Technical Reports on Language Technology
Report No 53, 2020
http://www.njas.helsinki.fi/salama

Out of vocabulary guesser: Swahili[1]

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

Abstract

Free texts include also such words, which are not listed in the analysis system. Yet they need to be treated as part of the vocabulary, so that the unknown elements in text do not unnecessarily disturb the translation process. They cannot be fully treated as the known lexical items, but if we know some basic propertied of the words, we can figure out the structure of the sentence kore precisely. Traditionally, the heuristic guessing of such unknown words was done on the basis of the morphological form of the word only. In this report it is suggested that the unknown words should be treated in two phases. First, we give a tentative assignment of the word in the word-level guesser. In the second phase we test the assignment in context. The first assignment may have two or more assignment candidates, and in the second phase we test which one is the correct one in the context.

Key Words: *morphology, word analysis guesser, machine translation.*

1 Introduction

Unknown words constitute a constant problem in machine translation. Although we cannot give a translation of an unknown word, we can greatly reduce the problems in translation, if we can give the unknown word such interpretations as POS and syntactic role. Then we can treat the word as belonging to a certain POS and syntactic category. If no gloss for the word can be found, we can treat the word as such without translation and treat it as an untranslated item with normal inflection paradigms, as we do with proper names.

It depends very much on the language type, how the heuristic guesser should be constructed. A language such as English, which has only a marginal morphological structure, is very difficult to handle on the word level, because there are often no morphological features, which would help in guessing. Languages with rich inflectional paradigm, such as Swahili and Finnish, have many such features, which give hints for guessing. In this report, examples are from Swahili.

2 Morphological features as indicators of POS

In Bantu languages, such as Swahili, the noun class system gives hints to POS assignment. Nouns of various noun classes have typically class prefixes, which distinguish them as belonging to a certain noun class. There is ambiguity, but the prefix often at least helps to rule out many such possibilities, which in no case can come into question. Some classes, such as classes 7, 8, and 11 have unambiguous class prefixes and need no further disambiguation. Some other classes, such as classes 9 and 10 get easily mixed with class 5, because they normally are without prefix. In such cases an unknown word, which is not a verb, should be given all three alternative assignments.

The unknown word, which seems to have a noun class prefix, can be either a noun or an adjective. Adjectives get the prefix according to the subject head, and often the prefix is identical in both cases. It is not possible to decide between these two alternatives on the word level. One must look at the context and make the decision on that basis.

On the other hand, in Swahili as well as in other Bantu languages, the number of grammatical adjectives is rather small, and it is unlikely that such unknown adjectives are encountered in text. Therefore, the ambiguity between adjectives and nouns is very limited.

Because such POS categories as pronoun, preposition, and conjunction are closed classes, the unknown word cannot be any of them. The decision is very likely between the noun classes of nouns. Also, adverbs come into question, but the only possibility in them are the classes 7 and 8, which often are used as adverb markers in derived adverbs.

On the basis of the above considerations, we get the following possibilities for classification.

(1)

Prefix	Class	POS
mC	1, 3,	N, A
mi	4	N, A
wa	2	N, A
0	5, 9, 10	N, A ADV
ma	6	N, A
ki	7	N, A, ADV
vi	8	N, A, ADV
ch	7	N, A, ADV
vy	8	N, A, ADV
u	11	N,
ku	15	N, V

Verbs are a rather limited class, and it very seldom happens that a new verb occurs in text. If this happens, however, there are good indicators for a verb, because the inflection markers are before the verb stem.

To include the full inflection paradigm into the guesser is an excessively heavy process and hardly practical. It is enough that the subject prefix and the TAM marker are included for identifying that the word is a verb. These are compulsory elements in verbs, except for infinitive forms.

For verbs, we get the following set of indicators (2). The table contains only the most obvious combinations of prefixes.

