Nimaan Abdillahi Kaourah

40 years later: automatic transcription of somali[46]

1. Introduction

Nowadays, we are witnessing renewed interest in the oral traditions and African history. The works of [1] and [2] have found proof verifying the approach that oral traditions can be considered as one of the sources of African history. However, this ancestral knowledge that has been accumulated for centuries is threatened due to the globalisation process, social transformation and lack of safeguarding means. This is what brought the eminent defender of the oral tradition, Cheick Amadou Hampâthé Bâ to say that "*When an old man dies in Africa, it's like a library that burns*".

Today, most of the concerned countries have at their disposal important audio databases that local radio stations have recorded many decades ago. These countries are faced with two questions: safeguarding this patrimony through a digitalisation programme and thus making it more accessible. Regarding the first point, the techniques are well known, and the problem of digitalisation which is being carried out in several countries is limited to a logistical level.

The second point is more sensitive because the utilisation of the audio databases, particularly those with large sizes requires a high level computer processing for all the languages of the concerned countries such as automatic transcription tools and indexing. The development of such tools requires large quantities of transcribed speeches and texts for different modelling. This is by itself a handicap for countries, particularly those with an oral tradition and do not have enough textual corpus. Ever since the attainment of their independence, several African countries have developed more or less complete writing systems, although the spelling is not standardized in most of them [3] e.g. the case of the word "eight" in Mandingo is written *segin,* in Mali *seyin,* in Guinea and *séegin* in Burkina Faso. Sometimes, in the same country different spellings are applied for the same word. This lack of

[46] I would like to express my gratitude to Idris Said Mïguil (Somali-Speaking Pen) and Abdallah Haji (Somali-Speaking Pen & RTD)

standardisation in the spelling is something usual in African languages.

This paper exposes the first signs of automatic processing of the oral cultural heritage of the Republic of Djibouti. As a first step, a presentation is made on the Somali language and the lack of spelling standardisation as well as on the solutions proposed. Then we describe the experiences of the Automatic Speech Recognition (ASR) in the Somali language affected in words, in roots and in a hybrid system combining the words and the roots as well as the different corpus constituted for this study. Finally, we draw conclusions from these works and state possible future research lines.

2. The somali language

Four languages are spoken in Djibouti: French and Arabic, which are the official and minority languages; Afar and the Somali, which are the indigenous and majority languages. Our present works focus solely on the Somali language that concerns half of the audio-targets records[47]. An estimated 12 to 15 million people speak this language in several East African countries. It is itemised under the Cushitic sub-family of the Afro-Asiatic languages in the international classification. The Somali-Somali variant, commonly called the Somali language and spoken in Djibouti is more specifically targeted in our researches. Its phonetic system is composed of 22 consonants and 20 vowels (5 long and 5 short with ATR: Advanced Tongue Root).

Table 1 presents the phonetic structure of the consonants. It is also a pure-tone language with two or three different tones [6], [7]. Its graphic form is relatively young, as it is only written in Latin letters since 1972. There was no written document prior to this date. The transcription of a word emanates directly from its phonetic realisation (each phoneme is represented by a letter or by two letters for some phonemes such as /dh/ /sh/ and /kh/.

3. Problem of standardisation

The Somali language is a "young" language in its written version, and the same word can have different spellings. The spelling variations of the words in this language can be grouped together in three categories. The first category consists in doubling or not doubling a consonant. Thus, the same word, like *director*, appears indifferently with two /d/ (GUDDOOMIYAHA) or with only one /d/ (GUDOOMIYAHA). The same author or journalist often uses the same spelling.

The second category is the appearance of a word under the form of compound-word or not. Thus, the word *communication* appears under the form of a single word ISGAADHSIIN or two words IS GAADHSIIN or compound-word IS-GAADHSIIN. The same goes for KA DIB and KADIB (after), KUXIGEEN (*Deputy* or *Assistant*) and The third category, which is the most frequent, is the same word which is written in two different ways, like WEYDIIYAY and WAYDIIYAY (to ask), JABUUTI and JIBUUTI (Djibouti), RAYSAL and RA'IISAL (President), etc. These multiple transcriptions cannot be considered as errors, since there is no standardisation imposed todate. However, they disturbthe quality of the language models as well as the robustness of the Automatic Speech Recognition systems. In order to circumvent this problem, and due to the lack of an official standardisation of the transcription of the words in the Somali language, we have adopted the following strategy. The latter is not meant to carry out any choice between the differenttranscriptions on the basis of whatever criteria, but opts for a given transcription in order to be able to move forward in our study. The transcription thus accepted does not have anything particular in relation to the others that are not accepted. The only chosen criteria are of quantitative or strategic nature.

