Jama Musse Jama

SOMALI LANGUAGE AND ITS STANDARDIZATION AS A WRITTEN LANGUAGE: DESCRIPTION AND COMPUTER TOOLS

Abstract

In Africa, it is recognized that the use of Information and Communication Technology in its development and education is a viable way to reduce the gap between the continent and the rest of the world, especially the Western countries. The creation of localized content in African languages, and tools for the development of such content, might be the first step to bridge the so called digital divide. However, there are still a number of obstacles, some of them technical, relating for example to how to handle diverse scripts and extended characters for some African languages (see the UNICODE philosophy in UC [2007]). While the Somali language does not encounter these technical issues, because it is written in basic Latin alphabet (Jama Musse Jama [2006]), there are nevertheless still other problems to be faced in the creation of adequate Somali content. One of the main obstacles for Somali is the unavailability of computer tools that can help and promote the use of the Somali language in storing electronic content. This paper addresses recent developments in ongoing research activities on the Somali language dealing with tools for Information and Communication Technology.

1. *Introduction*

The central theme of the research activities presented here is to develop computer tools to help the Somali language to approach a standardized form as a written language, and to produce localized content stored as documents or data on computers, and then eventually distributed through the Internet.

This paper should not be considered as a research presentation on linguistic aspects of Somali, but instead it addresses the current state of the art of the linguistic tools developed by the author, which are available for free to both ordinary personal computer users and professionals working on corpus linguistic research. Among the focal issues of these activities, the paper describes how to build an accurate corpus for spell checking dictionaries, tools for syntactic and morphological parsing and utilities converting text to speech for the Somali language.

REDSEA-ONLINE.COM started its activities on the Somali language 10 years ago. As a first tangible contribution of its activities, free word processing software, which includes a spell checker with more than 180,000 Somali mostly used words (list of word roots plus corresponding derivations), has recently been released in a beta version under the name of *Ubbo* 1.0. This word processor is being used by Somalis, mostly those authors who write in the Somali language, to spell check their works before being published. In this paper we will briefly describe how its corpus has been built and how it is continuously updated.

2. *'Ubbo' - Somali Language Online Spell Checker and Word Processor*

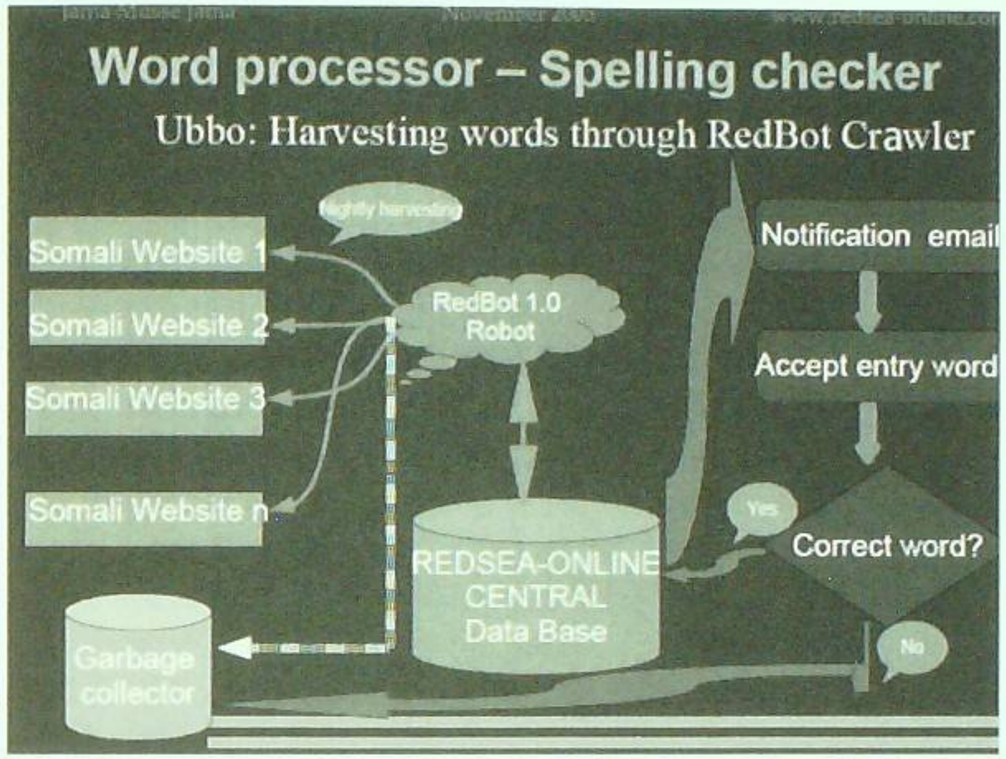
Ubbo is a freely downloadable word processor that runs on Microsoft Windows operating systems. It has a localized Somali language user interface, and is mostly used by Somali authors and website managers basically for the purpose of spell checking. It is also available as a free online tool that allows everyone to check spelling errors in the Somali language text without necessarily installing the *Ubbo* desktop application. This project was the first milestone for REDSEA-ONLINE.COM in contributing to the development of African languages in the era of Information and Communication Technology.