(2)

S-Prefix	TAM	POS
ni SG1	na	V
u SG2	a	V
a SG3	ta	V
tu PL1	li	V
m PL2	me	V
wa PL3	ki	V
a 1	ka	V
wa 2	nge	V
u 3	nga	V
i 4	si	V
li 5		V
ya 6		V
ki 7		V
vi 8		V
ch 7		V
vy 8		V
i 9		V
zi 10		V
u 11		V
ku 15		V, N

Any of the subject prefix in column 1 can be combined with any of the TAM markers in column 2, except for the last one (ku), which is the marker of infinitive, and also the marker of noun, if the verb is used as noun.

Note that the list of the first two verb morphemes is not complete. It only includes the most likely sequences of characters in the beginning of verbs, and they are as such sufficient indicators of verbs for a guesser.

The word-level guesser can be constructed on the basis of the two above tables. Note that although the subject prefixes *ki* and *vi* would be ambiguous also in relation to adverbs, nouns and adjectives, this does not happen, because the compulsory TAM marker follows.

3 Implementing the verb finder

On the basis of the table (2), we can construct a Perl script, which identifies and marks unknown verbs (3).

(3)
```
perl -pe
's/^("\<\*?(ni|u|a|tu|m|wa|a|wa|u|i|li|ya|ki|vi|ch|vy|i|zi|u)(na|a
|ta|li|me|ki|ka|nge|nga|si)[a-z]+\>" \<Heur\>)/$1 V/gm'
```

After analysis we get the following kind of analysis (4).

(4)
```
"<ameona>" <Heur>
"<tumeona>" <Heur>
"<waliona>" <Heur>
"<ningeona>" <Heur>
"<ningona>" <Heur>
```

When we apply the rule (3), the result is in (5).

(5)
```
"<ameona>" <Heur> V
"<tumeona>" <Heur> V
"<waliona>" <Heur> V
"<ningeona>" <Heur> V
"<ningona>" <Heur>
```

We see that all forms, except the last one, were recognised as verbs. In this way we can find unknown verbs, but we cannot produce appropriate analysis tags. Such new verbs should be included into the morphological analyser.

4 Capital-initial words

Proper names are written with capital-initial letters in all positions in the sentence. If the word in a non-initial position is written with a capital-initial letter, it is very likely a proper name. In sentence-initial position all words are written with capital-initial letters. This makes the disambiguation problematic[2].

The safe approach is that any unknown word, which is written with a capital-initial letter, is given two possible interpretations, one for proper name and another for an ordinary word. Then, on the basis of context, the correct one is selected.

In (6) are two sentences, where an unknown word with capital-initial is in two different positions.

(6)
```
"<<s>>"
       "<s>" { <s> }
"<*magufuli>"
       "*magufuli" <Heur> PROPNAME { *magufuli } CAP
       "magufuli" <Heur> N 5/6-PL { magufuli } CAP
```

[2] The problem of sentence-initial proper names is discussed in Report No. 28, 2018.

```
"<ni>"
      "ni" V V-BE INIT { it , he , she } { is }
      "ni" V V-BE INIT { they } { are }
      "ni" V V-BE NOSUBJ { is , are , am }
"<rais>"
      "rais" N 9/6-SG HUM { the } { *president } MALE AR HUM
      "rais" N 1/2-SG { *president } MALE AN HUM CAP
"<siku>"
      "siku" N 9/10-SG { the } { day } TIME
      "siku" N 9/10-PL { the } { day } TIME
"<hizi>"
      "hizi" V IMP VFIN z [hizi] { disgrace , dishonour , insult }
SVO AR
      "hizi" V <kwisha z [hizi] { disgrace , dishonour , insult }
SVO AR
      "hizi" PRON DEM :hV 10-PL { these }
      "hizi" <kwisha z [hizi] { disgrace , dishonour , insult }
SVO AR
"<.$>"
      ".$" { .$ } **CLB
"<<s>>"
      "<s>" { <s> }
"<*nilimwona>"
      "ona" V 1-SG1-SP VFIN { *i } PAST 1-SG3-OBJ OBJ { him , her
} z [ona] { see , feel } SVO HUM-ACT CAP
"<*magufuli>"
      "*magufuli" <Heur> PROPNAME { *magufuli } CAP
      "magufuli" <Heur> N 5/6-PL { magufuli } CAP
"<jana>"
      "jana" N 9-SG { yesterday } TIME
      "jana" ADV { yesterday } PREFR TIME
"<.$>"
      ".$" { .$ } **CLB
```