For the third category of words, the spelling which is most frequently found in the corpus is selected. Thus, if WAYDIIYAY appears x times and WEYDIIYAY *y* times, and therefore if x > y, WAYDIIYAY is selected. For the second category of the words the forms in two words such as KU XIGEEN or IS GAADHSIIN are chosen. This choice has been made in order to later allow the recognition of the speech at the syllabic level (roots). For the first category of the words, the double consonants are replaced by simple consonants. Thus, we think that we have "fixed" the spelling for us to be able carry on with our study.

4. Automatic recognition of the somali language

We presented in [8] the first system of automatic recognition of speech in the Somali language. A trigram language was trained on a corpus of texts called WARGEYS (*Newspaper*) composed of almost 3 million words and of 121k different words. This corpus is made up of "broadcast news" type documents collected

[47] Djibouti Radio Television Audio Archives

from the World Wide Web. A lexicon composed of the most frequent 20k words from WARGEYS corpus has been extracted and later entirely transcribed into phonetics by SOMPHON phonetiser which is inspired by the French LIAPHON [4] phonetiser and developed to this effect. The language model thus obtained is made up of 726k bi-grams and 1.75M trigrams. The acoustic analysis is made on 30 ms windows taken every 10 ms. The acoustic signal emanating from the ASAAS (*foundation*) audio corpus entirely transcribed with a Transcriber [5], is parameterized by 39 coefficients: 12 MFCC coefficients and the energy, plus their primary and secondary derivatives. The parameters are centred and reduced. The acoustic models are composed of 3 states per phoneme, except for the "glottal occlusive" phoneme, which is coded with 1 state taking into account its execution briefness. For the experiments described in this paper, we used non contextual models with 128 gaussians per state.

The first experiments of ASR? were carried out on the corpus of speech test read for one hour HAATUF. The perplexity of this corpus calculated on WARGEYS corpus is 51.52% and the rate of Out of Vocabulary (OOV) is 4.90%. The large vocabulary speech recognition engine SPEERAL [6] has been used. The Word Error Rate (WER) obtained with a language model trained from the gross WARGEYS corpus (without standardisation of the spelling) and a lexicon of 20k words is 28.3%. The results of the analysis of the system have allowed us to bring it to the fore.

Ref: GUDDOOMIYAHA		gobolka oo uu WERIYAHAYAGU wax			
Hyp: GUDOOMIYAHA		gobolka oo uu WARIYAHAYAGU wax			
Ref: sheegay in waqti kale ay		**	*****		**U BALLAMEEN**
Hyp: sheegay in waqti kale ay		KU	TIMID		**BALAMEEN**
Ref: **DHAMMAYSTIRKA**		HESHIISYO		*****	hore U
Hyp: **DHAMAYSTIRKA**		BISHII		SIIYO	Hore UGU
Ref: dhexmaray oo	*	aanu	**FAAH**		**FAAHIN**
Hyp:dhexmaray oo	U	aanu	*****		**FAAHFAAHIN**

Table 1 – Examples of errors owed to lack of standardisation

Standardisation	Cor(%)	Sub(%)	Sup(%)	Ins(%)	WER(%)
None	76.4	**20.8**	2.8	4.7	**28.3**
HAATUF	83.7	14.7	1.6	5.2	**21.5**
WARGEYS. HAATUF	85.1	13.4	1.5	5.3	**20.2**

Table 2 – Results of experiments of ASR for different standardisations;

A large number of errors owed to different spellings of the same word between the differences and hypothesis. As is shown in example figure 1, pairs of words such as /GUDDOOMIYAHA/GUDOOMIYAHA, WEYDIIYAY/WAYDIIYAY, FAAH FAAHIN:

- FAAHFAAHIN etc. are counted as errors while it is only a question of different transcriptions of the same word.

In order to settle this problem and to estimate the actual error rate, we standardised only the spelling of the hypothesis supplied by the system as well as the one of the references (test corpus HAATUF). The WER has shifted from 28.3% to 21.5% (relative gain of 24%). Then after, we proceeded to the standardisation of the WARGEYS corpus. The results were then improved (WER=20.2%). A relative gain of 28% of the WER is acquired when the two corpus (WARGEYS and HAATUF) are standardised. All the results are grouped together in table 1.

