people were selected for recordings according to their reading fluency. Table 2 shows information on their social class, their study level and their age. All recorders were volunteers.

Spakers Initials	Profession	Study Level	Age
Aaa	Technician	secondary	30-40
Abg	Researcher	University	40-50
Ahd	Journalist	secondary	40-50
Hha	Businessman	secondary	30-40
Hhdj	Technician	secondary	20-30
Hnm	Jobless	secondary	30-40
Ind	Technician	secondary	20-30
Ism	Policemen	University	40-50
Mar	Writer	University	50-60
sha	Writer	University	50-60

Table 2: Speakers Characteristics

3.3 Recordings

The recordings took place in the Djibouti Institute of Science and Information Technologies. They were recorded in mono in an office without any environment noise (fan, air-conditioner, phone, etc.) with a standard microphone with a sampling frequency of 16 KHz and a 16-bit encoding.

3.4 Corpus characteristics

The duration of the Somali read-speech corpus is 10 hours and 26 minutes. We named it ***Asaas***, which means "beginning" in the Somali language. Its phonetic distribution is given in Figure 1. We can consider that this distribution is well-balanced because it is quite similar to the one of the huge amount of the Somali text corpus (3 million of words). The phoneme /**a**/ occurs approximately 20 % of all the phonemes. The glottal phoneme /**'**/ (ʔ in IPA) represents about 0,2%.

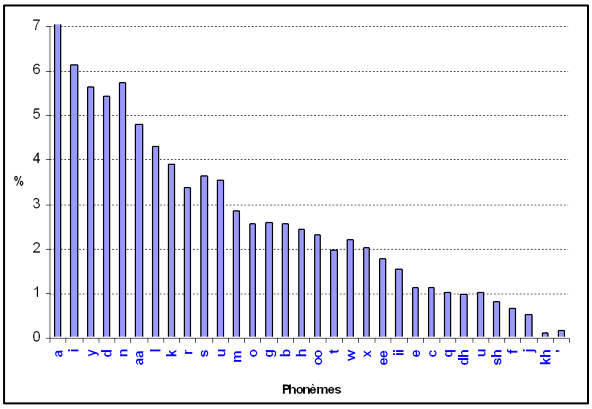

Figure 1: Phonetic distribution of Asaas Corpus

The duration of phonemes varies according to the speakers. However, in general, long vowels and fricatives are the longest ones. The short vowels and plosives have smaller duration. The phoneme which has the longest duration is /***sh***/ (ʃ in IPA). The average duration of all phonemes is given in Figure 2. Plosives phonemes are split into two parts: the burst (*Btt* for the /***t***/ phoneme, *Bkk* for the /***k***/ one and the occlusion *Ott* for the /***t***/ one and *Okk* for the /***k***/).

The average rate of the speech is 1.93 words per second.

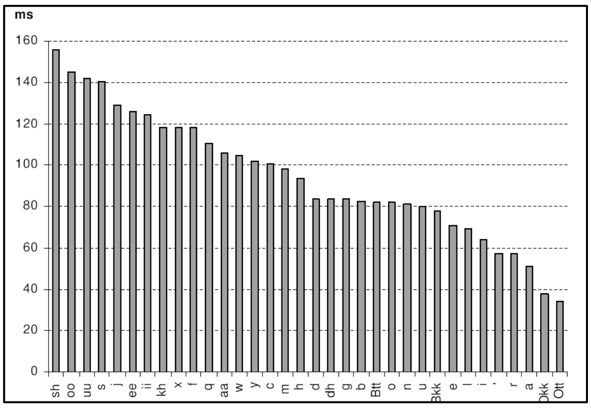

Figure 2: average duration of phonemes

There is a real difference between the duration of short and long vowels. Thus, long /***a***/ (written ***aa*** is Somali) is 2 times longer than the short /a/ (written ***a*** in Somali) and the /***uu***/ is 1.75 times longer than the /***u***/. On average, the ratio of duration between long and short vowels is 1.86. A comparison of the duration of long and short vowels is given in Figure 3. This feature can be used in acoustic modeling by creating two separate models for each vowel (long and short). It is also possible to recognize them with the language model. But both language model and separate acoustic models will probably improve the Word Error Rate.

