PRESERVATION OF AFRICAN CULTURAL HERITAGE BY AUTOMATIC TRANSCRIPTION OF AFRICAN LANGUAGES

Abdiilahi NIMAAN, Pascal NOCERA***

* Institut des Sciences et des Nouvelles Technologies (Centre de Recherche de Djibouti)

** Laboratoire Informatique d'Avignon (Université d'Avignon et des Pays du Vaucluse)

nimaan.abdillahi@gmail.com, pascal.nocera@univ-avignon.fr

ABSTRACT

Most African countries follow an oral tradition system to transmit their cultural, scientific and historic heritage through generations. This ancestral knowledge accumulated during centuries is today threatened of disappearing. Automatic transcription and indexing tools seem potential solution to preserve it. This paper presents the first steps of automatic speech recognition (ASR) of Djibouti languages in order to index the Djibouti cultural heritage. This work is dedicated to process Somali language, which represents half of the targeted Djiboutian audio archives. We describe the principal characteristics of audio (10 hours) and textual (3M words) training corpora collected and the first ASR results of this language. Using the specificities of the Somali language, (words are composed of a concatenation of sub-words called "roots" in this paper), we improve the obtained results. We also discuss future ways of research like roots indexing of audio archives.

Index Terms— less-equipped languages, corpora, african languages, asr, hybrid language model

1. INTRODUCTION

Nowadays, we are witnessing that the interest for the oral tradition and the African history is being absolutely renewed. The works of [1] and [2] have found proofs verifying the approach that the oral tradition can be considered as one of the sources of the African history. This ancestral knowledge accumulated during centuries is threatened to disapearing 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 data bases that more often local radio stations have recorded for many decades. These countries are faced with two questions : safeguarding this patrimony through digitalisation programme and make it more accessible. Regarding the first point, the techniques are well known, and the problem for the digitalisation which is under realisation in several countries is limited to a logistical level. The second point is more sensitive because the utilisation of the audio data bases of large sizes requires a high level computer processing for all the languages of the concerned countries such as the automatic transcription tools and indexing. The development of such tools requires large corpora of transcribed speeches and texts for different modelling. This is by itself a handicap for countries labelled as being countries of oral tradition and that do not have enough textual corpus. Even if since the attainment of their independence several African countries have developed more or less completed writing systems, 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 standardisation in the spelling is something usual for the African languages.

This paper exposes the first steps of the automatic processing of the oral cultural heritage of the Republic of Djibouti. As a first step, presentation is made on the Somali language and the lack of the spelling standardisation as well as on the solutions proposed. Then, we make the description of the experiences in the Automatic Speech Recognition (ASR) in the Somali language effected 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 the future research main lines.

2. THE SOMALI LANGUAGE

Four languages are spoken in Djibouti : French and Arabic which are the official languages, the Afar and the Somali which are the indigenous and mostly used ones. Our present works focus solely on the Somali language that concerns half of the audio-targets records. 12 to 15 million people speak this language in several East-African countries [1]. It is clas-

[1] Ethnologue 2005

sified under the Cushitic sub-family of the Afro-Asian languages [2] . 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) [14]. It's also a tonal language with two or three different tones [6], [7], [13]. Its graphic form is relatively young, as long as it is written only since 1972 in Latin letters. 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 STANDARDIZATION

The Somali language is a "young" language in its written version, and the same word can be found written in different manners. The spelling variations of the words in this language can be grouped together in three categories.

The first one consists in doubling a consonant. (4). 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 variations are implemented from author to author.

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 *ISGAADSIIN* or two words *IS GAADHSIIN* or compound-word *IS-GAADHSIIN*. The same goes for *KA DIB* and *KADIB* (after), *KUXIGEEN* and *KU XIGEEN* (Deputy or Assistant).

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 till to-date. However, they confuse the quality of the language models as well as the robustness of the ASR systems. In order to circumvent this problem, and due lack of an official standardisation of the transcription of the words in the Somali language, we have adopted the following strategy. The later is not meant to carry out any choice between the different transcriptions 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 doesn't have anything particular in relation to the others that are not accepted. The only chosen criteria are of quantitative or strategic nature.

[2] Ethnologue 2005

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 chosen if not the opposite.

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 [4] the first system of automatic speech recognition in the Somali language. A trigram language model 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 from World Wide Web [5]. 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 one LIAPHON [6] and developed to this effect. The language model obtained is composed 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 Transcriber [7], is parametrized (parametré) 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 it 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 words (OOV) is 4.90%. The large vocabulary speech recognition engine SPEERAL [8] has been used. The Word Error Rate obtained with a language model trained on the WARGEYS corpus and a lexicon of 20k words is 28.3%. The results of the analysis of the system has allowed us to bring to the fore a large number of errors owed to different spellings of the same word between the dif-

Normalisation	Cor(%)	Sub(%)	Sup(%)	Ins(%)	WER(%)
Aucun	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 1. Results of experiments of RAP for different standardisations.

ferences and hypothesis. As is shown in example 1, pairs of words such as *GUDDOOMIYAHA/GUDOOMIYAHA*, *WEYDIIYAY/WYDIIYAY*, *FAAH FAAHIN/FAAHFAAHIN* etc. are counted as errors while it's 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 WARGEY 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.

Ref : **GUDDOOMIYAHA** gobolka oo uu **WERIYAHAYAGU** wax
Hyp : **GUDOOMIYAHA** gobolka oo uu **WARIYAHAYAGU** wax

Ref : ka **WEYDIIYAY** ARIMAHA AY ka wada hadleen WUXUU
Hyp : ka **WAYDIIYAY** MASTAR ILAAHAY ka wada hadleen UGU

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**

Fig. 1. Examples of errors owed to lack of standardisation.

