Technical Reports on Language Technology
Report No 82, 2022
http://www.njas.helsinki.fi/salama

Rule-based language technology and self-tutored language learning systems[1]

Arvi Hurskainen
Department of Languages
FIN-00014 University of Helsinki, Finland
arvi.hurskainen@helsinki.fi

Abstract

Web-based language learning has become increasingly popular, especially during the Covid pandemic, when contact teaching was very difficult. Artificial learning environments can be arranged via the web, but the creation of natural and interactive learning environments is challenging. Computer-assisted language learning (CALL) is a field, where web-based learning methods have been developed for decades. So far, the possibilities of the rule-based language technology (RBLT) for facilitating the development of self-tutored interactive learning systems have not been extensively studied.

The benefits of the RBLT in developing learning applications depend very much on the language type. For example, English is the type of language, where RBLT is not very helpful, because its morphology is very simple and it has no concordance rules. However, the large majority of world's languages have such features, which make them ideal for developing self-tutored language learning systems.

In this report I discuss various possibilities for developing language learning applications for languages with rich morphology and a concordance system. The example language is Swahili.

Key Words: *morphological analysis, disambiguation, language learning.*

1 Introduction

The Covid pandemic forced many universities and other learning institutions to find solutions for arranging learning environments through the web. Although such learning environments lose many benefits of contact teaching, they have also such benefits that contact teaching does not have.

One of the benefits is that learning does not require traveling to the learning place, and learning can be arranged using all those possibilities that web-based teaching can offer. These include the almost unlimited number of participants in learning sessions, and the possibility to listen to recorded lessons any time that is suitable to the learner.

The web-based learning can be further extended by developing such learning environments, which do not require a teacher at all. A number of such learning

[1] The report is issued under licence CC BY-NC

environments have been on the market for years. However, they are typically restricted to a set of tasks, guided strictly by the designer of the system. Also, the vocabulary used in learning tasks is very restricted.

Rule-based language technology provided means for developing such learning systems, where the full vocabulary of the language can be used. The system analyses each entered word and disambiguates it if the entry is a phrase, consisting of more than one word. The correctness of disambiguation depends on how much environment there is available. That is, the long sequences are more precisely disambiguated than the short ones.

The learning system does not only check the correctness of the phrase. It advises the learner on various types of mistakes. It controls the correct word order in the phrase. It also checks whether each word is correct Swahili and points out the mistakes. The more challenging feature is the control of concordance, which in languages such as Swahili can be very tricky, due to its complex noun class system.

In case the entered phrase has different kinds of mistakes, the system advises the learner to correct them one by one in a certain order. If the learner follows instructions, finally the phrase is correct.

The system also outputs a detailed analysis of each word-form, including the glosses in English.

Although the learning system that allows for unrestricted vocabulary sounds promising, it has its own disadvantages. One of them is that it is too free, too open, offering almost limitless possibilities. Learning loses a systematic method of how to proceed. An open system requires that the learner creates one's own learning procedure.

A guided learning system offers a route for going systematically through all structures that need to be learned. It is possible to construct also such guided learning systems. In fact, I have constructed a guided tour through all significant phrase structures in all noun classes. The learner can try to construct correct structures, and assistance is given when needed. The guided tour consists of 43 learning modules, with several individual tasks each.

One further method for helping the learner to find the words to practise with is to use the frequency lists of each part of speech. The learner can start with the most common words in each category and proceed towards less common ones. Using this method, the learner learns the correct structures, and at the same time learns to use new vocabulary.

The best procedure in learning is to first go through the guided tour through all the important structures. This gives experience on how the system operates. After the initial introduction it is easier to start constructing phrases independently.

I have been developing the language learning system described here for many years. These developments were described in Technical Report No 3[2] and 9[3]. The current description describes its use through a browser application. In addition, this report describes also other types of language learning aids.

[2] http://www.njas.helsinki.fi/salama/language-learning-system.pdf
[3] http://www.njas.helsinki.fi/salama/language-learning3.pdf

2 Demonstration on how the language learning system works

Using a private server, below I will demonstrate how the Swahili language learning system works. The SALAMA opening web page is in (1)

(1)

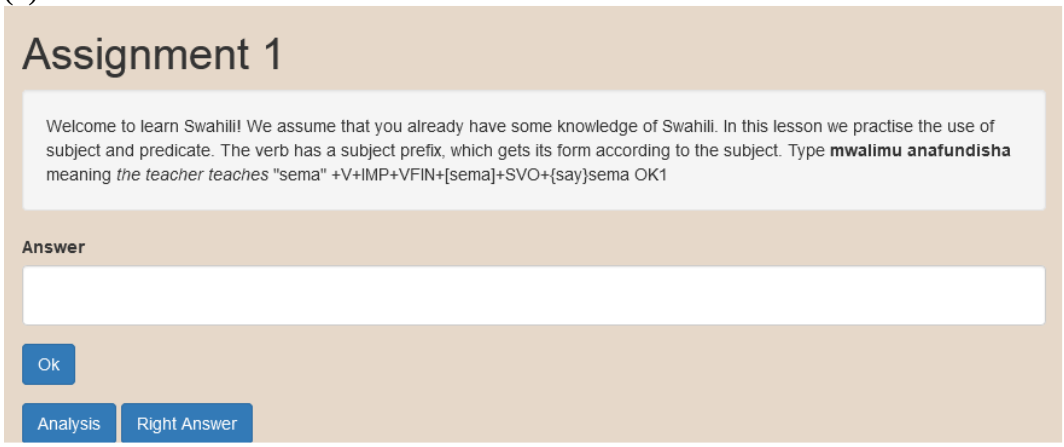
The web page has different functions, such as Translator, Dictionary, Tagger, and Learn. We are here interested in what is under Learn. When we open it, we get the view as in (2).

(2)

The window above starts the guided tour through the main phrase structures of Swahili. The phrases typically consist of the main verb and its subject with all possible types of modifiers. The whole series of phrase types has 40 modules. In addition, there are three modules for practising greetings.

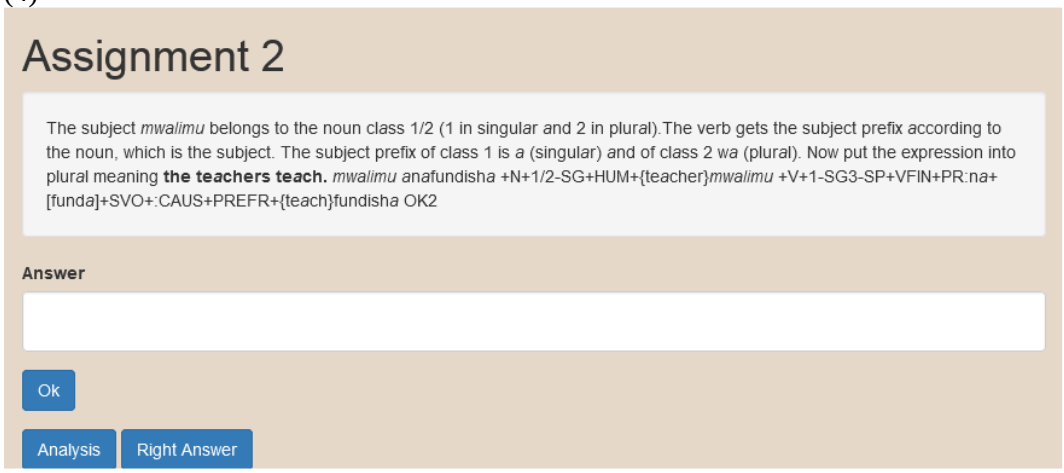
The opening window of the module Verbs 1 is in (3).

(3)



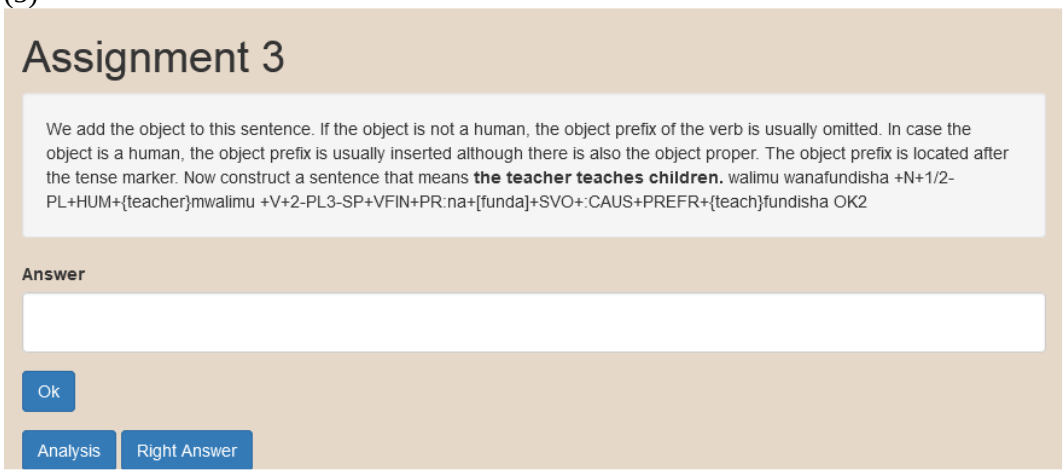
The learner is advised to type the Swahili words *mwalimu anafundisha*. The answer is in (4).