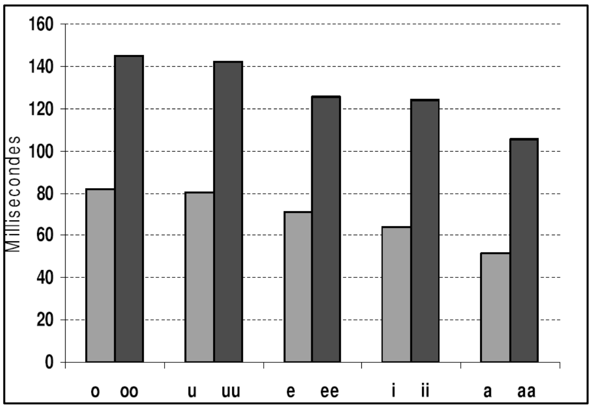

Figure 3: Duration of long and short vowels

4 Somali Text Corpus

Corpus containing millions of words or even billions of words is usually used for language modeling. If these data are mostly available for English and French languages, it is quite different for newly written languages like African languages. The lack of textual data constitutes a real handicap.

Grefenstette (2002) shows that African languages are gaining ground on the Web, even if European languages are largely dominant. Despite this growth their presence on the web is insufficient. This growth is relative, because in large part due to South African websites. Shigeaki (2008) clearly shows the prominence of South African sites on the continent.

Grefenstette and Nioche (2000) propose a formula to estimate the number of words found on the Internet for a given language. For this, we divide the number of times a word has been found by a search engine in cyberspace by the relative frequency of the word in that language. The average result on a predefined list of words provides an estimation of the number of words on the web. Estimation calculated in March 2014 for the Somali language gives about 500 million Somali words on the Web. The frequency of Somali words used was calculated on the textual corpus WARGEYS.

Many researchers are involved on how to create automatically textual corpora from the Internet for under-resourced languages (Ghani et al, 2001. Vaufreydaz et al, 1999). For our purposes we selected and downloaded a set of Somali newspapers on the Internet. As the audio corpus, the selection criteria were the variant of the language, the diversity of topics and the number of errors.

4.1 Formatting

The text downloaded is not directly usable. After removal of HTML tags, we proceeded to some transformations dealing with abbreviations, dates and times, numbers, proper names, foreign words and punctuation.

4.2 Spelling normalization

The Somali language as most of African languages is written after the independence period (1960-1970). So the orthography is not yet stabilized. The same word can be written in different ways according to people. Calvet, L. (1987) shows that the word eight in Mandingo is written *segin* in Mali, *seyin* in Guinea and *seegin* in Burkina Faso. In Somali Language the word President appears like *madaxweyne* or *madaxwayne*. This lack of standardization is common for most of the African languages and disrupts the language models quality as well as the Automatic Speech Recognition systems accuracy.

To resolve this problem, for a given simple word like *madaxweyne* we considered that the most frequent spelling is the good one. If *madaxweyne* appears 17 times in the corpus and *madaxwayne* 9 times, *Madaxweyne* is selected and all the *madaxwayne* were changed to *madaxweyne*.

For the component words like *iskumid* and *isku mid*, we choice the separate orthography like *isku mid*. This choice is made to separate syllable because if *iskumid* (Perhaps Out of Vocabulary Word) is not recognized by a ASR system, *isku* can be recognized or *mid* can be recognized.

4.3 Corpus Characteristics

Somali web-based textual corpus was named WARGEYS (Newspaper in Somali) because this corpus contains News topics and is similar to the French one called BREF (Lamel, L 1991) and gathered from the French newspaper Le Monde. Table 4 shows the characteristics of this corpus. WARGEYS contains 2 820 000 words with 121

000 different words and 84 000 sentences with an average of 33 words per sentence.

