Jama Musse Jama

Can a Computer compose a Somali poem after 40 years of Somali Language written experience?

1. Introduction

The idea of computer assisted poetry writing is not new to literature. For French poetry, Nell [2003] has developed a method using the computer to detect rhythmic patterns in the traditional *isometrical alexandrines*. For English literature, the use of the data processing machines started with the early use of computers. In 1951, in fact, a team of scientists tested the capabilities of the world's first computer, *Mark One "Baby"*, which was used to experiment in composing love poetry [SSEM, 1948]. Recent studies showed that, more than purely computer generated poetry, the interesting aspect of the use of computers in poetry is the valuable assistance that a poet can get from computer engagement in poetry composition. As Hartman (1996) writes the issue "isn't exactly whether a poet or a computer writes the poem, but what kinds of collaboration might be interesting" and he describes in his book *Virtual Muse: Experiments in Computer Poetry* how computer programming helped him probe poetry's aesthetic potentialities [Hartman, 1996]. For the Somali language, and for many other African languages, little has been done so far in IT related research and development [*see* Jama Musse (2009), Cabdiraxmaan C Faarax 'Baraaqo' & Xuseen Yuusuf Liibaan (*see* References).]

Somali verse is marked by alliteration and the use of scansion metrical system.

Alliteration: rules of alliteration that govern the Somali poetry may be simplified within these three fundamental points: i. poems with short lines, which have no caesura, each line should contain at least one alliterative word; ii. poems with long lines, divided into two parts (*hojis* and *hooris*), each part should contain at least one alliterative word; iii. poems of all types, the same alliteration should be used in all lines (and parts of each line when divided into two parts) [*see* for example, Andrzejewski 1982]. For this limitation of the alliteration word and the rhyme scheme that these words should pattern, a poet should have an accumulation of large vocabulary in order to express his or her idea.

Scansion: Somali scansion metric system was articulated in the early 1970s [*see* Gaarriye 1976; Carraale 1978, 2004; Johnson 1979, 1984; Banti and Giannattasio 2004] and it was recently further developed in research on consonants and their role as virtual geminates [*see* Orwin and Gaarriye, 2010]. A different and a very interesting aspect of metrical system is also presented by Faarah Ali Gaamuute in his upcoming book [*see* Faarah Ali, 2012]. Somali poetry is therefore 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 and "the time of establishing the metric units, the rules and procedures that make Somali poetry possible has finally come." [*see* Faarah Ali, *id.*]

It is not the purpose of this work to go into detail on the study of alliteration and metrical system in Somali poetry. Suffice to say that we will use *mora* (*morae* in plural) as a unit of measure in phonology that determines the weight of a syllable in each word of the verse; this measurement also determines stress and timing in the verse. More specifically we will count one *mora* for a short syllable and two *moroe* for a long syllable. The determination of the type of genre of a *maanso* consists of totalling the number of *moroe* in each half line of the verse; and these time units (*morae*) should be a constant series of distribution. For example in the case of *jiifto,* the following pattern can be observed (2-2, 1-2-2-1):

2 2 , 1 2 2 1 (total 8 *morae*)

Naa yaa, dayooy hee dheh

So the total number of *morae* should be 8 or 9, with the above mentioned distribution. If a verse respects this condition, we will say that it is a metrically correct *maanso* verse in the *jiifto* genre. The same rules will apply in other genres (for instance *gabay, geeraar, hees,* etc.)

This paper proposes a new algorithm and presents a set of stand-alone modules based on the structure of the Somali language and on the metrical system of the Somali poetry. This algorithm uses a large Somali corpus (list of base word lemmas) which also includes synonym and antonym definitions; syntactical and morphological parsers for the Somali language; and a base knowledge data, which is a rich database containing a collection of both prose and poetry texts. Other auxiliary modules include grammar checking for Somali; and a parser module for Somali poetry.

The idea is to put together all these modules to give an affirmative answer to the very ambitious question that is: Can a computer compose some semantically correct verses in Somali? And can it produce verses that also make sense in the spoken language?