(4)



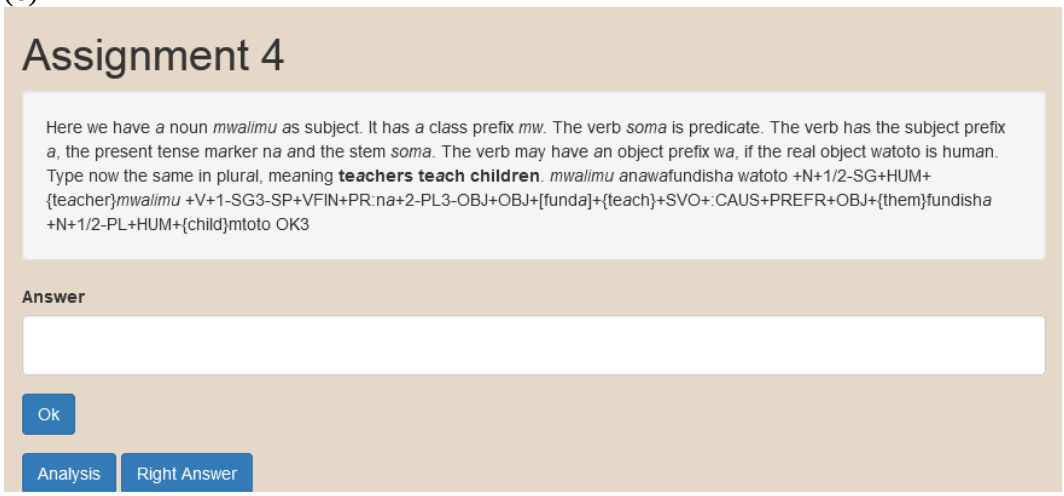
If the reply is correct, the system moves to the next assignment. Now the system asks to translate the English phrase into Swahili. When we type *walimu wanafundisha*, we get the response as in (5)

(5)



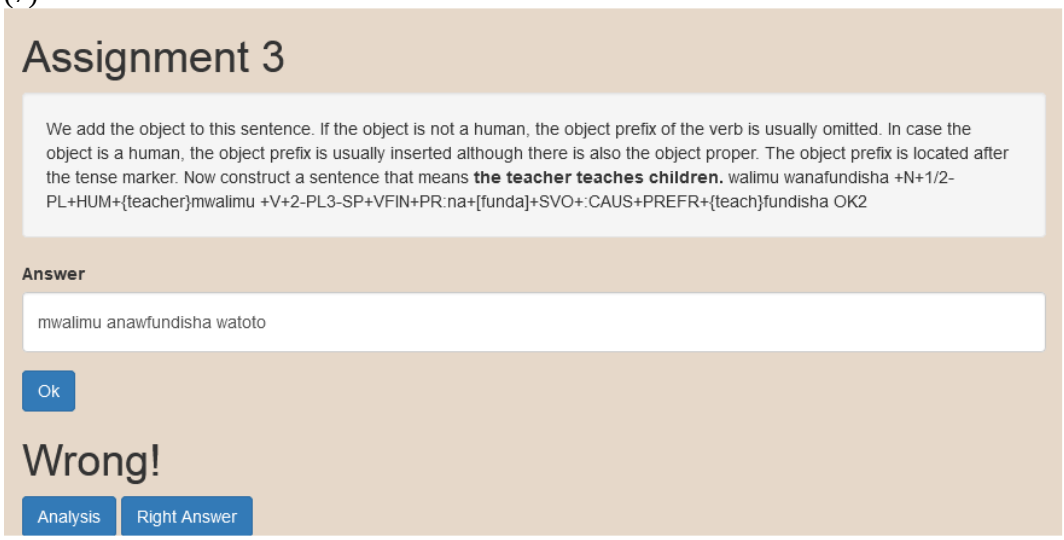
Now the system asks to add an object to the phrase. We type *mwalimu anawafundisha watoto* (6)

(6)



The reply is correct again, and the system moves to the next assignment.
What about if we make a mistake? (7)

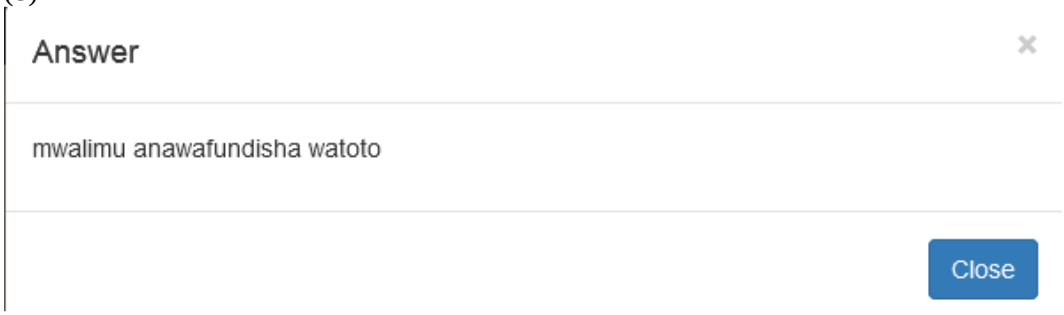
(7)



We get a warning, and we should correct the string. When the string is correct, the control moves to the following assignment.

If the learner does not find the correct form, one can check it by pressing the tab Right Answer. (8)

(8)



There are hundreds of exercises in the guided tour into Swahili grammar. They include also such structures that do not follow general rules. The weakness in this learning method is that it does not allow any deviations from the course. Each exercise must be done precisely as expected.

Such restrictions do not exist in the other learning environment, where the learner is entirely free from vocabulary and of the types of structures that the learner wants to use. Next we look into the behaviour of the free learning environment.

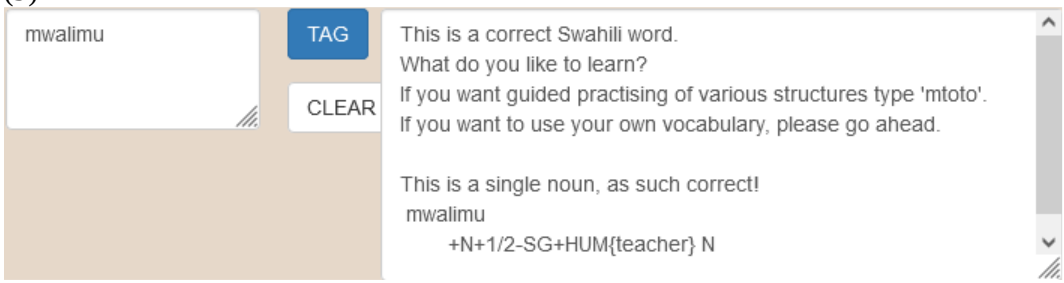
3 Free language learning environment

Because the learner in the free learning system is prone to many types of mistakes, the system provides a wide array of instructions in the course of learning process. Below we take a look at the free learning system and see how its guidance system works in each case.

On the SALAMA server, the free learning system is found under the tab Tagger, on its last row, which is SWA-LEARN.

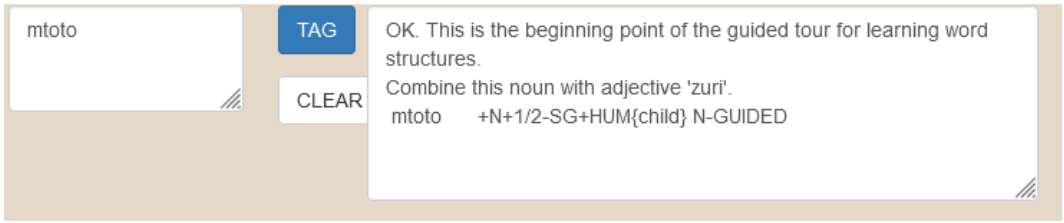
The learner can start by typing any Swahili word, such as in (9).

(9)



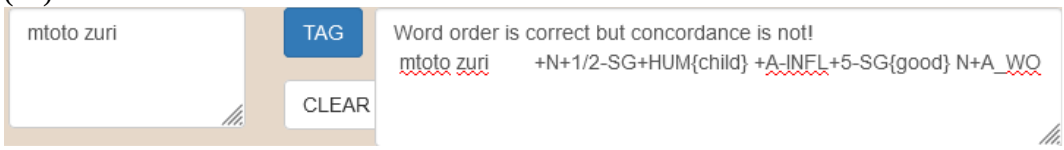
As we see above, the learner can either start guided practising or free learning. For the purpose of demonstration, we choose the first alternative and type *mtoto* (10).

(10)



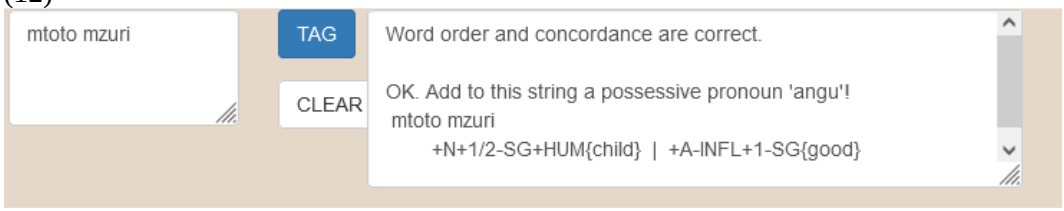
We are asked to add the word *zuri* (11).