In the latter sentence the word *Magufuli* is obviously a proper name, because it is inside the sentence. In the first sentence the decision cannot be made on the basis of its position. Both interpretations would be possible. The position of the word shows that it must be the subject. The postmodifier of the verb is a human being. Therefore, it is very likely that *Magufuli* is a proper name.

Note that the decision can be made only in the disambiguation process, as we see in (7).

(7)

```
"<<s>>"
      "<s>" { <s> }
"<*magufuli>"
      "*magufuli" <Heur> PROPNAME { *magufuli } CAP @SUBJ
"<ni>"
      "ni" V V-BE NOSUBJ { is } @FMAINVintr-def
"<rais>"
```

```
      "rais" N 1/2-SG { *president } MALE AN HUM CAP @<P
"<siku_hizi>"
      "siku_hizi" ADV { these days } @ADVL
      >MW
"<.$>"
      ".$" { .$ } **CLB
"<<s>>"
      "<s>" { <s> }
"<*nilimwona>"
      "ona" V 1-SG1-SP VFIN { *i } PAST 1-SG3-OBJ OBJ NO-OBJ-GLOSS
z [ona] { see } SVO HUM-ACT CAP PROP-CAND PROP-CAND @FMAINVtr+OBJ>
"<*magufuli>"
      "*magufuli" <Heur> PROPNAME { *magufuli } CAP @OBJ
"<jana>"
      "jana" ADV { yesterday } PREFR TIME @ADVL
"<.$>"
      ".$" { .$ } **CLB
```

5 Lower-case initial words

If the word is written with lower-case characters, its classification must be done using the features described in tables (1) and (2). For each word type we must give as many interpretation alternatives as the table shows.

In (8) we have examples of unknown words. The words themselves are not rare. They are in fact in the analysis system, but for the sake of demonstration they were temporarily removed from the system.

(8)

```
"<*hii>"
      "hii" CAP PRON DEM :hV 4-PL { these }
      "hii" CAP PRON DEM :hV 9-SG { this }

"<katiba>" <Heur>

"<nzuri>" <Heur>

"<inayopendekezwa>" <Heur>

"<inaondoa>" <Heur>

"<ukomo>" <Heur>

"<wa>"
      "wa" GEN-CON 3-SG { of }
      "wa" GEN-CON 11-SG { of }
      "wa" GEN-CON 1-SG { of }
      "wa" GEN-CON 2-PL { of }
```

```
"<muda>"
      "muda" N 3/4-SG { the } { time , time period } TIME

"<wa>"
      "wa" GEN-CON 3-SG { of }
      "wa" GEN-CON 11-SG { of }
      "wa" GEN-CON 1-SG { of }
      "wa" GEN-CON 2-PL { of }

"<kuhudumu>" <Heur>

"<kwa>"
      "kwa" PREP { with , for , to }
      "kwa" PREP { by , on , in }
      "kwa" PREP { from , at }
      "kwa" GEN-CON-KWA 15-SG { of }
      "kwa" GEN-CON-KWA 17-SG { of }

"<rais>"
      "rais" N 9/6-SG HUM { the } { *president } MALE AR HUM
      "rais" N 1/2-SG { *president } MALE AN HUM CAP

"<.$>"
      ".$" { .$ } **CLB
```

The words *katiba, nzuri, inayopendekezwa, inaondoa, ukomo*, and *kuhudumu* are unknown in the sentence. When we apply the guessing rules, we get the result as in (9).

