Technical Reports on Language Technology
Report No 64, 20020
http://www.njas.helsinki.fi/salama

Language learning system with unconstrained vocabulary[1]

Arvi Hurskainen
Department of World Cultures, Box 59
FIN-00014 University of Helsinki, Finland
arvi.hurskainen@helsinki.fi

Abstract

In 2009 I described a computerised language learning system for Swahili, using the Xerox FST package. In that system, there was no specific disambiguation component included, and the disambiguation was performed using pattern matching rules for each type of noun phrase. As a result, the analysis results were often so massive that it was not possible to print the whole result into the report.

Here I present a more sophisticated version of the language learning system, where a specific disambiguation module is included as part of the analysis system itself. I will also introduce the adaptation of the system as a web-based interactive learning system and demonstrate in detail its operation.

Key Words: *language learning, CALL, computer-assisted language learning.*

1 Introduction

Learning Bantu languages, as well as other class languages, requires the command of the class system and concord system of the language. It is not sufficient to know the particular class of a noun. One must also know precisely how each of its modifiers is formed, and in which order they come in relation to the noun. In some classes the rules are clear, but there are several cases where the correct form cannot be derived on the basis of basic rules. Especially in the cases, where the form of the noun and its semantic content clash, it is not clear, which of the modifiers follow the formal rule, and which follow the inflection required by the semantic content. There are also numerals and adjectives that do not inflect. In case of adjectives especially, it is difficult to remember whether the adjective inflects or not.

A noun may have several meanings. If one meaning or more of the noun falls into a different semantic category than the other meaning(s), they trigger separate inflection patterns. A typical case is that one meaning is animate and other meanings are inanimate. In some cases, the noun belongs to three semantic categories, each of which requires its own inflection pattern. It is quite complicated to construct such a training system that handles also such cases. This can be done, however. It requires a disambiguation system that can resolve such problems.

The strength of the learning system discussed here is that it includes all the vocabulary of Swahili, about 39,000 words. Only basic words are included; compounds and multiword-expressions are excluded as irrelevant in this application. The learner can train with any of these words and with any of their forms.

The learning system is interactive. When the learner enters a phrase, or even a single word, the system gives feedback. It checks three things, (a) typing errors, (b) correct word order, (b) concordance of the phrase, in this order.

The system inspects first whether each word is a Swahili word. If not, the incorrect word is marked and the response asks to correct the word.

Next the system checks the word order. If there is a problem in word order, the response notes it and asks to correct the order. If the word order is correct, no response is given and the process goes to the next phase.

Finally, the system checks whether each member of the noun phrase has the form required by the rules. This is the heaviest part of processing, because each noun class, including cases with deviant inflection, must be handled separately.

If the phrase is correct, the response states that the word order and concordance are correct. In case the word order is correct, but the concordance is not, this will be reported. In case the concordance would seem to be correct, but the word order is not, the process does not come this far. It stops in the phase where word order is checked and a note on wrong word order is given.

The learning system is interactive in all forms of learning. It tries to guess what the learner intends to do. Especially mistakes in word order trigger many kinds of comments for assisting the learner to correct the phrase. In the case of word order, such comments are easy to implement. In case of concordance, the implementation is more difficult, because there is a huge number of possibilities for constructing the wrong concordance. In spite of these difficulties, I have implemented a system, where the check of concordance proceeds word by word. The system tells in which word precisely the concordance error is. When the first concordance error is corrected, the system looks further and if a new concordance mistake is found, the system tells in which word the mistake is. So the process continues until the whole phrase has the correct concordance. I think that this advice system is much more helpful than the sheer statement that the concordance is wrong.

This report is an extension to the earlier report (2009)[2] on the computer-assisted language learning system

The rest of this report is divided into the following sections: (1) description of the learning system components, (2) description of various learning methods, and (3) demonstrating the learning system in SALAMA web-based environment.

2 Implementation of the learning system

The core of the learning system is the language analyser, which first analyses each word and gives it all linguistically legitimate interpretations. Context is not concerned in this

[2] http://www.njas.helsinki.fi/salama/language-learning-system.pdf

phase. The analyser was implemented with the two-level morphology system developed by Koskenniemi .[3] Example of the analysis result is in (1).