(11)



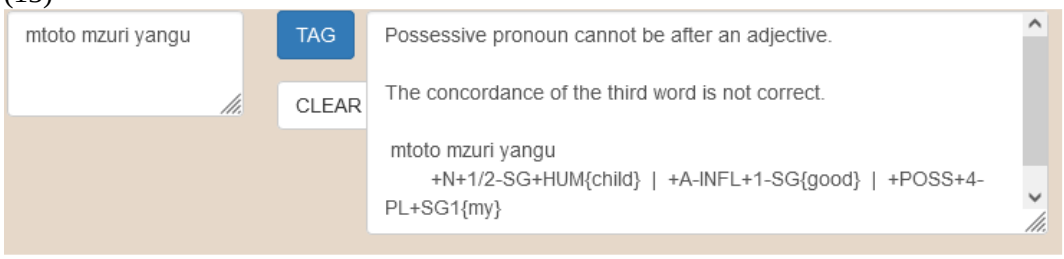
The report says that the word order is correct but concordance is not. The adjective *zuri* is of inflecting type, and it should have the correct prefix (12).

(12)



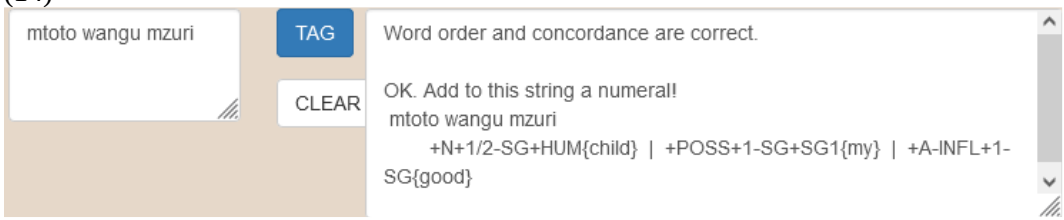
Now the string is correct, and we are asked to add the possessive pronoun *angu*. Now we must remember to put the correct prefix. We try (13).

(13)



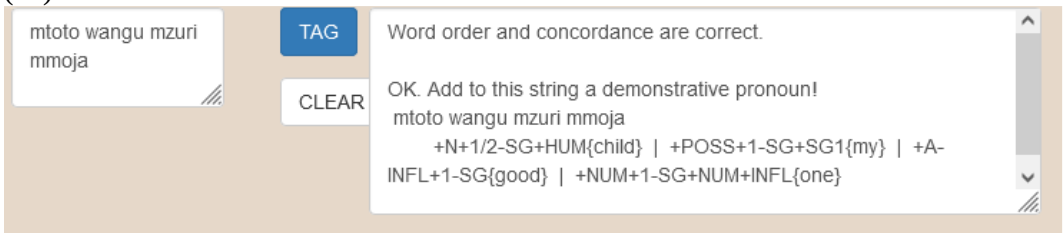
We get two warnings. Possessive pronoun cannot be after an adjective, and the concordance of the third word is not correct. We correct both mistakes (14).

(14)



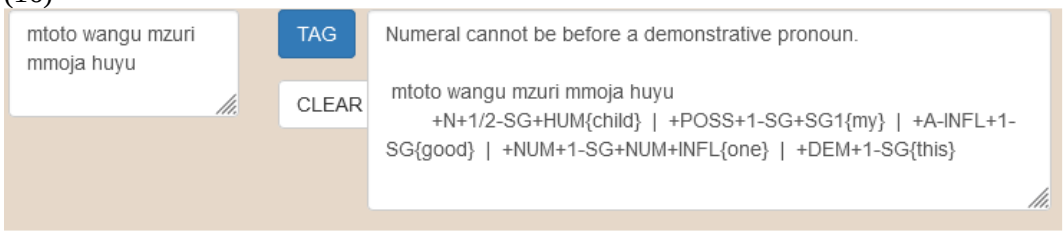
Now the string is correct, and we are asked to add a numeral (15).

(15)



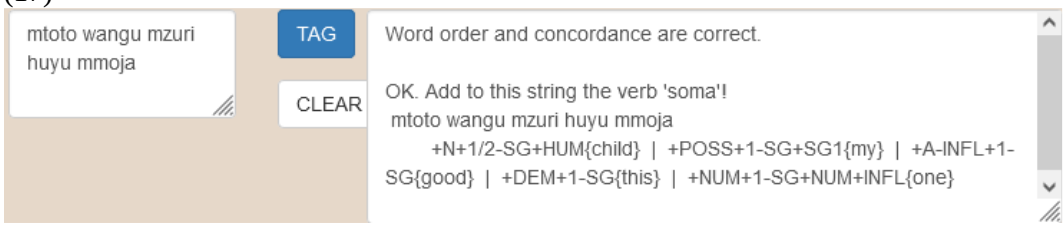
We succeeded with first try. Then we should add a demonstrative pronoun. We have three alternatives, and we choose the word *huyu* (16).

(16)



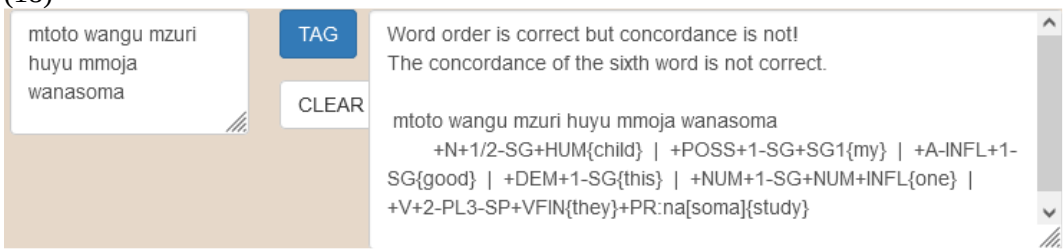
We got the warning that numeral cannot be before a demonstrative pronoun. We correct it (17).

(17)



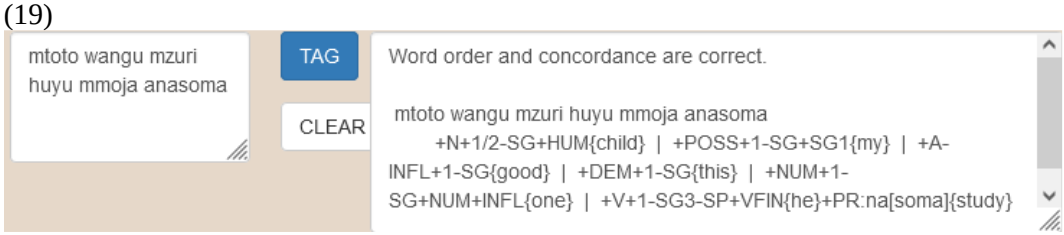
We are asked to add the verb *soma* (18). The verb can be in any of its thousands of forms, provided that the form is grammatically correct.

(18)



The word order is correct, but the concordance of the sixth word is wrong. We correct it (19).

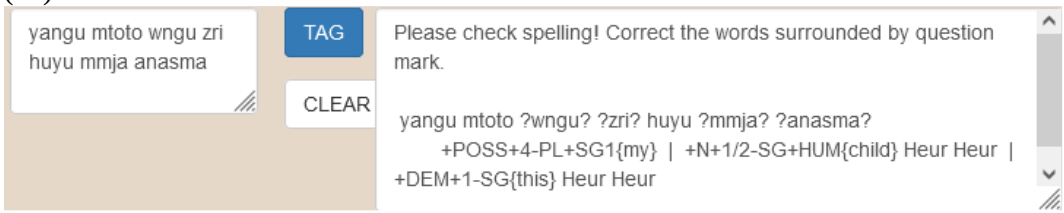
(19)



4 Different types of mistakes in the same string

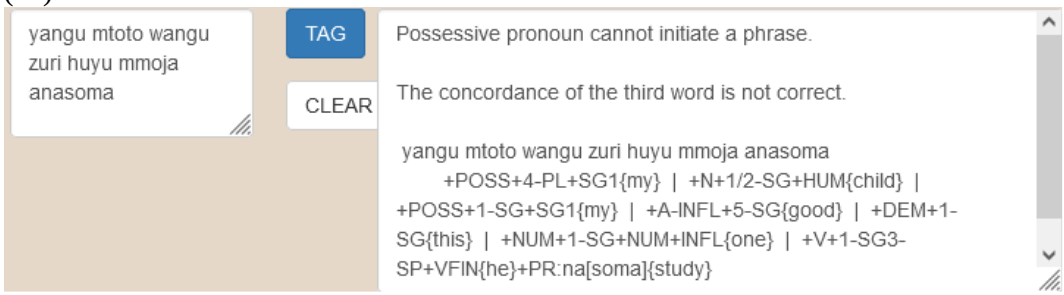
Next, we test how the system behaves when we have all three types of mistakes, word order, concordance and typos (20).

(20)



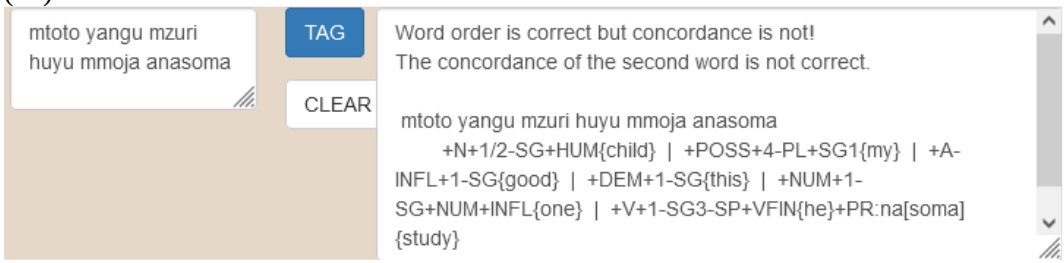
Very first, the system asks to correct misspelled words. Typos are surrounded with question marks (21).