5. AUTOMATIC TRANSCRIPTION OF THE ORAL HERITAGE

5.1. Automatic transcription of the Djiboutian oral heritage

The RTD[3] corpus is composed of an extract from 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 access to are of a format similar to RTD corpus (speech of dialogue-conversation, multi-speakers type etc.) Therefore, the RTD will be the target of our research works. Let's point out here that similar corpus 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 structured speech (read, journalistic, etc.). While an error rate of approximately 20% was obtained with the read speech, the later 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).

[3]Radio Television of Djibouti.

5.2. Automatic transcription in syllables-roots

The previous experiments show the difficulty to transcribe automatically the oral heritage data that are distanced from the training corpora. However it will be difficult to find training corpora that are linguistically close to the data we wish to deal with. To the obstacles we usually face for the languages-τ [10] are added in our case (country with an oral tradition) the absence of written data 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 oral heritage [4]. This is why, the study of the recognition in syllables-roots, which number is limited, seemed to us an interesting way to explore. Indeed, the roots are the base of words formation and are found inmost 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 to audio archives.

The WARGEYS corpus has been split into roots as well as the reference files and this, 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 in 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 language model has been trained from the WARGEYS-roots corpus. This model is composed of 189.000 bi-grams and 996.000 trigrams of roots. AnOOV roots rate of 0.03% is obtained. The Root Error Rate of the system for a transcription of the RTD corpus ba-

sed on the roots (RER : Root Error Rate) represents 47%. The hypothesis obtained are of course illegible, because they are not words. 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 have been split into roots. The WRER (Word-Root Error Rate) (46.4%) is slightly better than the RER, despite the important OOV rate. This 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 words 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 trigrams 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, 12, 13]. We used a method similar to that of [12] where the words and the roots are not differentiated. The roots are considered as words. It means that neither 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 the training of a hybrid language models of n words will be noted as HLM_n. Thus, we train different language model, $HLM_{0,2k}$ to HLM_{20k} . In the same manner, we wanted to know the WRER , the words of the hypothesis supplied by the hybrid systems were transformed into 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.

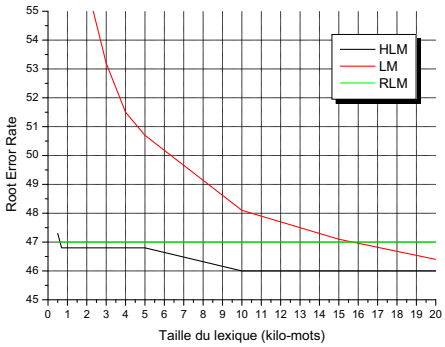

Fig. 2. Word-Root Error Rate for different HLM_n, compared with the WRER and RER obtained with system based on words (LM_n) and on roots (RLM).

Table 2 shows a few OOV words recognized by the HLMn systems or the system in roots. The words between parentheses are not OOV words. We notice that in the system in words (WLM_{20k}) the word *shiinaha* (chinese) which had to be normally recognized is disturpted by the OOV word that comes right after it *qudhooda* (themselves). The HLM_{20k} 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" that words systems. Indeed, the back-off phenomenon makes the systems in words rigid. As soon as we are faced with an OOV word, it 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 give 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 fluctuation of the African languages spelling. 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 of this first phase is the difficulty to transcribe automatically the oral tradition records, knowing it will be difficult to find training corpus that linguistically close to these data. In the face of this result we searched for a representation that is sufficiently strong to the

Référence	WLM 20k	HLM 20k	HLM 1k	RLM
asnaamtaasi[4]	wasaaradaasi[5]	as naam ta si	as naam ta si	as naam ta si
tafaraaruqa[6]	taf abaabulka	taf ar aar uq a	taf ar aar uq a	taf ar aar uq a
faaqidi[7]	nafaqada	faaq id i	faaq id i	aq ad i
(shiinaha) qudhooda	bishii lagu looga	shiinaha qudh ood a	bishii ina qudh ood a	bish iib a qudh ood a
(laba) dakhare[8]	labadaba	laba dakh ar e	labada kale	lab ad a sar e

Table 2. Some OOV words recognized by the RLM and HLM systems. The words between parentheses are not OOV words.

temporal and thematic mismatch (?). Therefore, we turned towards the roots which number is limited and that are the basis of the formation of the words in 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 Word-Root Error Rate (WRER) thus obtained is 46.4%. The errors made are not situated on 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. It results from these experiments the hybrid system's error rate in roots are better than those exclusively in words or in roots whatever their n size of the lexcion (WRER=46% for HLM20k).

The 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.

7. 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*, Hachette Littératures, 1987.

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

[5] L. Viet-Bac, B. Bigi, L. Besacier, and E. Castelli, "Using the web for fast language model construction in minority languages," in *Eurospeech*, Genève (Suisse), 2003.

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

[7] 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.

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

[9] B. Ramabhadran, J. Huang, and M. Picheny, "Towards automatic transcription of large spoken archives - english asr for the malach project," Hong-Kong, China, 2003, IEEE International Conference on.

[10] Vincent Berment, "Méthodes pour informatiser des langues et des groupes de langues "peu dotées".," 2004.

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

[12] Ali Yazgan and Murat Saraclar, "Hybrid language models for out of vocabulary word detection in large vocabulary conversational speech recognition," Montreal, Canada, 2004.

[13] M. Bisani and H. Ney, "Open vocabulary speech recognition with flat hybrid models," in *In Eurospeech 2005*, Lisbon, Portugal, 2005.