This paper is organized as follows: in Section 2, we describe briefly major components that constitute the bases of the algorithm, and for each component the reader is reminded of the previous work from which this component comes from. Section 3 constitutes the core of the paper, and it outlines the chart-flow of the algorithm, and the major passages of the logical flow are explained. Section 4 describes further the algorithm steps by using a specific example to generate a poem. We conclude in Section 6 with a discussion of the main insights gained from these experiments and also by reporting on the different tests carried out, and then the machine produced verses are compared with parts of well-known Somali poems.

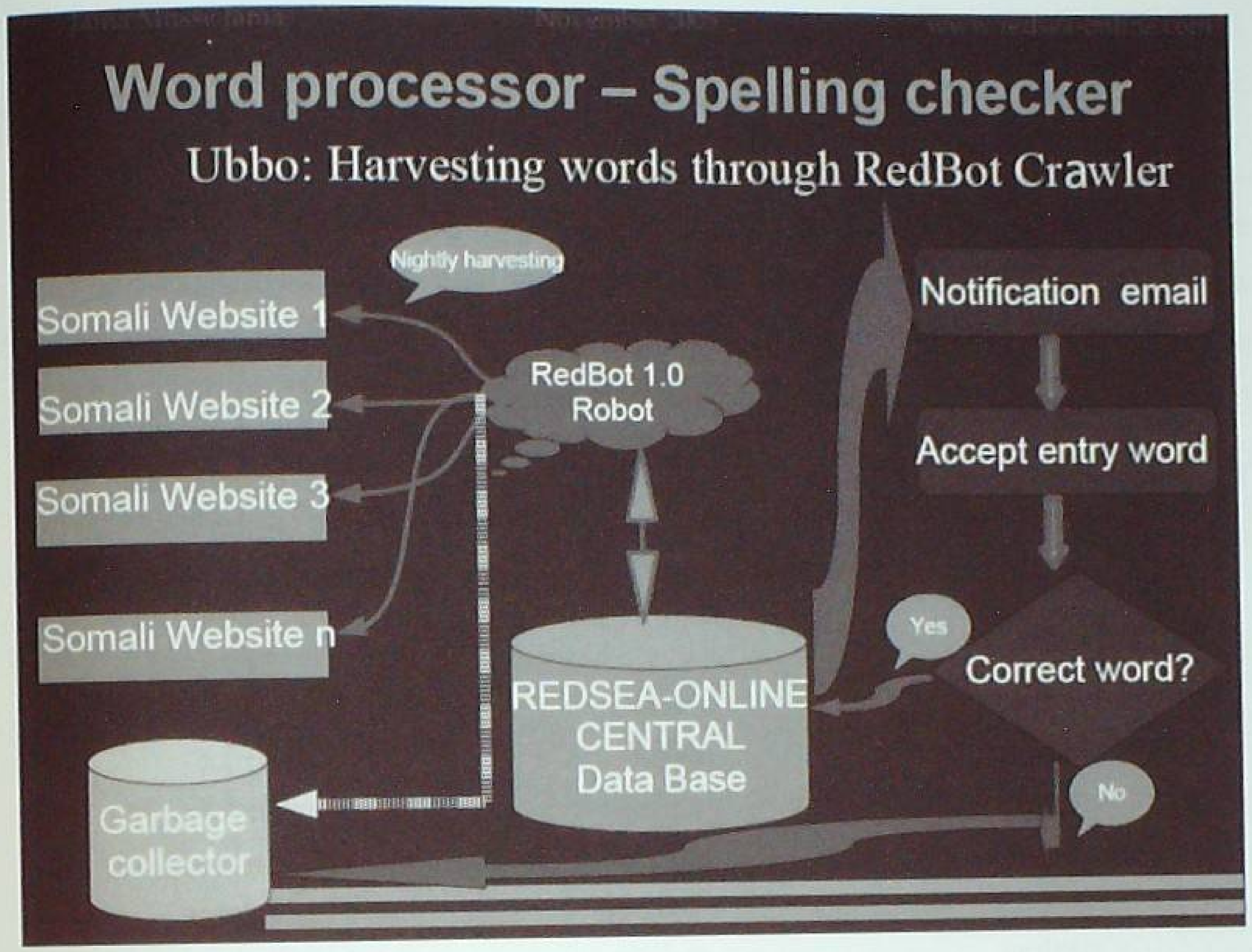
2. Brief description of the components:

2.1. Word List Database (WLD):

WLD module is a comprehensive and accurate Somali corpus consisting of a list of base word lemmas, which includes also synonym and antonym definitions of the words. Redsea Online has collected nearly 150,000 unique terms from its online spell-checker and through a web crawler. It contains about 42,000 lemmas and their derivations through a simple morphological parser. These words have been confirmed as either lemmas (base forms) or morphemes (grammatical forms), and part of them have link relationships with other words. Of these, nearly 5,000 have been matched to a synonym and/or antonym definition. *WLD* is growing on a daily basis, thanks also to the visitors of the online edition of Redsea Online dictionary. 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. *WLD* has also the capability to self-feed and collect new words from its usage while the programme is analysing new poems. The idea is to enrich the database so all words match in grammatical relationships (parts of speech, combination with articles, etc.) and have defined equivalent meanings (synonyms, antonyms, etc.).

The *WLD* is also growing through the *REDBot* crawler, which is a web crawler or web bot[45] that retrieves text from web pages and follows up the 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 all the 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.

[45] 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].



Word processor – Spelling checker

Ubbo: Harvesting words through RedBot Crawler

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

2.2. Parser for Somali:

PfS is a syntactical and morphological parser for the Somali language. It is a system which analyses an input text syntactically and morphologically and converts each sentence into a structured meta-data representing as its parts of speech. In the tokenized word list, we have verbs and nouns written in their basic form. This mini parses uses non-exhaustive rules, using simple resources with morphological information, a derivative suffix and part of speech category definitions.

For example the following rule applies to identify the 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 PfS rule definitions file contains, for instance, the following three lines:

- nin /R
- cad /R
- qalin /R

In this case, the mini parser will identify and/or generate during the verse composing, 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

The following is another example, to identify derivative words through inflection of verbs ending with the vowel i. In this case:

a. The rule defined in the rules file is:

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

b. The *PfS* rule definitions file contains the following line (for example for the verb sahwi)

sahwi /S

c. The *PfS* module recognizes dynamically the following words and can trace back the verb "sahwi" to 'figure out' the meaning of the current verse; or to build the suitable word for the poetry metric rule: 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. These rules are taken from Andrzejewski [1964, 1979], Keenadiid [1976], Banti [1988], Saeed [1993], Cabdalle and Puglielli [1999], Rabi [2004]. The idea is to enrich these rules, and make robust and modular morphological and syntactical parser for Somali.

2.3. Base Knowledge Data (BKD)

BKD consists of a rich database containing a collection of both prose and poetry texts in Somali. This unstructured text database includes an index of a wide variety of Somali literature texts; a collection of Somali poems in different genres; a collection of proverbs, traditional wisdom sayings. This module helps the algorithm to "guess" words making sense in combination with other words.

2.4. Grammar Checker (GC)

GC is a module for basic grammar checking for Somali language. It is a simplified grammar checker for poetry verse syntax control. For instance, if there is a subject in the verse, which is masculine singular third person, it verifies that the form mode used in the verb is correct grammatically. The idea is to develop a comprehensive grammar checker for the Somali language.

2.5. Poetic Verse Templates Collector (PSTC)

There are a minimum set of verse sentence templates at the initial stage of the programme. Every time a verse is approved by the administrator as a metrically and grammatically correct verse in a specific genre, the *PSTC* splits the verse in macro-area of type of speech, and it identifies the positions of the main words: subject, verb, object, etc. It removes the identified words, substituting them with place holders with the meter of each word. The result of this operation is a poetic sentence template. For example, if the administrator accepts the following verse:

- *Shimbiraha gabraartiyo hadday, galow wax weydiiso*

The names (*shimbiraha, galow*) and the verbs (*gabraartiyo, weydiiso*) in the verse are substituted with the following place holders: CN-P-4 (common name, plural, counting 4 *morae*); V-5 (verb counting 5 *morae*); SN-S-2 (specific name, singular, counting 2 *morae*); V-4 (verb counting 4 *morae*).

CN-P-4 V-5 hadday, SN-S-2 wax V-4

This template is added into the poetic verse templates database for the future use of the composition phase of the poetry construction process (*see* paragraph 3.)