(21)



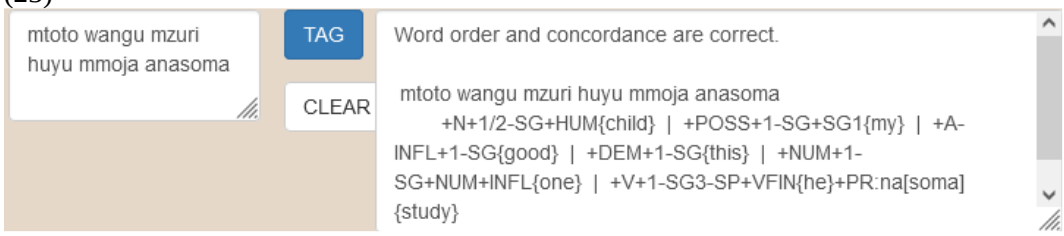
Now we have two warnings. Possessive pronoun cannot initiate a phrase, and the concordance of the third word should be corrected (22).

(22)



Now the word order is correct, but we should correct the concordance of the second word (23)

(23)

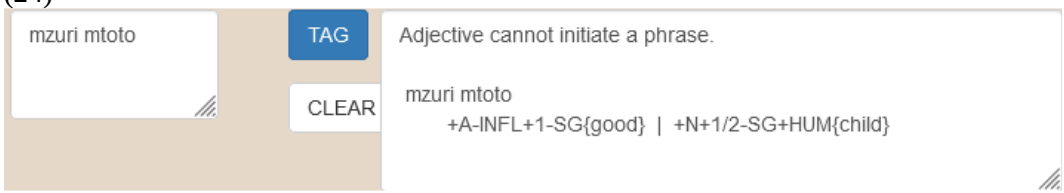


Now the system accepts the phrase as correctly typed. We noticed that feedback on mistakes was given gradually and in a certain order. The typos were checked first, because without recognizable word forms it is not possible to give sensible advice. Next come the mistakes in word order. And the last set of instructions concerns the concordance.

5. More types of mistakes

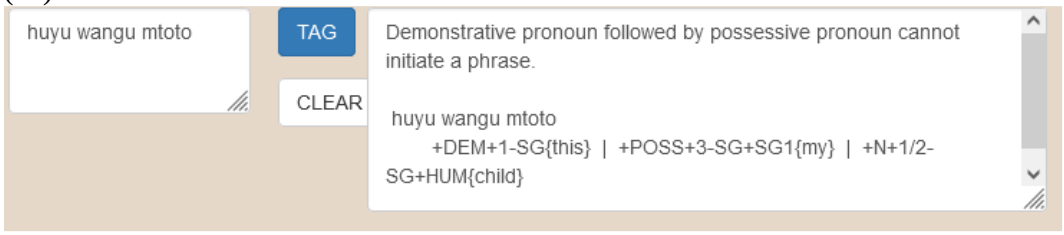
There are several ways for making mistakes in writing. Below I will give more examples of them (24).

(24)



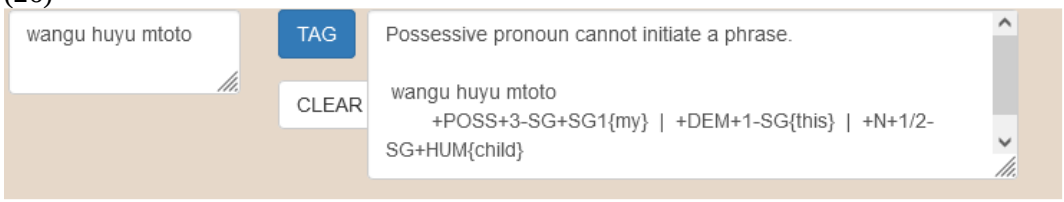
Adjective cannot initiate a phrase because it always follows the noun.
Next, we try to type a phrase meaning this my child (25).

(25)



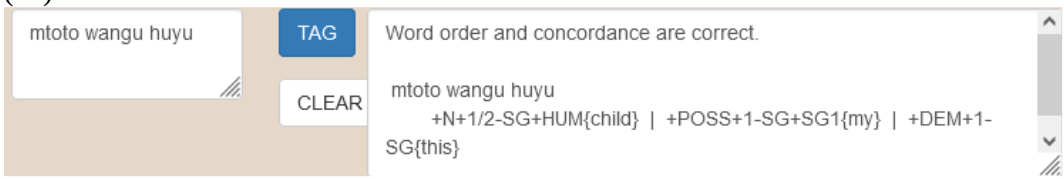
The demonstrative pronoun could initiate a phrase, but if it is followed by possessive pronoun, it is not acceptable. We try to correct it (26).

(26)



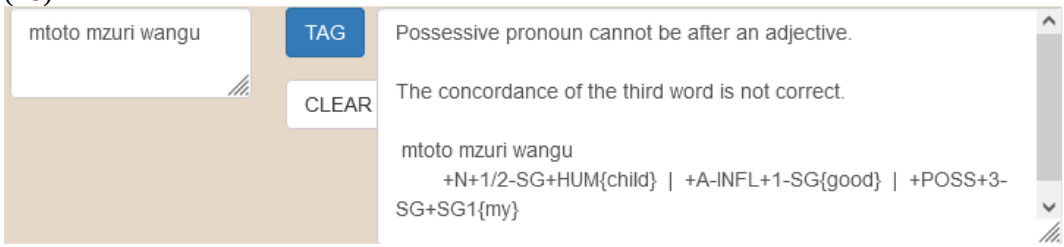
Shifting the order of words did not help. The possessive pronoun is in the wrong place. We try further (27).

(27)



Now finally the phrase is correct. There are more possibilities for mistakes (28).

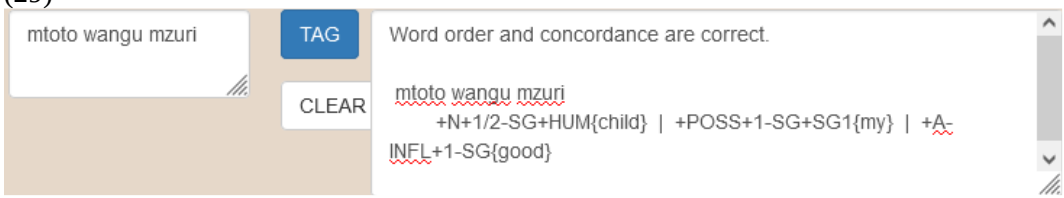
(28)



This is an interesting case. First, we are warned that possessive pronoun cannot be after an adjective. But we are also warned that the concordance of the third word is not correct. This is not true, however, because the concordance is in fact correct. The reason for the erroneous warning is that because the word order is ungrammatical, the disambiguator was not able to select the correct interpretation for the word *wangu*.

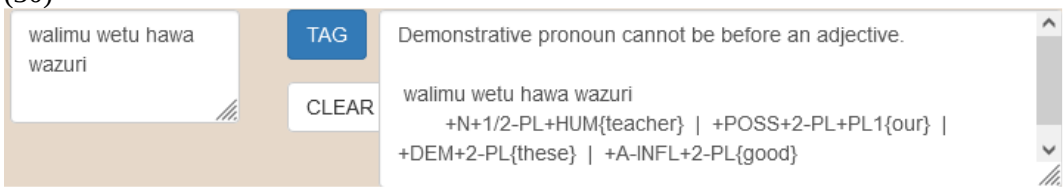
We correct the string according to the first warning (29).

(29)



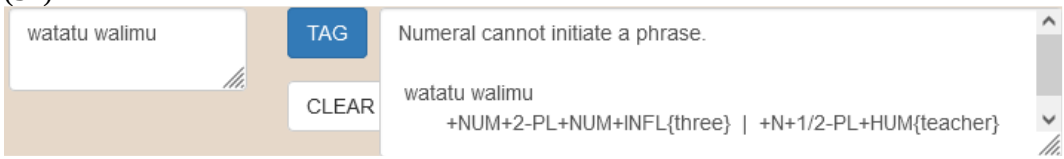
Now the string is correct, and also the word *wangu* was disambiguated correctly. More mistakes are below (30).

(30)



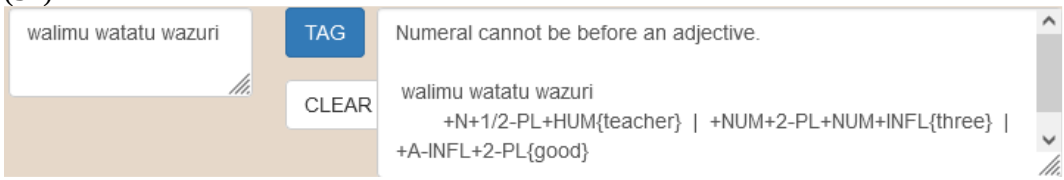
Also numerals can be in a wrong place. In (31) a numeral initiates a phrase.