The project is a community initiative to develop a comprehensive Somali word list corpus. The created corpus will be

incorporated into several open source software packages to spell-check the Somali language. Therefore the main target of the project is to create an official Somali word list for universal and standard spell checkers.

As of today, the corpus main dictionary contains the 183,520 most used Somali words (around 42,000 lemmas and their derivations through a simple morphological parser), and it is growing on daily basis, thanks also to the visitors who are using these tools. In fact every new word not yet listed in the main dictionary will be collected from inserted text, and then a notification e-mail will be sent to a group of Somali language experts. The group members will examine the new words and decide whether or not to add these words to the main dictionary.

The dictionary is also growing through the *REDBot* crawler, which is a web crawler or web bot[1] that retrieves text from web pages and follows up hyperlinks they contain. The *REDBot* crawler starts at midnight every night on randomly selected websites, harvesting new words recursively from other websites linked to the just visited pages. The Somali Language Spell Checker accepts all harvested text from the *REDBot* periodically and generates indexes of the all new words found therein. Whenever *REDBot* collects a new word for a predefined number of times (for example, a word which has been found 50 times from different web pages), it assumes that the word is Somali, and therefore it automatically sends a notification e-mail to the dictionary maintainers, who will approve or decline the correctness of the word. If the word is accepted, it will be added to the word list, otherwise it will be put in a "garbage" collection database. Words considered "incorrect" will therefore be put in the garbage collection database and will not be collected during future "word harvesting" through the crawler.



The above scheme shows how the *REDBot* crawler works to collect new words from the Internet.

When used as a desktop word processor, *Ubbo* permits users to add their words in their local dictionary. It also permits them to upload their collected words to the central main dictionary on the REDSEA-ONLINE.COM server. In this case all the suggested words will be notified to the dictionary management committee who will then decide whether to accept the new entries. Once approved, a notification e-mail is sent back to the contributor who can automatically update his/her version of *Ubbo* with his/her words.

3. *Syntax and morphological mini parser for the 'Ubbo' corpus*

The collected corpus mentioned above contains a tokenized word list which comprises only the base words. Nouns and verbs are written in their basic form. In order to maintain an acceptable size of the corpus, we defined a mini parser that uses simple resources with morphological information, a

derivative suffix and part of speech category definitions. These resources are defined in a rules file that *Ubbo* uses to expand its corpus.

For example, the following rule, defined in the rules file, allows the *Ubbo* spell checker to recognize more words than those that are statically included in its corpus dictionary.