Component	Amount
Words	2 820 000
Different words	121 000
Sentences	84 000

Table 1: Distribution WARGEYS corpus

The figure 4 shows that the phonetic distribution of the two corpus **Asaas** and **WARGEYS** are similar.

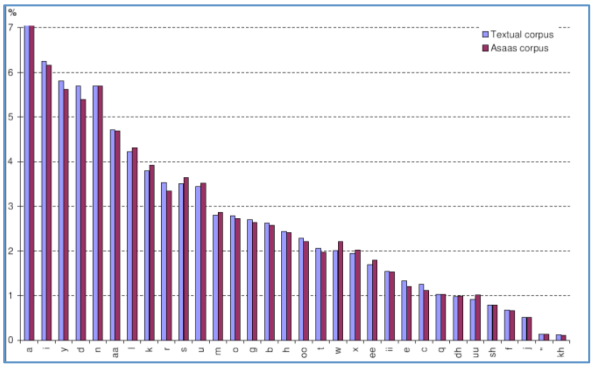

Figure 4: Phonetic distribution of Asaas and WARGEYS corpora

5 Conclusion

Somali read-speech corpus **Asaas,** consists of 10 hours and 26 minutes of good quality signal fully transcribed and well corrected. The phonetic distribution of this corpus is well-balanced. The Web-based Somali textual corpus **WARGEYS** contains 3 million of words and more than 120 000 different words. This corpus is formatted and the spelling fluctuation is standardized.

Reference

Badenhorst, J. Heerden, C. Marelie, D. Barnard. E. 2009. AfLaT '09 Proceedings of the First Workshop on Language Technologies for African Languages. Pages 1-8 ACL, Stroudsburg, PA.

Bendjaballah, S. 1998. La palatalisation en somali. *Linguistique africaine*, 21, 5-52.

Calvet, L. J. 1987. *Guerre des langues*. Payot, Paris.

Gac, D. L. 2002. Tonal alternations and prosodic structure in Somali. In *Speech Prosody 2002, International Conference*.

Ghani, R., Jones, R., & Mladenić, D. 2001. Mining the web to create minority language corpora. In *Proceedings of the Tenth International Conference on Information and Knowledge Management* (pp. 279-286). ACM.

Grefenstette, G., & Nioche, J. 2000. Estimation of English and non-English language use on the WWW. *arXiv preprint cs/0006032*.

Grefenstette, G., & Nioche, J. 2002. The WWW as a Resource for Lexicography. Lexicography and Natural Language Processing. A Festschrift in Honour of BTS Atkins. Göteborg, EURALEX, 199-215.

Hughes, T., Nakajima, K., Ha, L., Vasu, A., Moreno, P. J., & LeBeau, M. 2010. Building transcribed speech corpora quickly and cheaply for many languages. In *INTERSPEECH*, 1914-1917.

Hyman, L. M. 1981. Tonal accent in Somali. *Studies in African linguistics*, 12(2).

Jelinek, F. 1976. Continuous speech recognition by statistical methods. IEEE. 64(4), 532-556

Lamel, L. F., Gauvain, J. L., & Eskénazi, M. 1991. BREF, a Large Vocabulary Spoken Corpus for French1. *Training*, *22*(28), 50.

Mori, R. D. 1998. Spoken Dialogue with computers. Academic Press, London.

Saeed, J. I. 1987. *Somali reference grammar*. Dunwoody Press, Wheaton, MD.

Saeed, J. 1999. *Somali*. John Benjamins Publishing, Amsterdam.

Schlippe, T., Djomgang, E. G. K., Vu, N. T., Ochs, S., & Schultz, T. 2012. Hausa large vocabulary continuous speech recognition. *Proc. of SLTU*.

Shigeaki. K. 2008. Languages on the Asian and African Domains. *Proceedings of the International Symposium on CDG*.

Vaufreydaz, D., Akbar, M., & Rouillard, J. 1999. Internet documents: a rich source for spoken language modeling. In *IEEE Workshop ASRU'99 (Automatic Speech Recognition and Understanding)*.