(31)



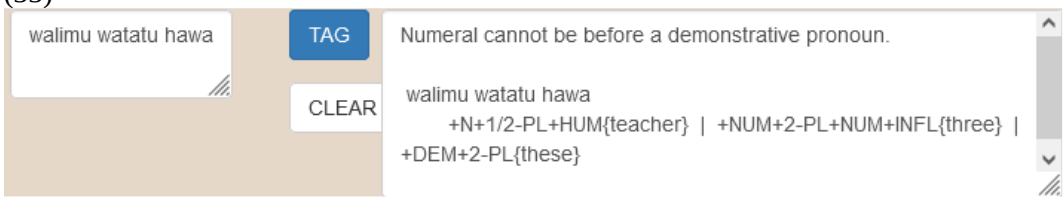
In (32) a numeral is before an adjective.

(32)



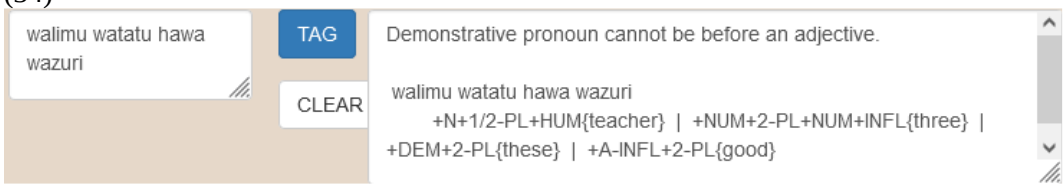
A numeral must also be in the correct place in relation to the demonstrative pronoun (33).

(33)



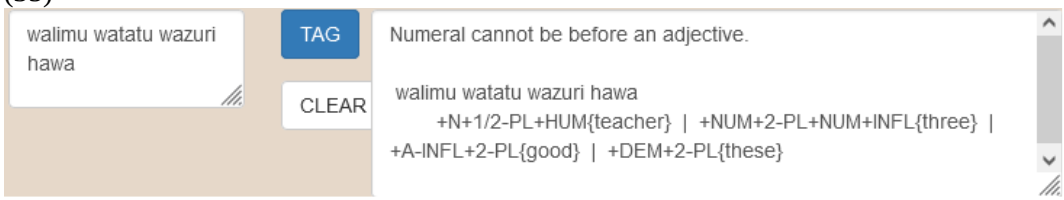
If there are more than one word in a wrong place, the system warns about them in a certain order. We add an adjective to the above phrase (34).

(34)



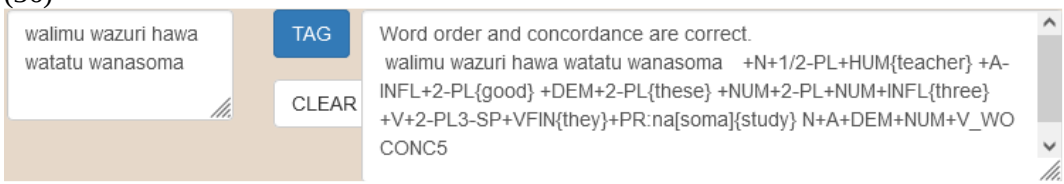
Now the warning concerns the position of the demonstrative pronoun. We correct it (35).

(35)



Now we were given a warning about the place of the numeral. We correct it (36).

(36)



After correction the string is correct.

6 Using frequency lists in training

In order to properly fulfil its task, the fully self-tutored learning system requires some sort of strategy for organising the training. An ad hoc learning system leaves easily parts of vocabulary and grammar out of the learning program. When everything is possible, the learner needs to become a planner of training.

A useful approach is to use word lists of various part-of-speech categories in constructing phrases. For example, frequency lists, such as the ones below (lists 1, 2 and 3), help the learner to train with such vocabularies, which often occur in texts. The lists contain nouns, adjectives and verbs, because these are open part-of-speech categories. On the other hand, possessive pronouns, demonstrative pronouns, and numerals are closed sets that can be handled by memorising.

The list of nouns includes also the noun class specification for each noun. Because the noun is the basic unit for defining the concordance patters of the whole phrase, the class specification helps the learner to formulate the correct forms for the rest of the members of the phrase.

If it is needed, separate lists for nouns of each noun class can also be retrieved, so that a specific noun class can be trained with several nouns of the same class. The class-specific noun lists would force the learner to train with the whole range of noun types.

Below are frequency lists of Swahili nouns, adjectives and verbs. Only 100 most frequent words of each type is listed. Nouns show also their noun class affiliation, which should help in constructing phrases.

Frequency list 1: 100 most common nouns

51577 "mtu" N 1/2 { man }
38463 "mwaka" N 3/4 { year }
30604 "wakati" N 11/10 { time } AR
30585 "serikali" N 9/10 { government } PERS
26147 "nchi" N 9/10 { earth }
23370 "nchini" N 9/10 { in the country}
21915 "mtoto" N 1/2 { child }
21553 "chama" N 7/8 { party }
19099 "jambo" N 5/6 { matter }
18013 "kiongozi" N 7/8 { leader } AN
15699 "kazi" N 9/10 { work }
14925 "eneo" N 5/6 { area }
14455 "taifa" N 5/6 { nation } AR
14433 "habari" N 9/10 { news } AR
14380 "siku" N 9/10 { day }
14018 "mjini" N 3/4 { in the town } LOC
13253 "rais" N TITLE { President } AN
12994 "hali" N 9/10 { state } AR
12589 "mwananchi" N 1/2 { citizen }
11244 "muda" N 3/4 { time }
11027 "mkutano" N 3/4 { meeting }
10773 "mkuu" N 1/2 { head }
10714 "shirika" N 5/6 { organization } AR
10699 "fedha" N 9/10 { money } MASS
10149 "maisha" N 6SG { life } AR
9980 "sheria" N 9/10 { law } AR
9935 "tatizo" N 5/6 { problem } AR
9861 "kanisa" N 5/6 { church } AR
9841 "mwanamke" N 1/2 { woman } FEM
9359 "taarifa" N 9/10 { report } AR
9303 "suala" N 5/6 { issue } AR
9264 "sehemu" N 9/10 { part } AR
9063 "hatua" N 9/10 { step } AR
8858 "mpango" N 3/4 { plan }
8819 "maendeleo" N 6SG { development }
8730 "mara" N 9/10 { time } AR
8712 "mkoa" N 3/4 { region }
8625 "mwandishi" N 1/2 { writer }
8510 "maji" N 6SG { water } MASS
8408 "mahakama" N 9/10 { court } AR
8226 "timu" N 9/10 { team } ENG
8072 "nafasi" N 9/10 { opportunity } AR
8031 "njia" N 9/10 { way }
7969 "kijana" N 7/8 { youth } AN

7936 "tukio" N 5/6 { event }
7695 "chakula" N 7/8 { food } MASS
7620 "uchaguzi" N 11/10 { election }
7611 "wilaya" N 9/10 { district } AR
7461 "jamii" N 9/10 { community } AR
7410 "haki" N 9/10 { right } AR
7202 "mji" N 3/4 { town }
7156 "ugonjwa" N 11/6 { disease }
7092 "nyumba" N 9/10 { house }
7079 "polisi" N 9/10 { police } ENG
6880 "mwezi" N 3/4 { month }
6856 "msaada" N 3/4 { assistance } AR
6793 "mchezo" N 3/4 { play }
6793 "kitu" N 7/8 { thing }
6749 "jeshi" N 5/6 { army }
6697 "aina" N 9/10 { kind }
6607 "jijini" N 5/6 { in the city}
6549 "umoja" N 11 { unity }
6542 "huduma" N 9/10 { service }
6497 "kesi" N 9/10 { case }
6479 "amani" N 9/10 { peace } AR
6396 "mfanyakazi" N 1/2 { worker }
6236 "jina" N 5/6 { name }
6148 "kituo" N 7/8 { station }
6088 "mkazi" N 1/2 { inhabitant }
5960 "waziri" N 9/6 { Minister } AR
5923 "gari" N 5/6 { car }
5917 "lengo" N 5/6 { target }
5902 "polisi" N 9/10 { police officer } AN ENG
5761 "mwanafunzi" N 1/2 { pupil }
5757 "nguvu" N 9/10 { power }
5714 "kundi" N 5/6 { group }
5686 "kipindi" N 7/8 { period }
5683 "upande" N 11/10 { direction }
5589 "rais" N 9/6 { President } MALE AR
5559 "uwezo" N 11 { ability }
5551 "mwanachama" N 1/2 { party member }
5495 "elimu" N 9/10 { education } AR
5492 "wiki" N 9/10 { week } ENG
5472 "gazeti" N 5/6 { newspaper } ENG
5398 "kamanda" N 9/6 { commander } MALE ENG
5376 "shughuli" N 9/10 { activity } AR
5340 "kamati" N 9/10 { committee } ENG
5309 "mazingira" N 6SG { environment }
5280 "neno" N 5/6 { word }
5239 "biashara" N 9/10 { trade }

5234 "mbunge" N 1/2 { member of parliament }
5189 "kijiji" N 7/8 { village }
5132 "wizara" N 9/10 { ministry } AR
5128 "shule" N 9/10 { school } GER
5117 "asilimia" N 9/10 { percent }
5039 "uongozi" N 11/6 { leadership }
5007 "kampuni" N 9/10 { company } ENG
4997 "silaha" N 9/10 { weapon }
4953 "duniani" N 9/10 { on the earth } AR
4909 "kitendo" N 7/8 { action }

Frequency list 2: 100 most common adjectives

Note that adjectives are either inflecting (A-INFL) or uninflecting (A-UNINFL). The system controls that adjectives are used correctly. It warns you if you try to inflect an uninflecting adjective, or if you try to add an inflection prefix to an uninflecting adjective..