(1)
```
"<mtoto>"
      "mtoto" N 1/2-SG HUM { the } { child , :young person ,
juvenile }
"<wangu>"
      "angu" PRON POSS 3-SG SG1 { my , mine }
      "angu" PRON POSS 3-SG SG1 <PRE { me }
      "angu" PRON POSS 11-SG SG1 { my , mine }
      "angu" PRON POSS 11-SG SG1 <PRE { me }
      "angu" PRON POSS 1-SG SG1 { my , mine }
      "angu" PRON POSS 1-SG SG1 <PRE { me }
      "angu" PRON POSS 2-PL SG1 { my , mine }
      "angu" PRON POSS 2-PL SG1 <PRE { me }
"<mzuri>"
      "zuri" ADJ A-INFL 1-SG { good , beautiful , pretty }
      "zuri" ADJ A-INFL 3-SG { good , beautiful , pretty }
      "zuri" ADJ A-INFL 11-SG { good , beautiful , pretty }
      "zuri" ADJ A-INFL 18-SG LOC { good , beautiful , pretty }
"<anasoma>"
      "soma" V 1-SG3-SP VFIN { he } PR:na z [soma] { read , study
, educate } SVO SVOO
      "soma" V 1-SG3-SP VFIN { she } PR:na z [soma] { read , study
, educate } SVO SVOO
"<$>"
      "$" DOLLAR-SIGN { $ }
```

The word *mtoto* has only one interpretation, but the words *wangu* and *mzuri* have several. Without going into details of the meaning of each tag, we go to the next phase, where we disambiguate the analysis result. The disambiguation is performed using the Constraint Grammar rules. Note that in addition to disambiguation, also syntactic tags are included. The disambiguated text is in (2).

(2)
```
"<mtoto>"
      "mtoto" N 1/2-SG HUM { the } { child } @<P
"<wangu>"
      "angu" PRON POSS 1-SG SG1 { my } @GCON
"<mzuri>"
      "zuri" ADJ A-INFL 1-SG { good } @<NADJ
"<anasoma>"
```

[3] Koskenniemi, K. 1983. Two-level Morphology: A General Computational Model for Word-Form Recognition and Production. Department of General Linguistics. University of Helsinki. Publication No. 11.

```
      "soma" V 1-SG3-SP VFIN { he } PR:na z [soma] { study } SVO
SVOO @FMAINVtr-OBJ>
"<$>"
      "$" DOLLAR-SIGN { $ }
```

Each word has now only one interpretation. Now we put the whole phrase to one line so that first comes the whole surface phrase, as we typed it, and then the analysis of each word, glued together with a plus sign. At the end of the string is a sign '&' to help in marking the end of line (3).

(3)

```
 mtoto wangu mzuri anasoma  +N+1/2-SG+HUM{child} +POSS+1-SG+SG1{my}
+A-INFL+1-SG{good} +V+1-SG3-SP+VFIN{he}+PR:na[soma]{study} &
```

In the string in (3) above, the analysis of each word starts with its part-of-speech (POS) code. This is important information for judging whether the word order is correct. On the basis of the analysis sequence in (3), we can see whether the word order is correct or not. If it is correct as in this case, we mark the whole string with a code, which contains information on its correctness (4).

(4)

```
 mtoto wangu mzuri anasoma  +N+1/2-SG+HUM{child} +POSS+1-SG+SG1{my}
+A-INFL+1-SG{good} +V+1-SG3-SP+VFIN{he}+PR:na[soma]{study}
N+POSS+A+V_WO
```

At the end of the string in (4) there is the code N+POSS+A+V_WO. This means that the phrase contains the following word categories: a noun, followed by a possessive pronoun, followed by an adjective, followed by a verb. This is a grammatical word order, and it is marked with the code _WO.

What about if the word order is not grammatical? In (5) we have such a case.

(5)

```
mtoto mzuri wangu anasoma  +N+1/2-SG+HUM{child} +A-INFL+1-SG{good}
+POSS+3-SG+SG1{my} +V+1-SG3-SP+VFIN{he}+PR:na[soma]{study}
N+A+POSS+V_!WO
```

The correctness code in (5) is _!WO, which means that the word order is not correct. This code prompts a further check. In giving information to the learner, we need only to look at the word order code and then give appropriate guidance. If there is the code _WO, the learner can be informed that the word order is correct. If there is the code _!WO, the learner can be informed that word order should be checked.

Next we check whether the concordance is correct (6).

(6)

```
 mtoto wangu mzuri anasoma  +N+1/2-SG+HUM{child} +POSS+1-SG+SG1{my}
+A-INFL+1-SG{good} +V+1-SG3-SP+VFIN{he}+PR:na[soma]{study}
N+POSS+A+V_WO CONC4
```

The system added the tag CONC4 to the end of reading. The tag means that the concordance is correct, and that the phrase has four words.

Let us see what happens when the concordance is not correct (7).

(7)
mtoto wangu wazuri anasoma +N+1/2-SG+HUM{child} +POSS+2-PL+SG1{my} +A-INFL+2-PL{good} +V+1-SG3-SP+VFIN{he}+PR:na[soma]{study} N+POSS+A+V_WO

The system did not produce the concordance code at all. On the basis of this missing information we can advise the learner about concordance problems.

Now we come to the stage, where the system reports on the phrase and gives guidance if needed (8).