(9)

```
"<<s>>"
      "<s>" { <s> }

"<*hii>"
      "hii" CAP PRON DEM :hV 4-PL { these }
      "hii" CAP PRON DEM :hV 9-SG { this }

"<katiba>"
      "katiba" <Heur> N 5/6-SG { katiba }
      "katiba" <Heur> N 9/10-SG { katiba }
      "katiba" <Heur> N 9/10-PL { katiba }
      "katiba" <Heur> A 9-SG { katiba }
      "katiba" <Heur> A 10-PL { katiba }

"<nzuri>"
      "nzuri" <Heur> N 9/10-SG { nzuri }
      "nzuri" <Heur> N 9/10-PL { nzuri }
      "nzuri" <Heur> A 9-SG { nzuri }
      "nzuri" <Heur> A 10-PL { nzuri }
```

```
"<inayopendekezwa>" <Heur> V

"<inaondoa>" <Heur> V

"<ukomo>"
      "ukomo" <Heur> N 11-SG { ukomo }

"<wa>"
      "wa" GEN-CON 3-SG { of }
      "wa" GEN-CON 11-SG { of }
      "wa" GEN-CON 1-SG { of }
      "wa" GEN-CON 2-PL { of }

"<muda>"
      "muda" N 3/4-SG { the } { time , time period } TIME

"<wa>"
      "wa" GEN-CON 3-SG { of }
      "wa" GEN-CON 11-SG { of }
      "wa" GEN-CON 1-SG { of }
      "wa" GEN-CON 2-PL { of }

"<kuhudumu>" <Heur> V INF

"<kwa>"
      "kwa" PREP { with , for , to }
      "kwa" PREP { by , on , in }
      "kwa" PREP { from , at }
      "kwa" GEN-CON-KWA 15-SG { of }
      "kwa" GEN-CON-KWA 17-SG { of }

"<rais>"
      "rais" N 9/6-SG HUM { the } { *president } MALE AR HUM
      "rais" N 1/2-SG { *president } MALE AN HUM CAP

"<.$>"
      ".$" { .$ } **CLB
```

We see that the words *katiba* and *nzuri* were assigned to noun or adjective. The word *katiba* has three class alternatives for noun, that is, class 5, 9 and 10. It also has two class alternatives as adjective, that is, class 9 and 10, but not class 5, because the class prefix for adjective in this class is different.

For the word *nzuri* there are the class alternatives 9 and 10 for noun and adjective alike.

The word *ukomo* was interpreted as noun in class 11. The words *inayopendekezwa*, *inaondoa* and *kuhudumu* were guessed as verbs, the last one in infinitive form.

The verbs, except for the last one, need much more morphological information so that they can be processed properly.

The non-verbs are ambiguous between noun and adjective, except for *ukomo*, which could be only a noun, because the adjective prefix in this class is different.

The disambiguation of the nouns with double or triple interpretation is performed in the conventional disambiguation process.

If there are many adjacent guessed words, as in our case above, it is difficult to disambiguate, because there are very few words around to rely on. Our example is artificial, and in normal development process such situations hardly occur.

When we disambiguate the sentence, the result is as in (10).

(10)

```
"<<s>>"
       "<s>" { <s> }
"<*hii>"
       "hii" CAP PRON DEM :hV 9-SG { this } @NDEM>
"<katiba>"
       "katiba" <Heur> N 9/10-SG { katiba } @<P
"<nzuri>"
       "nzuri" <Heur> A 9/10-SG { nzuri } @<P
"<inayopendekezwa>" <Heur> V
"<inaondoa>" <Heur> V
"<ukomo>"
       "ukomo" <Heur> N 11-SG { ukomo } @<P
"<wa>"
       "wa" GEN-CON 11-SG { of } @GCON
"<muda>"
       "muda" N 3/4-SG { the } { time } TIME @<NH
"<wa>"
       "wa" GEN-CON 3-SG { of } @GCON
"<kuhudumu>" <Heur> V INF @-FMAINV
"<kwa>"
       "kwa" PREP { with } @ADVL
"<rais>"
       "rais" N 1/2-SG { *president } MALE AN HUM CAP @P>
"<.$>"
       ".$" { .$ } **CLB
```

The word *nzuri* was interpreted as adjective, on the basis of its position in the sentence. Verbs are helplessly defective and cannot be used in this way. They must be included into the dictionary. However, even the fact that verbs are identified among other words helps in including them into the appropriate place in the dictionary.