20422 "kubwa" ADJ A-INFL { big }
12791 "kuu" ADJ A-INFL { great }
11153 "mbalimbali" ADJ A-UNINFL { various }
10643 "pya" ADJ A-INFL { new }
8013 "ingine" ADJ A-INFL NE 1/2 { other }
7575 "dogo" ADJ A-INFL { small }
7525 "ingine" ADJ A-INFL NE 9/10 { other }
7468 "zuri" ADJ A-INFL { good }
6535 "muhimu" ADJ A-UNINFL { important }
5915 "ingine" ADJ A-INFL NE 5/6 { other }
5577 "ingine" ADJ A-INFL NE 1/2 { others }
5245 "tayari" ADJ A-UNINFL { ready }
4580 "zima" ADJ A-INFL { whole }
4375 "refu" ADJ A-INFL { long }
4288 "bora" ADJ A-UNINFL { excellent }
3902 "kadhaa" ADJ A-UNINFL { some }
3403 "kali" ADJ A-INFL { sharp }
3367 "chache" ADJ A-INFL { few }
3304 "fulani" ADJ A-UNINFL { certain }
3292 "maalum" ADJ A-UNINFL { special } AR
3143 "baya" ADJ A-INFL { bad }
2721 "binafsi" ADJ A-UNINFL { private } AR
2582 "fupi" ADJ A-INFL { short }
2543 "geni" ADJ A-INFL { foreign }
2511 "gumu" ADJ A-INFL { hard }
2468 "ema" ADJ A-INFL { good }
2326 "ingine" ADJ A-INFL NE 7/8 { other }
2198 "pekee" ADJ A-UNINFL { unique }

2122 "huru" ADJ A-UNINFL { free } AR
2015 "katoliki" ADJ A-UNINFL { catholic } ENG
2002 "halisi" ADJ A-UNINFL { real } AR
1872 "wazi" ADJ A-UNINFL { clear } AR
1686 "safi" ADJ A-UNINFL { clean } AR
1662 "takatifu" ADJ A-INFL { holy } AR
1654 "kamili" ADJ A-UNINFL { complete }
1565 "rasmi" ADJ A-UNINFL { official } AR
1556 "mojawapo" ADJ A-INFL { one }
1435 "kali" ADJ A-INFL { angry }
1408 "zito" ADJ A-INFL { heavy }
1387 "rahisi" ADJ A-UNINFL { easy } AR
1387 "ingine" ADJ A-INFL NE 9/10 { others }
1303 "maalumu" ADJ A-UNINFL { special } AR
1271 "maarufu" ADJ A-UNINFL { famous } AR
1265 "ingine" ADJ A-INFL NE 11 { other }
1227 "ingine" ADJ A-INFL NE 3/4 { other }
1197 "halali" ADJ A-UNINFL { lawful }
1180 "tofauti" ADJ A-UNINFL { different } AR
1005 "changa" ADJ A-INFL { young }
990 "dhahiri" ADJ A-UNINFL { evident }
928 "dogodogo" ADJ A-INFL { small }
907 "hai" ADJ A-UNINFL { alive }
887 "ingine" ADJ A-INFL NE 7/8 { others }
869 "ema" ADJ A-INFL { good } AR
846 "chafu" ADJ A-INFL { dirty }
792 "imara" ADJ A-INFL { strong } AR
784 "eupe" ADJ A-INFL { white }
784 "duni" ADJ A-UNINFL { inferior } AR
739 "mosi" ADJ A-UNINFL { first }
738 "ekundu" ADJ A-INFL { red }
735 "kadha" ADJ A-UNINFL { some } AR
717 "tupu" ADJ A-INFL { empty }
675 "ingine" ADJ A-INFL NE 5/6 { others }
667 "makini" ADJ A-UNINFL { attentive }
641 "mashuhuri" ADJ A-UNINFL { famous } AR
640 "bovu" ADJ A-INFL { defective }
633 "maskini" ADJ A-UNINFL { poor } AR
602 "tukufu" ADJ A-INFL { glorious } AR
596 "eusi" ADJ A-INFL { black }
509 "ngapi" ADJ A-INFL INTERROG { how many }
493 "sahihi" ADJ A-UNINFL { correct } AR
489 "julikana" ADJ - { unknowing }
472 "sawa" ADJ A-UNINFL { same } AR
462 "dhaifu" ADJ A-INFL { weak } AR
448 "endelevu" ADJ A-UNINFL { sustainable }

447 "haramu" ADJ A-UNINFL { forbidden }
445 "ingine" ADJ A-INFL NE 16 { in another place }
444 "huria" ADJ A-UNINFL { free }
430 "katili" ADJ A-INFL { cruel } AR
422 "madhubuti" ADJ A-UNINFL { firm } AR
418 "salama" ADJ A-UNINFL { safe } AR
417 "aminifu" ADJ A-INFL { faithful } AR
396 "tajiri" ADJ A-UNINFL { rich } AR
392 "pana" ADJ A-INFL { wide }
392 "asilia" ADJ A-UNINFL { original } AR
389 "muafaka" ADJ A-UNINFL { acceptable }
374 "nafuu" ADJ A-UNINFL { modest } AR
373 "zee" ADJ A-INFL { old }
372 "masikini" ADJ A-UNINFL { poor } AR
352 "kamilifu" ADJ A-INFL { perfect } AR
351 "zazi" ADJ A-INFL { fruitful }
339 "epesi" ADJ A-INFL { light }
335 "kongwe" ADJ A-INFL { very old }
321 "ovu" ADJ A-INFL { evil }
317 "embamba" ADJ A-INFL { narrow }
311 "kavu" ADJ A-INFL { dry }
308 "bandia" ADJ A-UNINFL { artificial } AR
294 "tamu" ADJ A-INFL { sweet } AR
279 "nyeti" ADJ A-INFL { crucial }
274 "ja" ADJ { coming } NCL-PL
273 "heri" ADJ A-UNINFL { happy } AR

Frequency list 3: 100 most common verbs

91707 "sema" V { say }
91328 "wa" V { be }
40626 "ni" V { is }
32756 "fanya" V { do }
26028 "ni" V { are }
23062 "taka" V { want }
22277 "toa" V { give }
22102 "pata" V { get }
20669 "ni" V INIT { is }
17278 "endelea" V { continue }
17057 "tumia" V { use } APPL
16893 "weza" V { be able to }
16800 "pa" V { give }
16707 "anza" V { begin }
16649 "dai" V { claim } AR
16116 "toka" V { leave } STAT

15270 "enda" V { go }
14858 "ona" V { see }
14087 "weza" V { can }
10472 "fanyika" V { be done } STAT
10231 "saidia" V { help } AR
9948 "eleza" V { explain }
9692 "jua" V { know }
9587 "weka" V { place }
9210 "ni" V INIT { are }
9108 "tokea" V { appear } APPL
8806 "tolea" V { give } APPL
8740 "shiriki" V { participate } AR
8692 "omba" V { ask }
8666 "chukua" V { take }
8239 "amba" V { tell }
8071 "ingia" V { enter }
7752 "taja" V { mention }
7477 "ongeza" V { increase }
7288 "fikia" V { arrive } APPL
7017 "sema" V { speak }
6795 "fika" V { arrive }
6758 "onyesha" V { show }
6465 "zungumza" V { discuss }
6329 "ishi" V { live } AR
6288 "jenga" V { build }
6199 "takia" V { wish } APPL
6094 "acha" V { leave }
5904 "sababisha" V { cause } AR CAUS
5876 "leta" V { bring }
5833 "ja" V { come }
5809 "onekana" V { be seen } STAT REC
5804 "peleka" V { send }
5793 "ita" V { call }
5765 "lipa" V { pay }
5596 "jibu" V { answer } AR
5587 "anzisha" V { begin } CAUS
5330 "tangaza" V { announce } CAUS
5328 "kuna" V { there is }
5302 "amini" V { believe } AR
5287 "ongoza" V { lead }
5175 "amua" V { decide }
5136 "hakuna" V { there is not }
5017 "patikana" V { be available } STAT REC
4954 "tafuta" V { search }
4855 "uliza" V { ask }
4801 "pita" V { pass }

4730 "pinga" V { oppose }
4724 "tekeleza" V { implement }
4706 "fuatia" V { follow } APPL
4689 "zuia" V { prevent }
4612 "endesha" V { operate } CAUS
4610 "ondoka" V { leave } STAT
4601 "fuata" V { follow }
4540 "kamata" V { catch }
4515 "kubali" V { agree } AR
4496 "andika" V { write }
4413 "andaa" V { prepare }
4403 "hakikisha" V { ensure } AR CAUS
4395 "kutana" V { meet each other } REC
4345 "kuta" V { meet }
4337 "tumika" V { be used } STAT
4308 "funga" V { imprison }
4301 "penda" V { like }
4279 "piga" V { hit } ACT
4250 "husika" V { be concerned } STAT AR
4219 "pasa" V { suit }
4193 "pokea" V { receive }
4083 "tegemea" V { rely on }
4078 "hitaji" V { need } AR
4073 "tambua" V { recognize }
4071 "rudi" V { return } AR
4000 "punguka" V { decrease } CAUS
3980 "panga" V { arrange }
3977 "kabili" V { face } AR
3943 "kaa" V { stay }
3926 "pitika" V { review } STAT
3899 "soma" V { read }
3889 "patia" V { get } APPL
3868 "unda" V { construct }
3859 "ua" V { kill }
3828 "anzia" V { begin } APPL
3817 "ruhusu" V { permit } AR
3778 "zidi" V { increase } AR
3774 "chagua" V { choose }

7 Rule-based language technology in helping to understand the chosen text

The language analysis system can also be applied for helping the learners in comprehension. An important part of language learning consists of text reading and its translation. Translation itself is only a proof that the learner has comprehended the structures and words of the source text. Many learners may today use such aids as Google Translate for helping in comprehension. Although this service has improved over the years,

it has a serious disadvantage. It is not able to help the learner to understand the structure of individual words.