(8)
Word order and concordance are correct!

mtoto wangu mzuri anasoma
+N+1/2-SG+HUM{child} +POSS+1-SG+SG1{my} +A-INFL+1-SG{good} +V+1-SG3-SP+VFIN{he}+PR:na[soma]{study}

Note that the result was made more readable, and the word order code plus concordance code at the end of reading were removed as redundant.

3 Description of learning methods

The technical solution described above offers many kinds of learning applications. The strength of the system is that it operates using the linguistic information in making decisions and in giving guidance. Any noun of the given noun class serves as the training example. Also any adjective, pronoun, numeral or verb can be used in training.

The free learning method is assumably the most important method of using the system. It is probably not the easiest method, because the learner is not given precise tasks to do. The learner needs to collect the vocabulary for using in training. The task is not easy, because there are 39,000 words to choose from.

The ad hoc approach might work. The learner uses just the words and phrases that happen to come in mind. From experience we know that this method puts excessive stress to the leaner's imagination, and this is not the purpose of training.

A workable method would be to construct frequency lists for each class of nouns, for adjectives and for verbs. The learner could start with most common words and proceed to using less common ones. The lists would be bilingual, Swahili and English. This would help the student to concentrate on learning instead of trying to remember words. Also alphabetical lists to both directions would help in finding desired words.

Below I will give a number of examples of this free learning method. I also use examples of all noun classes, including such nouns that use exceptional concordance.

Let us start with a phrase with a typing error (9). The examples are from an intermediate phase of processing.

(9)
mtoto ?wagu? mzuri anasoma `+N+1/2-SG+HUM{child} Heur +A-INFL+1-SG{good} +V+1-SG3-SP+VFIN{he}+PR:na[soma]{study} HEUR2`

The faulty word is surrounded with question marks, and the analysis of that word is Heur. In the end of the phrase there is a code HEUR2 showing that there is a typing error in the second word.

We may take an example with two typing errors (10).

(10)
mtoto ?wagu? ?mzur? anasoma `+N+1/2-SG+HUM{child} Heur Heur +V+1-SG3-SP+VFIN{he}+PR:na[soma]{study} HEUR23`

Both faulty words are marked and their analysis is Heur. The code HEUR23 in the end means that the second and third word should be corrected.

We may take an even more faulty phrase (11).

(11)
?mtoo? ?wagu? ?mzur? anasoma `Heur Heur Heur +V+1-SG3-SP+VFIN{he}+PR:na[soma]{study} HEUR123`

All three faulty words are marked and their interpretation is Heur. The code HEUR123 at the end tells that three words in the beginning should be corrected.

Now we go to the final phase of processing (12).

(12)
Please check spelling! Correct the first word.

?mtot? wangu mzuri anasoma `Heur +POSS+3-SG+SG1{my} +A-INFL+11-SG{good} +V+1-SG3-SP+VFIN{he}+PR:na[soma]{study}`

The response asks to check the spelling. It also informs that one word is wrong.

When we add more mistakes, we get corresponding comments (13 and 14).

(13)
Please check spelling! Correct the first and second word.

?mtot? ?wngu? mzuri anasoma `Heur Heur +A-INFL+1-SG{good} +V+1-SG3-SP+VFIN{he}+PR:na[soma]{study}`

(14)
Please check spelling! Correct the first, second and third word.

?mtot? ?wngu? ?mzri? anasoma `Heur Heur Heur +V+1-SG3-SP+VFIN{he}+PR:na[soma]{study}`

What about if the wrongly spelled words are not consecutive? In the intermediate phase the result is as in (15).

(15)
?mtoo? wangu ?mzri? anasoma `Heur +POSS+3-SG+SG1{my} Heur +V+1-SG3-SP+VFIN{he}+PR:na[soma]{study} HEUR13`

The first and third word has something wrong. In the final phase the result looks like in (16).

(16)
Please check spelling! Correct the first and third word.

?mtoo? wangu ?mzri? anasoma `Heur +POSS+3-SG+SG1{my}`

Further examples of spelling errors are in (17-19).

(17)
Please check spelling! Correct the first, third and fourth word.

?mtoo? wangu ?mzri? ?hyu? anasoma `Heur +POSS+3-SG+SG1{my} Heur Heur +V+1-SG3-SP+VFIN{he}+PR:na[soma]{study}`

(18)
Please check spelling! Correct the first, third and fifth word.

?mtoo? wangu ?mzri? huyu ?anasom? `Heur +POSS+3-SG+SG1{my} Heur +DEM+1-SG{this} Heur`

(19)
Please check spelling! Correct the third and fifth word.

mtoto wangu ?mzri? huyu ?anasom?
`+N+1/2-SG+HUM{child} +POSS+1-SG+SG1{my} Heur +DEM+1-SG{this} Heur`

When we correct all spelling errors, we get the response as in (20).