5. Automatic transcription of the oral heritage

5.1. Automatic transcription of the Djiboutian Oral Heritage

The RADIO TELEVISION of DJIBOUTI (RTD) corpus is composed of an extract from a one hour broadcast about the awareness of the Djiboutian oral heritage. RTD is manually transcribed. 8 themes related to the historical events and personalities of the VIIth, XVIth, XIXth and XXth centuries were addressed. It is composed of 7,803 words with 2,378 different words. The OOV words rate is 12.48% for a lexicon of 20,000 words. This high rate is owed to the originality of the subjects treated. The records of the oral heritage that we wish to have access to are of a format similar to RTD Corpus (speech of dialogue-conversation, multi-speakers type etc.) Therefore, RTD will be the target of our research works. Let us point out here that similar corpuses/corpora in English or in French like those treated in the MALACH project [9] composed of stories and testimonies of the CHOA survivors are not easy to transcribe automatically. The error rates obtained and amounting to 40% are very far from those obtained with the read speech (read, journalistic, etc.). While an error rate of approximately 20% was obtained with the read speech, the latter goes up to 62% on the audio records of the cultural heritage. This can be explained by the important OOV rate (12.48%), the character "spontaneous speech" and "dialogue" of the RTD corpus as well as the temporal and thematic mismatch between the two corpora (training and test).

5.2. Automatic transcription in syllabus-roots

The previous experiments show the difficulty to transcribe automatically the oral heritage data that are distanced from the apprenticeships corpora. However, it will be difficult to find training corpora that are linguistically close to the data we wish to deal with. The usual obstacles faced in other languages-T [7] are intensified in our case due to the fact of these data are from an oral tradition country; with the absence of written forms prior to a certain date (1972 for the Somali). Consequently, we should find a sufficiently strong representation to the temporal and thematic gaps that give us the opportunity to directly have access to the old data that make up the heritage. [9]. This is why, the study of the recognition in syllabus-roots, whose number is limited, seemed to us an interesting avenue to explore. Indeed, the roots are the base of words formation and are found in most of the later (old or new, names of places, persons etc.). Moreover, even if the results that adopt a representation in roots are not readable, they could nevertheless allow an automatic indexing of audio archives.

The *WARGEYS* corpus has been split into roots as well as the reference files and this was through the SOMROOTS tool that was developed to this effect. WARGEYS-roots is composed of 6 million roots with 4,400 different roots. The words are on average composed of 2.14 roots. A lexicon composed of all the roots and entirely put into phonetic form has been used for the recognition of the roots. A model of language has been trained from the *WARGEYS-roots* corpus. This

model is composed of 189,000 bi-grams and 996,000 trigrams of roots. An OOV roots rate of 0.03% is obtained. The error rate of the system for a transcription of the RTD corpus based on the roots (RER: Root Error Rate) represents 47%. The hypotheses obtained are, of course, illegible because they are not words. But some OOV are entirely (*tafaraaruqa, qudhooda*) or partially *(asnaamtaasi)* recognised by the roots that compose them (Table 2).

In order to compare the results in words and that of in roots, the hypothesis obtained in section 5 has been split into roots. The WRER (46.4%) is slightly better than the RER, despite the important OOV rate and can be explained by the larger scope of the language model in words in relation to that in roots. Though this gap is relatively low, the errors produced by the two systems are not found at the same places. The system based on the words is good enough on the usual words (present in the lexicon) but make many errors on the OOV and in their surrounding while the system based on the roots has got a homogeneous behaviour for the two categories of words (in the lexicon or not).

5.3. Hybrid language model

The analysis of the previous results has led us to plan a recognition combining the words and the roots. This hybrid approach consists in learning a language model from a text composed at the same time of words and roots. The underlying idea is to benefit from the scope of language model in words, while enjoying taking advantage of the roots as far as the OOV words management is concerned. By choosing a restricted number of words – lexicon composed of more frequent words – we keep the bi-grams and tri-grams that appear more frequently. These structures make up the main "articulations" of the language. The remaining words not belonging to the lexicon are transformed into roots. This idea is implemented by [11]. We used a method similar where the words and the roots are not differentiated. The roots are considered as words. It means that neither both the distance and the proximity between the roots, nor those between the words and the roots are taken into account. The words of the lexicon are chosen among the most frequent *n* words of the WARGEYS corpus. These words are called In-Vocabulary (IV) words. All the other words are split into roots. The text thus obtained is composed of *n* words and almost 5,000 roots. This text will serve in for the hybrid language models of n words will be noted as *HLMn*.

Thus, we train different language models, *HLM0.2k* to *HLM20k*. In the same manner, we wanted to know the WRER (Word-Root Error Rate), the words of the hypothesis supplied by the hybrid systems were transformed in roots. These results are then compared with the previous results (former WRER emanating from the recognition in words and the RER obtained with the roots). The error rates in roots of the hybrid systems are better than those exclusively in words or in roots whatever the *n* size of the lexicon as is shown in diagram 2.