By using rule-based technology, we can help the learner in various ways for understanding the text.

One method is to analyse and disambiguate the text and provide each word with linguistic information. For the sake of demonstration, I have taken a short extract from BBC News Swahili on 29.4. 2022 (37).

(37)
Kwa nini Afrika imeshindwa kufikia lengo la chanjo ya kimataifa
"Hatushauri utumiaji wa barakoa pengini ujiskie mgonjwa," aliongeza Maria Van Kerkhove, Kiongozi wa Kiufundi wa Covid-19.
Lakini kile tulichojifunza tangu kuzuka kuanza kimebadilisha maoni hayo. WHO sasa inasema watu wanapaswa "kufanya kuvaa barakoa kuwa sehemu ya kawaida ya kuwa karibu na watu wengine."

Out of this text, we can produce various types of learning aids. We can produce the analysis inan xml-format (38)

(38)

```
<s>   <s>   {      <s> }
Kwa_nini    kwa_nini    INTERROG    { why }     @ADVL @ADVL
Afrika      Afrika      PROPNAME     SG PLACE   { Africa }  @SUBJ
imeshindwa  shindwa     V      SUB-PREF=9-SG TAM=PERF:me [shinda]
      { fail }    @FMAINVtr-OBJ>    SVO VFIN
kufikia     fikia V      [fika]     { arrive at }     @-FMAINV-n
      APPL SVO INF
lengo lengo N      5/6-SG     { target }  @OBJ
la    la    GEN-CON      5-SG { of }      @GCON
chanjo      chanjo      N      9/10-SG    { vaccination }   @<NH
ya_kimataifa      ya_kimataifa      ADJ   _     { international }
      @<NADJ
"<?
Hatushauri  shauri      V      TAM=NEG-a SUB-PREF=2-PL1 [shauri]
      { Advise }  @FMAINVtr+OBJ>    AR SVO VFIN
utumiaji    utumiaji    N      11-SG      { use }     @OBJ
wa    wa    GEN-CON      11-SG      { of }      @GCON
barakoa     barakoa     N      9/10-SG    { veil }    @<NH  AR
pengine     ingine      ADV   _      { perhaps } @<DN
ujiskie     ujiskie     N      Heur 11-SG { ujiskie } @OBJ
mgonjwa     mgonjwa     N      1/2-SG     { patient } @SUBJ
,     ,     COMMA _     { , }
aliongeza   ongeza      V      SUB-PREF=1-SG3 TAM=PAST [ongeza]
      { add }     @FMAINVtr+OBJ>    SVO VFIN
Maria Maria PROPNAME    _     { Maria }   @OBJ  FEM } FEM
Van   Van   PROPNAME     Heur { Van }     @OBJ
Kerkhove    Kerkhove    PROPNAME     Heur { Kerkhove }      @OBJ
,     ,     COMMA _     { , }
Kiongozi    kiongozi    N      7/8-SG     { Leader }  @<P
```

```
wa_Kiufundi wa_ufundi   ADJ   _     { technical }     @<NADJ
wa    wa    GEN-CON      3-SG { of }      @GCON
Covid-19    Covid-19    PROPNAME     Heur { Covid-19 }      @<NH
"<.
Lakini      lakini      CONJ  _     { but }     @CS   AR
kile  kile  PRON   DEM le 7-SG      { that }    @SUBJ
tulichojifunza    jifunza     V      SUB-PREF=2-PL1 TAM=PAST 7-SG-
REL   { what }    @FMAINVtr-OBJ> [funza]  { learn } SVO VFIN
tangu tangu PREP  _     { since }   @ADVL
kuzuka      zuka  N      15-SG [zua]      { emerge }  @<P   STAT
SV
kuanza      anza  V      [anza]     { begin }   @-FMAINV-n  SVO
INF
kimebadilisha     badilisha   V      SUB-PREF=7-SG TAM=PERF:me
[badili]    { change }  @FMAINVtr+OBJ>    CAUS AR SVO VFIN
maoni oni   N      5/6-PL     { view }    @OBJ
hayo  hayo  PRON   DEM 6-PL   { these }   @<NDEM
.     .     _     _     { . }
</sentence>
<s>   <s>   {      <s> }
WHO   WHO   PROPNAME     SG   { WHO }     @SUBJ
sasa  sasa  ADV   _     { now }     @ADVL AR
inasema     sema  V      SUB-PREF=9-SG TAM=PR:na [sema]     { say
}     @FMAINVtr+OBJ>    SVO VFIN
watu  mtu   N      1/2-PL     { man }     @SUBJ
wanapaswa   paswa V      SUB-PREF=2-PL3 TAM=PR:na [pasa]    { suit
to }  @FMAINVtr-OBJ>    PASS SVO VFIN
kufanya     fanya V      [fanya]    { do }      @-FMAINV-n  SVO
INF
kuvaa vaa   V      [vaa]      { dress }   @-FMAINV-n  SVO INF
barakoa     barakoa     N      9/10-SG    { veil }    @OBJ  AR
kuwa  wa    V      [wa] { be }      @-FMAINV-n  AUX-WA MONOSLB SV
INF
sehemu      sehemu      N      9/10-SG    { part }    @<P   AR
ya_kawaida  ya_kawaida  ADJ   _     { usual }   @<NADJ
ya_kuwa     ya_wa CONJ  _     { that }    @CS
karibu_na   karibu_na   ADV   _     { near }    @ADVL
watu  mtu   N      1/2-PL     { man }     @<P
wengine     ingine      ADJ    A-INFL 2-PL      { other }   @<DN
.     .     _     _     { . }
</sentence>
```

Another type of analysis result is in (39).

(39)

```
"<<s>>"
      "<s>" { <s> }
"<*kwa_nini>"
      "kwa_nini" INTERROG { why } @ADVL @ADVL
"<*afrika>"
      "*afrika" N PROPNAME SG PLACE { *africa } @SUBJ
```

```
"<imeshindwa>"
      "shindwa" V 9-SG-SP VFIN NO-SP-GLOSS PERF:me [shinda] PS {
be fail } SVO PREFR PASS-LEX AUXMOD @FMAINVtr-OBJ>
"<kufikia>"
      "fikia" V INF { to } z [fika] { arrive at } SVO APPL @-
FMAINV-n
"<lengo>"
      "lengo" N 5/6-SG { the } { target } @OBJ
"<la>"
      "la" GEN-CON 5-SG { of } @GCON
"<chanjo>"
      "chanjo" N 9/10-SG { vaccination } @<NH
"<ya_kimataifa>"
      "ya_kimataifa" ADJ { international } @<NADJ
"<?
"<*hatushauri>"
      "shauri" V NEG-a 2-PL1-SP VFIN { we } z [shauri] { *advise }
SVO AR CAP @FMAINVtr+OBJ>
"<utumiaji>"
      "utumiaji" N 11-SG { use } @OBJ
"<wa>"
      "wa" GEN-CON 11-SG { of } @GCON
"<barakoa>"
      "barakoa" N 9/10-SG { the } { veil } AR @<NH
"<pengine>"
      "ingine" ADV { perhaps } @<DN
"<ujiskie>"
      "ujiskie" <Heur> N 11-SG { ujiskie } @OBJ
"<mgonjwa>"
      "mgonjwa" N 1/2-SG HUM { the } { patient } @SUBJ
"<,>"
      "," COMMA { , }
"<aliongeza>"
      "ongeza" V 1-SG3-SP VFIN NO-SP-GLOSS PAST z [ongeza] { add }
SVO PREFR @FMAINVtr+OBJ>
"<*maria>"
      "*maria" N PROPNAME AN HUM { *maria } FEM } FEM @OBJ
"<*van>"
      "*van" <Heur> PROPNAME { *van } CAP @OBJ
"<*kerkhove>"
      "*kerkhove" <Heur> PROPNAME { *kerkhove } CAP PROP-CAND
PROP-CAND @OBJ
"<,>"
      "," COMMA { , }
"<*kiongozi>"
      "kiongozi" N 7/8-SG { the } { *leader } AN HUM CAP @<P
"<wa_*kiufundi>"
      "wa_ufundi" ADJ { technical } @<NADJ
"<wa>"
      "wa" GEN-CON 3-SG { of } @GCON
"<*covid-19>"
```

```
      "*covid-19" <Heur> PROPNAME { *covid-19 } CAP @<NH
"<.
"<*lakini>"
      "lakini" CONJ { but } **CLB AR CAP @CS
"<kile>"
      "kile" PRON DEM :le 7-SG { that } @SUBJ
"<tulichojifunza>"
      "jifunza" V 2-PL1-SP VFIN NO-SP-GLOSS PAST 7-SG-REL { what }
z [funza] { learn } SVO PREFR @FMAINVtr-OBJ>
"<tangu>"
      "tangu" PREP { since } @ADVL
"<kuzuka>"
      "zuka" N 15-SG z [zua] { emerge } SV :STAT PREFR @<P
"<kuanza>"
      "anza" V INF { to } z [anza] { begin } SVO @-FMAINV-n
"<kimebadilisha>"
      "badilisha" V 7-SG-SP VFIN { it } PERF:me z [badili] {
change } SVO :CAUS PREFR AR @FMAINVtr+OBJ>
"<maoni>"
      "oni" N 5/6-PL { the } { view } @OBJ
"<hayo>"
      "hayo" PRON DEM :hV ASS-OBJ 6-PL { these } @<NDEM
"<.$>"
      ".$" { .$ } **CLB
"<<s>>"
      "<s>" { <s> }
"<*w*h*o>"
      "*w*h*o" N PROPNAME SG { *w*h*o } @SUBJ
"<sasa>"
      "sasa" ADV { now } AR @ADVL
"<inasema>"
      "sema" V 9-SG-SP VFIN NO-SP-GLOSS PR:na z [sema] { say } SVO
@FMAINVtr+OBJ>
"<watu>"
      "mtu" N 1/2-PL HUM { the } { man } @SUBJ
"<wanapaswa>"
      "paswa" V 2-PL3-SP VFIN NO-SP-GLOSS PR:na [pasa] PS { be
suit to } SVO PASS @FMAINVtr-OBJ>
"<kufanya>"
      "fanya" V INF { to } z [fanya] { do } SVO @-FMAINV-n
"<kuvaa>"
      "vaa" V INF { to } z [vaa] { dress } SVO @-FMAINV-n
"<barakoa>"
      "barakoa" N 9/10-SG { the } { veil } AR @OBJ
"<kuwa>"
      "wa" V INF { to } z [wa] { be } AUX-WA SV MONOSLB @-FMAINV-n
"<sehemu>"
      "sehemu" N 9/10-SG { the } { part } AR @<P
"<ya_kawaida>"
      "ya_kawaida" ADJ { usual } @<NADJ
"<ya_kuwa>"
```

```
      "ya_wa" CONJ { that } @CS
"<karibu_na>"
      "karibu_na" ADV { near } @ADVL
"<watu>"
      "mtu" N 1/2-PL HUM { the } { man } @<P
"<wengine>"
      "ingine" ADJ A-INFL :INGINE 2-PL NOART { other } @<DN
"<.$>"
      ".$" { .$ } **CLB
```