Example 1 – *Definite article for Somali masculine nouns*

The rule defined in the rules file:
DIBRAAC: R 0 k? . dr:ka kii kee keer kaa kaygii keennii kiinnii koodii

The dictionary file contains the following three lines:
nin /R
cad /R
qalin /R

In this case, the *Ubbo* spelling checker recognizes dynamically the following words as corrected:

nin	cad	qalin
ninka	cadka	qalinka
ninkii	cadkii	qalinkii
ninkee	cadkee	qalinkee
ninkeer	cadkeer	qalinkeer
ninkaa	cadkaa	qalinkaa
ninkaygii	cadkaygii	qalinkaygii
ninkeennii	cadkeennii	qalinkeennii
ninkiinnii	cadkiinnii	qalinkiinnii
ninkoodii	cadkoodii	qalinkoodii

Example 2 – *Inflection of verbs ending with the vowel i*

The rule defined in the rules file:

DIBRAAC: S 0 i . dr:yi yey nnay day deen yeen yayey nayney yaysey yeyseen yeyney yayaa nnaa daa yaysaa yaan doonaa doontaa doonnaa

The dictionary file contains the following one line:
sahwi /S

In this case, the *Ubbo* spelling checker recognizes dynamically the following words as corrected:

sahwiyi sahwiyey sahwinnay sahwiday sahwideen sahwiyeen sahwiyayey sahwinayney sahwiyaysey sahwiyeyseen sahwiyeyney sahwiyayaa sahwinnaa sahwidaa sahwiyaysaa sahwiyaan sahwiyidoonaa sahwiyidoontaa sahwiyidoonnaa

A set of similar rules define different ways to recognize words derived from the base words regarding the vocative case, gender, number, pronoun and indicative mood and inflection of verbs. These are not exhaustive rules but they seem to cover a good percentage of the most used forms in Somali. They are taken from Andrzejewski [1964, 1979], Yaasiin Cusmaan Keenadiid [1976], Banti [1988], Saeed [1993], Abdalla Omar Mansur and Puglielli [1999], Rabi [2004] and Carab [2004].

4. *Word suggestions for misspelled entries*

The main technique for the *Ubbo* spell checker to list suggestion candidates is given by their 'closeness' to the misspelled word. This is established through a standard character swap and deletion method to calculate the "distance" of the candidates from the original word. However *Ubbo* also considers specific peculiarities of the Somali language. For example, the replacement of one or more characters by other characters. The following rule in the rules file defines a suggestion based to the replacement of R by Dh and vice versa.

BEDDEL: r 2 - dh

In this case if the misspelled word of the text is "ariighii", the following list of suggestions will be given:
arigii
adhigii
farriimihii
arrimahaagii

Other examples of Somali peculiarities are the double consonants and the long vs short vowels. *Ubbo* tries to suggest first when there are missing double consonants on the following BDGMNLR consonants as defined in the standard Somali grammar rules (Yaasiin Cusmaan Keenadiid [1976]), or if there is a wrongly doubled consonant. The same happens in the long vs short vowels.

5. *Text to speech utilities for the Somali language*

"Waa kuma?" (who is it?) is a software application of which the main purpose was to assist a Somali speaking person with reduced vision problems, to recognize the incoming calls from the home telephone. The software is able to read the calling number from the phone device and then to produce Somali speech using a synthesis method from an open source TTS (text-to-speech) engine. If the calling number is already registered in its address book, *"Waa kuma?"* announces the name of the caller, otherwise it just announces the telephone number in Somali language.

Starting from this tiny application, REDSEA-ONLINE.COM has developed a new TTS system that converts normal Somali language text into speech. The main idea remains to allow people with visual impairments or reading disabilities to listen to works on a home computer. However, there are other applications that can employ these tools (see the paragraph "further developments").

6. *Somali poetry parser*

Somali poetry is accomplished by a combination of scansion pattern rules and the sound alliteration rules, and although orally composed, a structural analysis similar to Western poetics can be applied to Somali verses. In fact, Somali verse is marked by alliteration and the use of metrical system. This system of metre has recently been documented in Cabdillaahi

Diiriye Guuleed [2003] and Faarax [2007] relying on the research initiated in 1976 by Maxamed Xaashi Dhamac "Gaarriye" (Maxamed Xaashi Dhamac [1976]) and Cabdillaahi Diiriye Guuleed "Carraale" (Cabdillaahi Diiriye Guuleed [1978, 2003]). It is not the purpose of this paper to present the metrical system of Somali poetry. For further details in English, see also Johnson [1979].