2.6. Parser for Somali Poetry (PSP)

PSP is a parser module to analyse and check the correctness of the poetry metric system. The scansion system used by this module is simplified. It has two methods of checking: the first method is *"simulation"* of existing verse definition (this is what we call in Somali *gabay dheeg)* and the second is building the verse according to the type of *maanso* chosen by the user.

2.6.1. *First method: gabay dheeg - simulation of alliteration and structure of an existing poem.*

- Step 1: The *PsP* different ways of writing PSP or PsP gets as input a reference verse from which to make a 'copy'.
- Step 2: Analyses the structure of the verse: *hojis*, vs. *hooris* in the case of Gabay; *Mooro* in case of *jiifto*, etc. identifying subject, verb, etc.
- Step 3: For each word in the original verse
  - ✓ Analysis grammaticality the word and finds its root (in case of verb or noun)
  - ✓ Removes all consonants from the original word (n.b. a new development of the Somali metric studies on the role of the consonants as virtual geminates [Orwin and Gaarriye, 2010] are to be considered here as well).
  - ✓ Searches all words from the database that are starting with the same character as the original word, and that are having the same vowel disposition
  - ✓ From the Knowledge database, searches best match according to the "similar words used in a discourse" (for instance, it looks at all words that have a distance from the word of less than 3 words in a text, and chooses those who match in my subset of words so far collected).
  - ✓ Builds the verse with the new words found
- Step 4: If the so built verse has been used in the past in an existing poem, the programme disqualifies it and changes the words so far selected and starts the loop again.

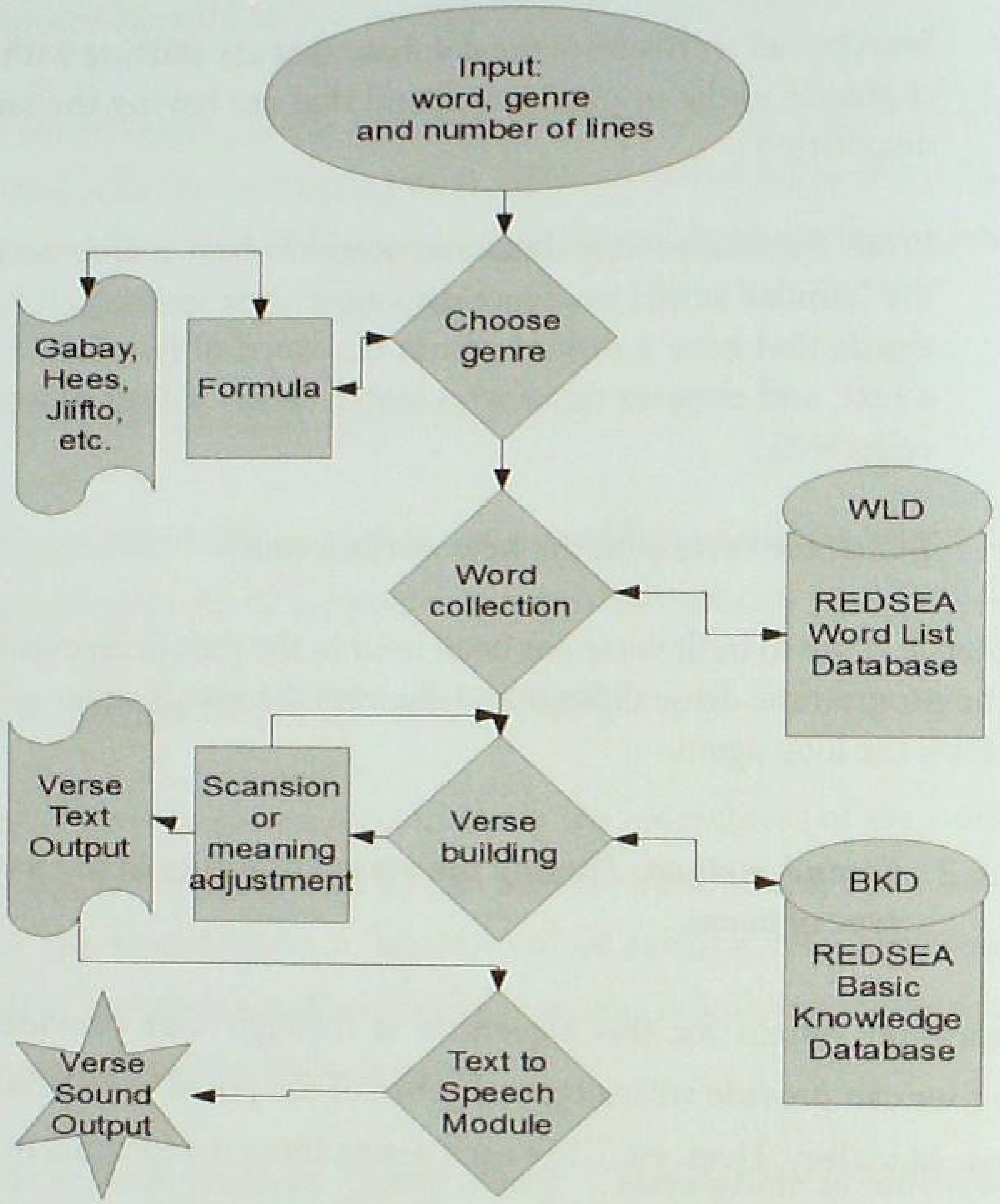
2.6.2. *Second method: Parsing the verse according to the user defined type of poem.*