(20)
Word order and concordance are correct!

mtoto wangu mzuri huyu anasoma
+N+1/2-SG+HUM{child} +POSS+1-SG+SG1{my} +A-INFL+1-SG{good} +DEM+1-SG{this} +V+1-SG3-SP+VFIN{he}+PR:na[soma]{study}

There is also another type of responding to spelling errors. Because each misspelled word is surrounded by question marks, no detailed information is needed in response. With this implementation, we get the following kinds or responses (21-23).

(21)
Please check spelling! Correct the word surrounded by question mark.

mtoto ?wang? mzuri huyu anasoma
+N+1/2-SG+HUM{child} Heur +A-INFL+1-SG{good} +DEM+1-SG{this} +V+1-SG3-SP+VFIN{he}+PR:na[soma]{study} HEUR1

(22)
Please check spelling! Correct the words surrounded by question mark.

mtoto ?wang? mzuri ?huy? anasoma
+N+1/2-SG+HUM{child} Heur +A-INFL+1-SG{good} Heur +V+1-SG3-SP+VFIN{he}+PR:na[soma]{study} HEUR2

(23)
Please check spelling! Correct the words surrounded by question mark.

mtoto ?wang? mzuri ?huy? ?anasom?
+N+1/2-SG+HUM{child} Heur +A-INFL+1-SG{good} Heur Heur HEUR3

Now we take an example with all three types of errors, that is, spelling, word order and concordance (24).

(24)
Please check spelling! The first and second word are not correct.

?mtot? ?mzri? wangu huu anasoma Heur Heur +POSS+3-SG+SG1{my} +DEM+3-SG{this} +V+1-SG3-SP+VFIN{he}+PR:na[soma]{study}

We see that the system asks first to correct spelling and ignores other mistakes. We correct it and the second type of error will be encountered (25). If all words are typed correctly, no comment is given.

(25)
Word order is not correct. Possessive pronoun cannot be after an adjective!

```
 mtoto mzuri wangu huu anasoma
       +N+1/2-SG+HUM{child} +A-INFL+1-SG{good} +POSS+3-SG+SG1{my}
+DEM+3-SG{this} +V+1-SG3-SP+VFIN{he}+PR:na[soma]{study}
N+A+POSS+DEM+V_!WO
```

The word order was checked and it was found wrong. The answer warns about wrong word order and tells precisely where the fault is. We correct the word order and get the result as in (26).

(26)
Word order is correct, but concordance is not!

```
 mtoto wangu mzuri huu anasoma
       +N+1/2-SG+HUM{child} +POSS+1-SG+SG1{my} +A-INFL+1-SG{good}
+DEM+3-SG{this} +V+1-SG3-SP+VFIN{he}+PR:na[soma]{study}
N+POSS+A+DEM+V_WO
```

Now the response is that the word order is correct. The concordance, however, is faulty. The response does not tell precisely what is wrong. It just tells that it is not correct.

This is the point in the learning system, where further development was needed. In fact, I implemented the precise feedback system, although it required a large number of new complex rules.

The approach in constructing the rules is that the starting point is the noun, because it defines the form of other components. The system checks first whether the second word conforms with the noun, regardless of whether it is a possessive pronoun, or adjective, or demonstrative pronoun, or numeral, of verb. If it is not correct, the system tells that the concord of the second word is not correct. When it is corrected, the system checks the third word, whatever it is, and warns if concordance is not correct. When the third word is corrected, the system looks at the fourth word, and so on, until the whole phrase is checked and corrected.

With this new implementation, the output of the example in (26) is in (27).

(27)
Word order is correct but concordance is not!
The concordance of the fourth word is not correct.

```
 mtoto wangu mzuri huu anasoma
       +N+1/2-SG+HUM{child} +POSS+1-SG+SG1{my} +A-INFL+1-SG{good}
+DEM+3-SG{this} +V+1-SG3-SP+VFIN{he}+PR:na[soma]{study}
N+POSS+A+DEM+V_WO
```

The system added a new line of warning. It warns about the concordance of the fourth word. Note that in order to get this prompt the faulty word must be a correct Swahili word

form. Otherwise, it interprets it as a typo, and the process stops already at the point, where typos are corrected.

Irregular concordance is handled in examples (28-37).

(28)
Word order and concordance are correct!

baba yangu mzuri huyu anasoma
+N+9/10-SG+HUM{father} +POSS+9-SG+SG1{my} +A-INFL+1-SG{good} +DEM+1-SG{this} +V+1-SG3-SP+VFIN{he}+PR:na[soma]{study}

Although the noun *baba* belongs to the class of humans, which should belong to class 1/2, this noun formally has no class marker. Such nouns, especially if their plural has the same form, are normally located to class 9/10. However, because they mean human beings, they do not naturally belong to that class. This causes strange modifications in concordance, so that the noun and the possessive pronoun (as well as the genitive connector *-a*) inflect according to class 9/10, and the other members according to class 1/2.