The idea of computer assisted poetry is not new to literature. For example, for French poetry, Nell [2003] has developed a method using the computer to detect rhythmic patterns in the traditional isometrical alexandrines. A similar project for Somali was announced recently in Faarax and Liibaan [to appear]. For English literature, the use of the computer started with the early use of computers. In fact, in 1951 a team of scientists tested the capabilities of the world's computer, Mark One "Baby", which was used to experiment in composing love poetry[2].

Based on these recent publications on the metrical system of Somali poetry, REDSEA-ONLINE.COM has developed a parser to recognize the correctness of the metrical structure and to identify the genre of a given Somali poem. The idea is to check not only the metrical system for Somali poetry, but also to develop a new module for a Robot being able to compose a Somali poem.

7. *Further developments*

Can a computer compose a Somali poem?

The challenging response to this question might be affirmative if we consider how the above mentioned tools can interact. In fact, the scheme of a robot being constructed by REDSEA-ONLINE.COM, and expected to be released by the end of the year 2009, is based on the following components:

- a comprehensive and accurate Somali corpus (list of base word lemmas) which includes also synonym and antonym definitions;

- a syntactical and morphological parser for the Somali language;
- base knowledge data: a rich database containing a collection of both prose and poetry texts. This includes a wide variety of Somali literature texts;
- grammar checking for Somali (a new module to be developed);
- a parser module for Somali poetry;
- text to speech converter (improvements on the *Waa kuma?* application).

Each module listed above is responsible for a specific part of the work. The following is a rough and simplified scheme of the algorithm to compose a poem:

1. The user provides the Robot with a letter and a subject. For example the letter 'g' and the word for the concept "Gobannimo". Gobannimo is a very complex word and has a broad meaning including freedom and liberty, but also implying a dignified, coherent, independent, giving, tolerant, and respectful approach. In this case all these concepts will be found in the antonym and synonym definitions of the word.
2. The Robot makes an index to obtain, from the lemmas of base words, different words starting with 'g' in the different types of speech (noun, verb, adverb, etc.) It also looks for other words that are synonyms and antonyms of the selected words (here we need to define a synonym correspondence of the lemmas word list).
3. For each word selected in step 2, the Robot performs a new search from the database of prose and poetry, to find sentences that include the word.
4. Each word in the identified sentences will be analysed by the morphological parser module, to find the root of the word, and such root words will be grouped according to their type of speech.
5. Using the grammar checker module on the grouped words, the robot composes new sentences.

6. Using the poetry parser, the robot composes metrically correct verses.
7. Finally, using the text to speech convertor, the robot recites a Somali poem which will certainly be semantically correct, but of which some of the verses may not make sense at all.
8. The recited text will be also printed as output. In the case of nonsense sentences, the user may make the necessary fixes, having at their disposal a list of words that can be used to substitute for the randomly generated words.

The ability of the Robot to produce acceptable results relies much on the knowledge database containing the prose and the poetry text, and the definition and accurateness of the base word list.

8. *Conclusions*

Linguistic software tools for the Somali language are much needed and, if they are available for free, their impact will be more profitable for the language itself to come to a standard written form. One of the missing modules, which is urgently needed, is the formalization of a morphological grammar. It is understandable that a Somali grammar checker will be difficult to be realized in the near future, because there is a need for more linguistic research on the Somali grammar itself before realizing a software that implements its rules.

The above listed activities are part of a first stage of assistance to those who have access to ICT and who need to write in Somali in their daily activities (spell checker and grammar checker). Of course they cannot cover the demands of a user who expects all the IT tools which are available for other languages. The fact that, as of today, there is no focal point or centre of research for the Somali language, which takes responsibility for coordinating all these kind of activities, makes the challenge yet more difficult. It is time to think seriously of establishing such a centre where experts may be employed in a wide variety of fields, including linguistics and information technology.