Another scansion system for this algorithm is through user provided type of genre. The user can provide with a type of genre of the poem to be produced (i.e. Jiifto, Gabay, Maqaleey, Hees, etc.). For each genre, there is a formula to count the "mora" which is, as said, a key unit for quantitative scansion.

2.7. Waa kuma:

Waa Kuma is Text to Speech converter device to produce Somali speech from a text, using a synthesis method from an open source TTS (text-to-speech) engine. The idea (still under development) is to make adaptation of the speech synthesizer to the melodic and rhythmic formalisation of the type of genre of Somali poetry. For this study, we developed the module as it can read the Somali language in "normal" speech intonation.

3. Graphical representation of the algorithm.



Each of the listed modules in Section 2 has a specific responsibility in the process of composing a verse. This process consists of five macro-activities: identifying, segmentation, composition, parsing and generation. The identifying module chooses the alliteration consonant and the "concept" word of the poem. The segmentation module finds possible ways to partition the sentences identified within the base knowledge into dictionary entries. Composition module arranges the verse in a structurally correct sentence. The parsing process approves the correctness of the poetry metric system; and finally the generation module converts the text verse into voice. The following is a rough and simplified scheme of the algorithm for a Robot to compose a poem:

1. A user provides the Robot a letter and a subject. The letter is intended for alliteration, and the word is a keyword for the theme about which the poem is. 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 rebuilds its internal data, and makes an index, to obtain from the lemmas of base words, all words starting with 'g' (in the case of this example) 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 (i.e. antonymous and synonyms of *Gobannimo* in this case but also synonyms of other words equivalent of *"Gobannimo"* and better if they are starting with 'g');

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. This search uses balanced methods to identify words: for instance considering the distance between the words in the sentence or if the word is in the sentence as a verb or as a noun etc.

4. Each word in the identified sentences is analysed by the morphological parser module, to find the root of the word, and such root words are grouped according to their type of speech (noun, verb, etc.);

5. Using the grammar checker module on the grouped words, the robot composes new sentences. The sentence composition consists of putting the appropriate word (verb, noun, adjective …) in the right place of the sentence template;

6. Using the poetry parser, the robot composes metrically correct verses including the listed words;

7. Finally, using the text to speech converter, the robot recites a Somali poem, which will certainly be semantically correct, but some of the verses may not make sense at all.

The recited verses will also be printed as text output. In the case of the creation of nonsense sentences, the users may make the necessary fixes, adjusting some words, having at their disposal a list of words that can be used to substitute from the randomly generated words. The ability of the Robot presented in this paper, 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.

4. An example: gabay dheeg

A simulation has been tested of the following verses from "Gudgude" of Mohamed Ibrahim Warsame "Hadraawi".

✓ *Gedgeddoonka maankiyo qalbiga, gocasho soo boodka,*
✓ *Xaajada garlayliga jartiyo, gun u saluuggayga,*
✓ *Gibladiyo mashxaraddayda iyo, farax la goohayga.*

The produced verses does not fit well in any meaning in any context, but every single verse makes sense in Somali, and it is quite correct in the metric scansion system of the Gabay genre.