We may put the example into plural (29).

(29)
Word order and concordance are correct!

baba zangu wazuri hawa wanasoma
+N+9/10-PL+HUM{father} +POSS+10-PL+SG1{my} +A-INFL+2-PL{good} +DEM+2-PL{these} +V+2-PL3-SP+VFIN{they}+PR:na[soma]{study}

We might expect that the problem is solved by remembering to put the possessive pronoun into the form required by class 9/10. We are wrong. We try with a similar noun *bwana* (30).

(30)
Word order is correct, but concordance is not!
The concordance of the second word is not correct.

bwana yangu mzuri huyu anasoma
+N+9/6-SG+HUM{Mister} +POSS+9-SG+SG1{my} +A-INFL+1-SG{good} +DEM+1-SG{this} +V+1-SG3-SP+VFIN{he}+PR:na[soma]{study}
N+POSS+A+DEM+V_WO

The word *bwana* belongs to class 9/6, and at least in singular it should require the same concordance as *baba*. But it does not. The correct form is in (31).

(31)
Word order and concordance are correct!

bwana wangu mzuri huyu anasoma
```
      +N+9/6-SG+HUM{Mister} +POSS+1-SG+SG1{my} +A-INFL+1-SG{good}
+DEM+1-SG{this} +V+1-SG3-SP+VFIN{he}+PR:na[soma]{study}
```

There are also nouns with semantically very different interpretations. Training with such nouns is subject to irresolvable ambiguities. Consider the examples in (32 and 33).

(32)
Word order and concordance are correct.

kiboko wangu
```
      +N+7/8-SG+AN{hippopotamus}    +POSS+1-SG+SG1{my}    N+POSS_WO
CONC2
```

(33)
Word order and concordance are correct.

kiboko changu
```
      +N+7/8-SG{whip} +POSS+7-SG+SG1{my} N+POSS_WO CONC2
```

In example (32) the noun *kiboko* means hippopotamus, and in (33) it means a whip. (Hippo skins were used for preparing whips for controlling slaves in slave caravans.) The disambiguation system selected the corresponding alternative on the basis of the form of the possessive pronoun. Therefore, both forms are correct.

We can add more modifiers to the noun (34). The string is not correct.

(34)
Word order is correct, but concordance is not!
The concordance of the second word is not correct.

kiboko changu mkubwa yule anakula
```
      +N+7/8-SG+AN{hippopotamus} +POSS+7-SG+SG1{my} +A-INFL+1-
SG{big} +DEM+1-SG{that} +V+1-SG3-
SP+VFIN+HUM{he}+PR:na+INFMARK[la]{eat}-ACT+MONOSLB N+POSS+A+DEM+
```

The third, fourth and fifth word would agree with the noun in the meaning of hippo. The second word, however, agrees with *kiboko* in the meaning of whip. The first interpretation is followed, and the system asks to correct the second word. After correction we get the result as in (35).

(35)
Word order and concordance are correct.

kiboko wangu mkubwa yule anakula

```
      +N+7/8-SG+AN{hippopotamus} +POSS+1-SG+SG1{my} +A-INFL+1-
SG{big} +DEM+1-SG{that} +V+1-SG3-
SP+VFIN+HUM{he}+PR:na+INFMARK[la]{eat}-ACT+MONOSLB
N+POSS+A+DEM+V_WO CONC5
```

In (36) we have a different type of faulty concord of the same words.

(36)
The concordance of the fourth word is not correct.

```
 kiboko changu kikubwa ule kinaumiza
      +N+7/8-SG{whip} +POSS+7-SG+SG1{my} +A-INFL+7-SG{big} +DEM+3-
SG{that} +V+7-SG-SP+VFIN{it}+PR:na[umiza]{hurt} &
```

Here the noun *kiboko* is in the meaning of whip, and it requires the concord of class 7. The fourth word needs to be corrected (37).

(37)
Word order and concordance are correct.

```
 kiboko changu kikubwa kile kinaumiza
      +N+7/8-SG{whip} +POSS+7-SG+SG1{my} +A-INFL+7-SG{big} +DEM+7-
SG{that} +V+7-SG-SP+VFIN{it}+PR:na[umiza]{hurt} N+POSS+A+DEM+V_WO
CONC5
```

4 Language learning environment

Computer-assisted language learning needs a web-based interface, where the learner can type words and the system gives needed feedback and advise. Such an interface was constructed into the SALAMA system, and below, using screenshots, I will demonstrate its learning possibilities.

4.1 Guided learning

There is an implementation of a guided tour to all noun classes and their concordance using various combinations of words. In all, there are 37 learning modules for verbs, each with up to 13 tasks. In addition, there are three modules for learning greetings. The image of the learning environment is in Figure 1.