Table 2 shows a few OOV words recognized by the HLM*n* systems or the system in roots. The words between parentheses are not OOV words. We notice that in the system in words (*WLM20k*) the word Shiinaha (*Chinese*) which had to be normally recognized is disrupted by the OOV word that comes right after it qudhooda(*themselves*). The *HLM20k*recognises the word Shiinaha, followed by the continuation of the roots of the word *qudhooda*. This shows the flexibility and the greatest fluidity in the *HLM n*, systems. This can be explained by the fact that the hybrid systems are more "flexible" than the words systems. Indeed, the back-off phenomenon makes the systems in words rigid. As soon as we are faced with an OOV word, its immediate neighbourhood is disrupted while in the hybrid systems, the representation in roots of OOV words makes the system more "fluid".

6. Conclusion

The automatic recognition of the read speech gave a word error rate of 28.3%. The reading of the hypothesis supplied by this system led us to proceed to the standardisation of the spelling. A 28 % relative gain was obtained by securing uniformity only to the spelling of the test and training corpora (WER=20.2%). This first result gives an indication on the errors produced by the spelling fluctuation of the African languages. In order to validate the ASR system, we proceeded to the recognition of the RTD corpus extracted from the Djiboutian oral heritage. A 62.1% error rate is obtained on this corpus. The lessons we draw from this first phase is the difficulty in transcribing automatically the oral tradition records, knowing that it will be difficult to find a training corpus that is linguistically close to the data obtained. In the face of this result, we searched for a representation that is sufficiently strong to the temporal and thematic mismatch. Therefore, we turned towards the roots whose number is limited and that are the basis of the formation of the Somali language. A recognition in roots has given a Root Error Rate of 47.0%. When we split in roots the hypothesis of recognition in words the root error rate (WRER) thus obtained is 46.4%. The errors made are not situated in the same places. The system based on the words is good enough on the usual words and the system based on the roots has got a homogeneous behaviour for all the words (including the OOV). Finally, we planned a hybrid approach by using at the same time, the words and the roots thus benefiting from the scope of the language model in words while taking advantage of the roots as far as the OOV management is concerned. In order to be able to compare the different results, we also calculated the WRER of the hybrid systems. The results from these experiments are that the hybrid system's error rate in roots are better than those exclusively in words or in roots whatever their *n* size of the lexicon? (WRER=46% for *HLM20k*).

Future works will focus on the audio indexing of the African oral heritage by comparing the three approaches of automatic transcription (in words, in roots and hybrid). We will also try to reconstitute the words starting from the roots in order to be able to compare the results within a words' space.

References

- [1] Ch. Anta Diop, Nations nègres et Culture - De l'Antiquité nègre égyptienne aux problèmes culturels de l'Afrique d'aujourd'hui, Editions Présences Africaines,Paris, 1954.

- [2] G. Mocktar, General History of Africa II. Ancient Civilizations of Africa, University of California Press, Berkeley, 1981.

- [3] Louis-Jean Calvet, La guerre des langues et les politiques linguistiques, Payot, 1987.

- [4] F. Bechet, "Lia_phon : Un système complet de phonétisation de textes " Traitement Automatique des Langues, vol. 2, no. 1, pp. 47–67, 2001

- [5] C. Barras, E. Geoffrois, Z. Wu, and M. Liberman,"Transcriber : development and use of a tool for assisting speech corpora production," Speech Communication, vol. 1-2, no. 33, pp. 5–22, 2001.

- [6] P. Nocera, G. Linares, D. Massonie, and L. Lefort, "Phoneme lattice based a* search algorithm for speech recognition," in TSD2002. 2002, Brno.

- [7] Vincent Berment, "Méthodes pour informatiser des langues et des groupes de langues «peu dotées».," 2004.

- [8] A. Nimaan, P. Nocera, and J-F. Bonastre, "Automatic transcription of somali language," Pittsburgh, USA, 2006, Interspeech 2006.

- [9] I. Bazzi and J. Glass, "Modeling out-of-vocabulary words for robust speech recognition," in ICSLP, Beijing CHINA, 2000.

- [10] Ali Yazgan and Murat Saraclar, "Hybrid language models for out of vocabulary word detection in large vocabulary conversational speech recognition," in International Conference on Acoustics, Speech, and Signal Processing, Montréal, Canada, 2004, vol. 1, pp. 745–8.

- [11] M. Bisani and H. Ney, "Open vocabulary speech recognition with flat hybrid models," in INTERSPEECH 2005, Lisbon, Portugal, 2005, pp. 725–728.