✓ *Gefkuddoonka maalkiyo qabsiga, gorayo soo haadka,*
✓ *Xaabada dabli'idiyo jabkiyo, dar u xanuunkayga,*
✓ *Gundhadiyo marxaladdayda iyo faham la' giidayda.*

5. Conclusion

Whether a machine can act intelligently, or solve a problem that a person would solve by thinking, are the questions to which the philosophy of artificial intelligence attempted to give an answer to since Alan Turing's provocative question "Can machines think?" [Turing, 1950]. Verse composing requires first of all thinking, and also emotion, intelligence, consciousness, intense feeling and a way to shape articulated form of precise representation that we get only in human languages, thinking and arts production. Computers possess none of these abilities; therefore they cannot produce the art of poetry. However, in today's advanced technology in simulation, and the very efficient procedures to store, organize and retrieve knowledge in real time, computers can represent a great support in poetry composition. They can simulate composing semantically correct verses in Somali, which also make sense in the spoken language.

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 are urgently needed and of which this algorithm could benefit, is the formalization of a morphological grammar. Tools for analytical research on Somali poetry can be a good research field for Somali literature students at university level.

References

- Andrzejewski B. W., 1964.*The Declensions of Somali Nouns*, School of Oriental and African Studies, University of London, London.
- Andrzejewski B. W., 1979. *The Case System in Somali*, School of Oriental and African Studies, University of London, London.
- Andrzejewski B. W., 1982. Alliteration and scansion in Somali oral poetry and its cultural correlates, Journal of the Anthropological Society of Oxford, 13/1, 68–83.
- 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, pp. 43-95, Roma.
- Banti G., Giannattasio F. 1996. *Music and metre in Somali poetry*, African Languages and Cultures. Supplement.
- Cabdillaahi Diiriye Guuleed 'Carraale', 2004. *Miisaanka Maansada Soomaaliyeed*, Abokers Förlag.
- Cabdillaahi Diiriye Guuleed 'Carraale', 1978. *Articles appeared bewteen January and March on Xiddigta Oktoober daily news paper*, Mogadishu.
- Cabdiraxmaan C Faarax 'Barwaaqo' and Xuseen Yuusuf Liibaan, *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).

- Farah Ahmed Ali (Gaamuute), 2012. *Coming of age: An Introduction to Somali Metrics,* Ponte Invisbile (redsea-online), Pisa.

- Hartman C., 1996. *Virtual Muse: Experiments in Computer Poetry*, Wesleyan University Press.

- Jama Musse Jama, 2009. *Somali language and its standardization as a written language: description and computer tools* in *Studi somali 13 - lessons in survival : the language and culture of Somalia, thirty years of Somali studies,* Annarita Puglielli (Ed), Roma.

- Martin Orwin and Mohamed Hashi Dhamac "Gaarriye", *Virtual geminates in the metre of Somali poetry*, in *Milk and Peace, Drought and War: Somali Culture, Society and Politics*, Markus Hoehne, Virginia Luling (Eds), London,

- Maxamed Xaashi Dhamac "Gaarriye", 2010. *Articles appeared bewteen January and March on Xiddigta Oktoober daily news paper*, Mogadishu, 1976.

- Nell S. D., *Toward a theory of rhythm in french poetry: Computer assisted recognition of rhythmic groups in traditional isometrical alexandrines*, in Computer and Humanities, Vol 27; pp 185-223; Springer Netherlands;

- Rabi H. M. (Maxamed Xaaji Raabbi), 2003. *Somali Syntax: Some Common Features*, in War Destroys – Peace Natures, Somali Reconciliation and Development, Ford, R. B.; Ismail, E. A. and Adam, H. M. (Ed.), selected papers from the 8th Somali Studies International Association Congress, pp 41-58, The Red Sea Press, Lawrenceville, NJ; 2004.

- Saeed J. I., 1993. Somali Reference Grammar. Dunwood Press, Kensington.

- SSEM, *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]

- Turing, Alan (October 1950), *"Computing Machinery and Intelligence"*, Mind LIX (236): 433–460.

- Yaasiin Cismaan Keenadiid . 1976. Qaamuuska Af-Soomaaliga, Firenze.