NOTES

[1] Internet bots, also known as web robots, WWW robots or simply bots, are software applications that run automated tasks over the Internet. Typically, bots perform tasks that are both simple and structurally repetitive, at a much higher rate than would be possible for a human alone [en.wikipedia.org/wiki/Internet_bot accessed 23/3/2009].

[2] The Small-Scale Experimental Machine, known as SSEM, or the "Baby", was designed and built at The University of Manchester, and made its first successful run of a program on June 21st 1948. See for the story of "Baby" on [www.computer50.org accessed 23/03/2009].

REFERENCES

Abdalla Omar Mansur, Puglielli, A. (1999), *Barashada Naxwaha Af Soomaaliga*, London: Haan Publications.

Agostini, F., Puglielli, A., Ciise Moxamed Siyaad (1985), *Dizionario Somalo Italiano*, Roma: Gangemi.

Andrzejewski, B.W. (1964), *The Declensions of Somali Nouns*, London: School of Oriental and African Studies, University of London.

Andrzejewski, B.W. (1979), *The Case System in Somali*, London: School of Oriental and African Studies, University of London.

Banti, G. (1988), *Reflections on Derivation from Prefix-conjugated Verbs in Somali*, in Puglielli A. (ed.), *Proceedings of the Third International Congress of Somali Studies*, Roma: Il Pensiero Scientifico, pp. 43-95.

Cabdillaahi Diiriye Guuleed (1978), *Hawraarey Ninba Si Kuu Qaaday*, Muqdishu: Xiddigta Oktoobar.

Cabdillaahi Diiriye Guuleed (2003), *Miisaanka Maansada Soomaaliyeed*, Sweden: Abokers Forlag.

Carab, S.H. (2004), *Qaamuus, Ereykoobaha Af Soomaaliga*, Jabuuti: Machadka Afafka ee Xarunta Cilmibaadhista Jabuuti.

Faarax, C.C. (2007), *Mahadhaada iyo waxqabadka Maxamed Xaashi Dhamac "Gaarriye"*, Calgary: Hal-aqoon publishers.

Faarax, C.C., Liibaan, H.Y. (to appear), *Laaxinjire: Qorme kombiyuuter oo kuu saxa Maansada*, To appear (the author announced their Computer Program in a radio interview during the 10th Somali Studies International Association Congress).

Jama Musse Jama (2006), *Proposal of Soomaali custom keyboard*, http://www.redsea-online.com/soomaali, August 2006.

Johnson, J.W. (1979), "Somali prosodic systems", in *Horn of Africa*, 2 (3), pp. 46-54.

Maxamed Xaashi Dhamac (1976), *Toddobaadkan iyo Suugaanta: Miisaanka Maansada*, Muqdishu: Xiddigta Oktoobar.

Nell, S.D. (2003), "Toward a theory of rhythm in French poetry: Computer assisted recognition of rhythmic groups in traditional isometrical alexandrines", in *Computer and Humanities*, 27, pp. 185-223.

Rabi, H.M. (2004), "Somali Syntax: Some Common Features, in War Destroys. Peace Natures, Somali Reconciliation and Development", in Ford, R. B., Adam, H. M., Ismail Edna Adan (eds.), *Selected papers from the 8th Somali Studies International Association Congress*, Lawrenceville, NJ: The Red Sea Press, pp. 41-58.

Saeed, John I. (1993), *Somali Reference Grammar*, Kensington, MD: Dunwood Press.

The Unicode Consortium. The Unicode Standard, Version 5.1.0, defined by: *The Unicode Standard, Version 5.0*, Addison Wesley, Boston, MA; 2007.

Yaasiin Cusmaan Keenadiid (1976), *Qaamuuska Af Soomaaliga*, Muqdisho/Firenze: Madbacadda Qaranka.

WEBSITES

http://redsea-online.com/ubbo

http://redsea-online.com/soomaali