If the learner is on an advanced stage, the output types above are perhaps too detailed. The leaner would need help only for more rare words. For this purpose, we can produce word lists of various levels. I have prepared vocabulary extraction programs that produce vocabularies for five different learning levels.

On the first level, the system produces a full vocabulary of the chosen text (40).

(40)
Afrika N PROPNAME SG PLACE { Africa }
anza V [anza] { begin } SVO
badilisha V [badili] { change } SVO CAUS AR
barakoa N 9/10 { veil } AR
chanjo N 9/10 { vaccination }
fanya V [fanya] { do } SVO
fikia V [fika] { arrive at } SV APPL
hayo PRON DEM 6-PL { these }
ingine ADJ A-INFL 2-PL { other }
ingine ADV { perhaps }
jifunza V [funza] { learn } SVO
karibu_na ADV { near }
kile PRON DEM le 7-SG { that }
kiongozi N 7/8 { leader } HUM
kwa_nini INTERROG { why }
la 5-SG { of }
lakini CONJ { but } AR
lengo N 5/6 { target }
Maria N PROPNAME { Maria } FEM
mgonjwa N 1/2 { patient }
mtu N 1/2 { man }
ongeza V [ongeza] { add } SVO
oni N 5/6 { view }
paswa V [pasa] { suit } SVO PASS
sasa ADV { now } AR
sehemu N 9/10 { part } AR
sema V [sema] { say } SVO
shauri V NEG-a [shauri] { advise } SVO AR
shindwa V [shinda] { fail } SVO
tangu PREP { since }

utumiaji N 11 { use }
vaa V [vaa] { dress } SVO
wa 11-SG { of }
wa 3-SG { of }
wa_ufundi ADJ { technical }
wa V [wa] { be } SV MONOSLB
WHO N PROPNAME SG { WHO }
ya_kawaida ADJ { usual }
ya_kimataifa ADJ { international }
ya_wa CONJ { that }
zuka V [zua] { emerge } SV STAT

On the next level of learning, the first 500 most frequent words are removed. The result is in (41).

(41)
Afrika - N PROPNAME SG PLACE { Africa }
badilisha - V <badili> { change } SVO CAUS AR
barakoa - N 9/10 { veil } AR
chanjo - N 9/10 { vaccination }
hayo - PRON DEM 6-PL { these }
ingine - ADV { perhaps }
jifunza - V <funza> { learn } SVO
karibu_na - ADV { near }
kile - PRON DEM le 7-SG { that }
kwa_nini - INTERROG { why }
la - 5-SG { of }
Maria - N PROPNAME { Maria } FEM
mgonjwa - N 1/2 { patient }
oni - N 5/6 { view }
shauri - V NEG-a <shauri> { advise } SVO AR
utumiaji - N 11 { use }
vaa - V <vaa> { dress } SVO
wa - 11-SG { of }
wa - 3-SG { of }
wa_ufundi - ADJ { technical }
WHO - N PROPNAME SG { WHO }
ya_kawaida - ADJ { usual }
ya_kimataifa - ADJ { international }
ya_wa - CONJ { that }
zuka - V <zua> { emerge } SV STAT

On the next level of learning we cut the 1000 most frequent words off (42).

(42)
Afrika - N PROPNAME SG PLACE { Africa }

badilisha - V <badili> { change } SVO CAUS AR
barakoa - N 9/10 { veil } AR
chanjo - N 9/10 { vaccination }
hayo - PRON DEM 6-PL { these }
ingine - ADV { perhaps }
jifunza - V <funza> { learn } SVO
kile - PRON DEM le 7-SG { that }
kwa_nini - INTERROG { why }
la - 5-SG { of }
Maria - N PROPNAME { Maria } FEM
utumiaji - N 11 { use }
vaa - V <vaa> { dress } SVO
wa - 11-SG { of }
wa - 3-SG { of }
wa_ufundi - ADJ { technical }
WHO - N PROPNAME SG { WHO }
ya_kawaida - ADJ { usual }
ya_kimataifa - ADJ { international }
ya_wa - CONJ { that }
zuka - V <zua> { emerge } SV STAT

On the even more advanced level, we cut the first 1500 words of the frequency list off (43).

(43)
Afrika - N PROPNAME SG PLACE { Africa }
badilisha - V <badili> { change } SVO CAUS AR
barakoa - N 9/10 { veil } AR
chanjo - N 9/10 { vaccination }
hayo - PRON DEM 6-PL { these }
kile - PRON DEM le 7-SG { that }
kwa_nini - INTERROG { why }
la - 5-SG { of }
Maria - N PROPNAME { Maria } FEM
utumiaji - N 11 { use }
wa - 11-SG { of }
wa - 3-SG { of }
wa_ufundi - ADJ { technical }
WHO - N PROPNAME SG { WHO }
ya_kawaida - ADJ { usual }
ya_kimataifa - ADJ { international }
ya_wa - CONJ { that }

Finally, we have the list for the most advanced level learners. We cut off the words that appear among the first 2000 words in the frequency list (44).

(44)
Afrika - N PROPNAME SG PLACE { Africa }
barakoa - N 9/10 { veil } AR
chanjo - N 9/10 { vaccination }
hayo - PRON DEM 6-PL { these }
kile - PRON DEM le 7-SG { that }
kwa_nini - INTERROG { why }
la - 5-SG { of }
Maria - N PROPNAME { Maria } FEM
utumiaji - N 11 { use }
wa - 11-SG { of }
wa - 3-SG { of }
wa_ufundi - ADJ { technical }
WHO - N PROPNAME SG { WHO }
ya_kawaida - ADJ { usual }
ya_kimataifa - ADJ { international }
ya_wa - CONJ { that }

We see that the vocabulary lists become shorter when we gradually cut off the words that appear on the top of the frequency list. In the two last examples the difference is minimal, because the words left appear seldom between 1500 and 2000 in the frequency list.

8 Discussion and conclusion

In this report I have demonstrated various possibilities for helping learners in self-tutored learning. The key components are the morphological analyser and disambiguator, without which such learning systems cannot be constructed. The more precise and comprehensive the basic language processing system is, the better applications can be constructed.

If the analysis is comprehensive, as is the case with Swahili, the whole language, including its vocabulary and all morphological forms, can be processed into such a form, on the basis of which the learning systems described here can be constructed.

The system currently has two learning branches. One is the guided tour through all important structures of the language, using a selected vocabulary. Another one is fully open in regard to vocabulary and structures. Various aids, such as vocabularies of various types can be extracted for helping the learner to find all such grammatical structured that need training.

An addition to vocabulary and structure training, the system also includes means for helping in text comprehension. Aids suitable for different levels of learners can be produced, and the learner can freely select the text for training comprehension.