Figure 1. The opening view of SALAMA interactive language learning interface.

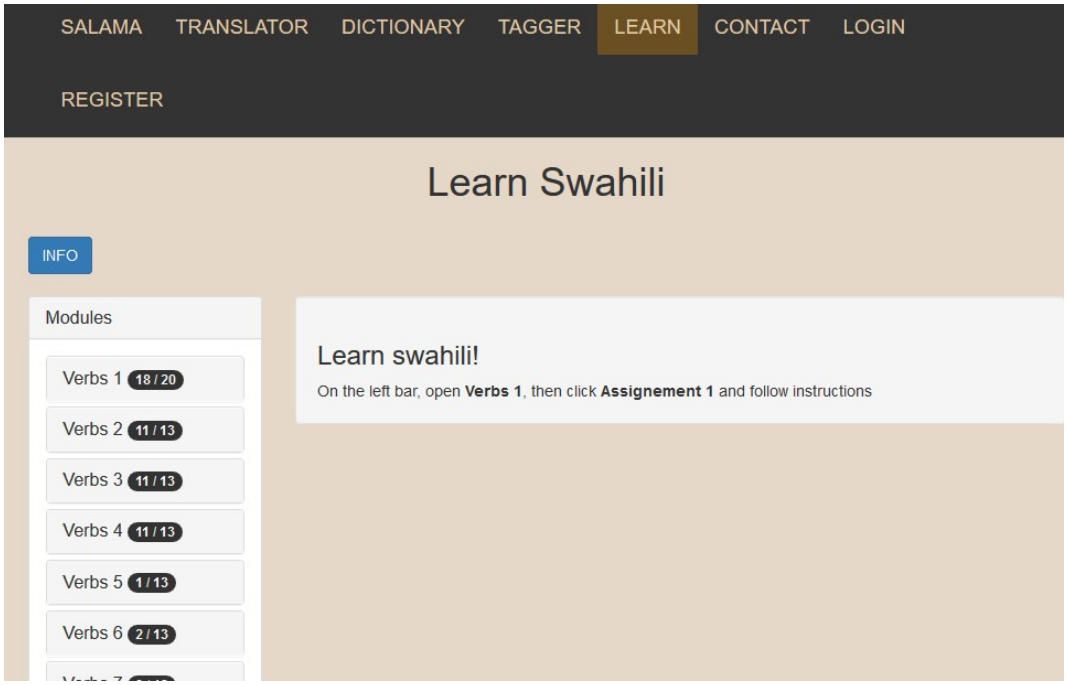


Figure 2. The first assignment in the learning system.

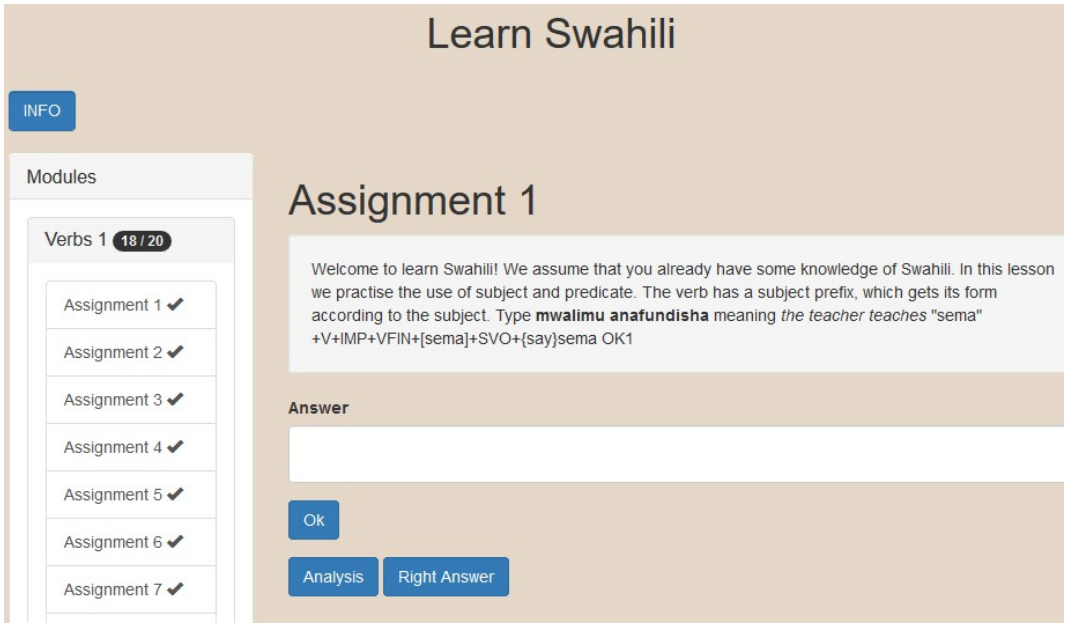
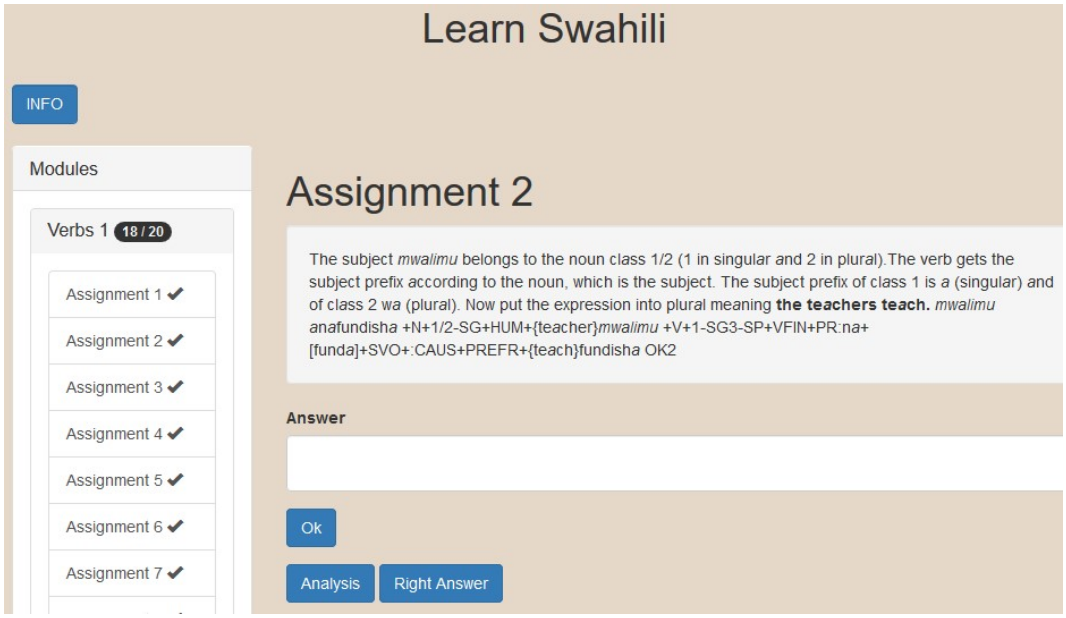


Figure 3. The correct answer prompts another assignment.



4.2 Free learning

A vastly more useful learning environment is the one, where one has limitless possibilities for learning correct structures using the whole vocabulary of the language. This environment is temporarily under Tagger with the name swa-learn. This opens up the environment, where noun phrases, including the verb, can be typed, and the result comes to another window. Examples of its use are in Figures 4-8.

Figure 4. The phrase is correct.

Figure 5. The phrase has a typing error.

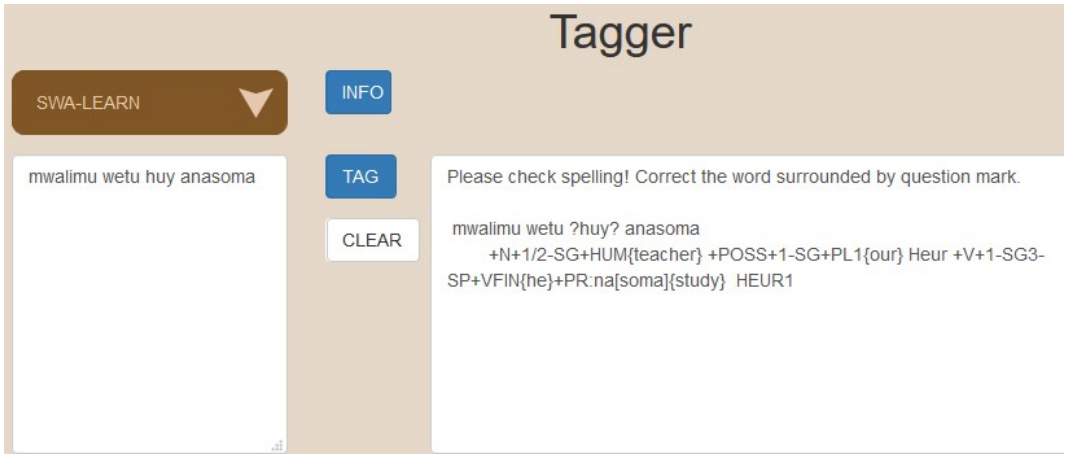


Figure 6. The phrase has a concord error.

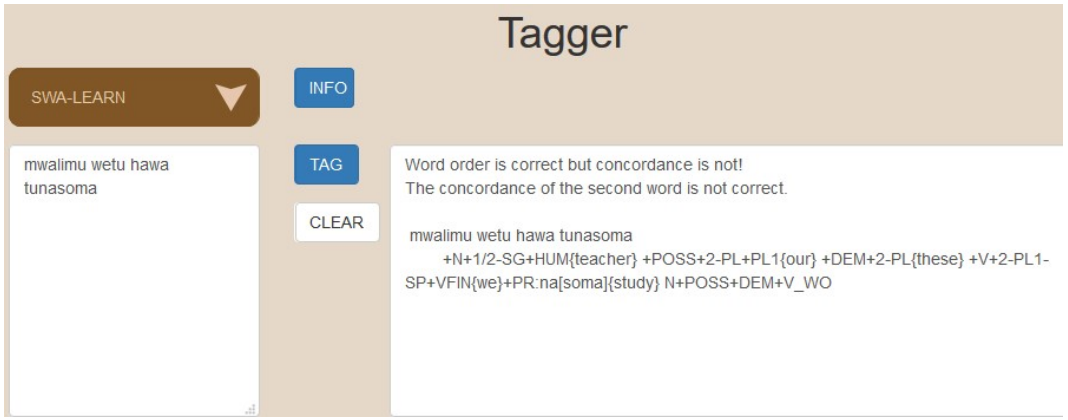


Figure 7. The phrase has another concord error.

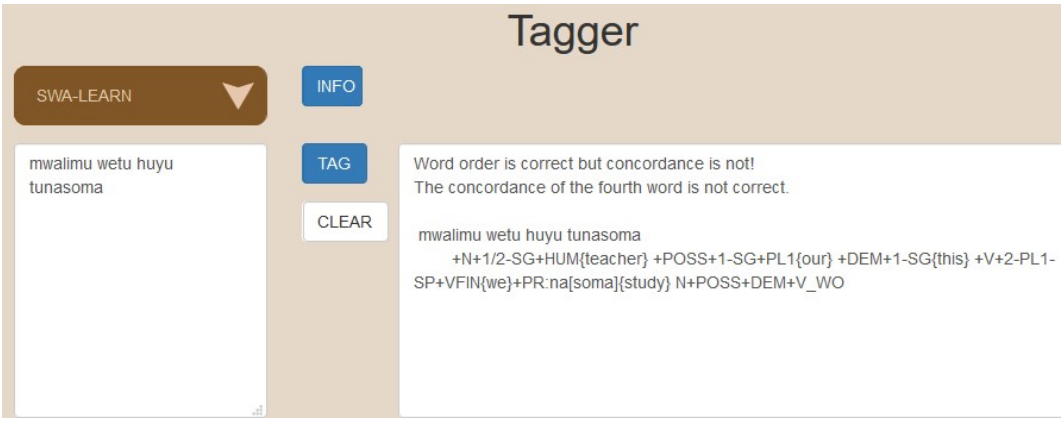
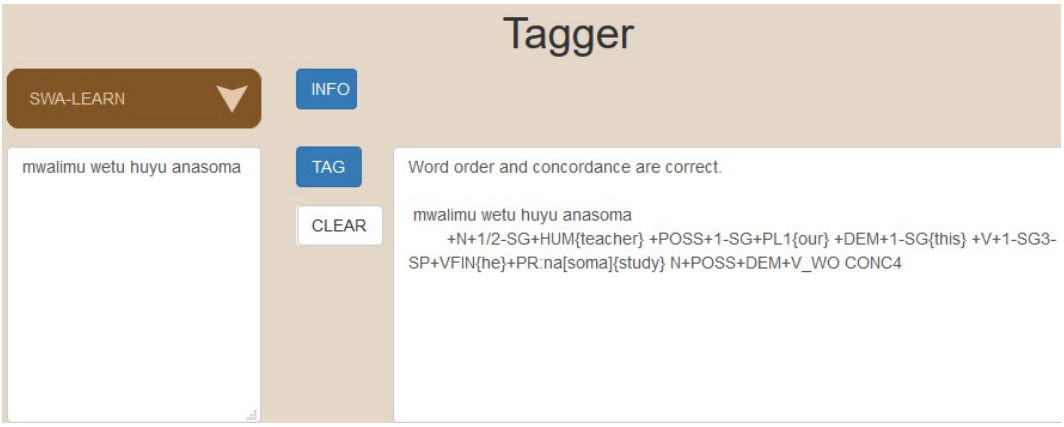


Figure 8. The phrase is now correct.



5 Conclusion

In this report I have described the current state of the computer-assisted language learning system of Swahili. There are several improvements to the earlier system described in the earlier report in 2009. There is now also a web based interactive learning environment implemented in SALAMA. The learning system is currently restricted to learning all kinds of noun phrases, including the verb. There is also a large number of guided learning modules for learning noun phrases of all noun classes, and also three modules for learning greetings.