Note also that for the noun candidates, the word as such was copied as gloss. For verbs the gloss was not given, because it is assumed that verbs must be included into the dictionary in any case and the gloss will be given in that connection.

The solution described above is temporary also for guessed nouns, because they have no proper gloss. They also should be included into the dictionary.

When we add the unknown words to the dictionary, we get the disambiguated result as in (11).

(11)

```
"<<s>>"
	"<s>" { <s> }
"<*hii>"
	"hii" CAP PRON DEM :hV 9-SG { this } @NDEM>
"<katiba>"
	"katiba" N 9/10-SG { the } { constitution } AR @<P
"<nzuri>"
	"zuri" ADJ A-INFL 9-SG { good } @<NADJ
"<inayopendekezwa>"
	"pendekezwa" V 9-SG-SP VFIN NO-SP-GLOSS PR:na 9-SG-REL {
which } z [pendekeza] { suggest } SVO STAT PREFR CAUS PASS SUB-REL
@FMAINVtr-OBJ>
"<inaondoa>"
	"ondoa" V 9-SG-SP VFIN { it } PR:na z [ondoa] { remove } SVO
@FMAINVtr+OBJ>
"<ukomo>"
	"ukomo" N 11-SG { limit } @OBJ
"<wa>"
	"wa" GEN-CON 11-SG { of } @GCON
"<muda>"
	"muda" N 3/4-SG { the } { time } TIME @<NH
"<wa>"
	"wa" GEN-CON 3-SG { of } @GCON
"<kuhudumu>"
	"hudumu" N 15-SG z [hudumu] { serve } SVO AR @<NH
"<kwa>"
	"kwa" GEN-CON-KWA 15-SG { of } @-FMAINV-n
"<rais>"
	"rais" N 1/2-SG { *president } MALE AN HUM CAP @<P
"<.$>"
	".$" { .$ } **CLB
```

After further processing, we get the translation (12)

(12)
This good constitution which is suggested removes the limit of the time of serving of President.

We can test the guesser with some other words. In (13) are three sentences with the word *zuri* in different forms and positions. The prefix *ki* refers to class 7, and prefix *vi* to class 8.

(13)

```
"<<s>>"
	"<s>" { <s> }
"<*vitu>"
	"kitu" N 7/8-PL { the } { thing , object , money } CAP
"<vizuri>"
	"vizuri" <Heur> N 7/8-PL { vizuri }
```

```
      "vizuri" <Heur> A 8-PL { vizuri }
      "vizuri" <Heur> ADV { vizuri }
"<vinanipendeza>"
      "pendeza" V 8-PL-SP VFIN { they } PR:na 1-SG1-OBJ OBJ { me }
z [pendeza] { please , attract , charm } SVO PREFR :CAUS
"<.$>"
      ".$" { .$ } **CLB
"<<s>>"
      "<s>" { <s> }
"<*kitu>"
      "kitu" N 7/8-SG { the } { thing , object , money } CAP
"<kizuri>"
      "kizuri" <Heur> N 7/8-SG { kizuri }
      "kizuri" <Heur> A 7-SG { kizuri }
      "kizuri" <Heur> ADV { kizuri }
"<kinanipendeza>"
      "pendeza" V 7-SG-SP VFIN { it } PR:na 1-SG1-OBJ OBJ { me } z
[pendeza] { please , attract , charm } SVO PREFR :CAUS
"<.$>"
      ".$" { .$ } **CLB
"<<s>>"
      "<s>" { <s> }
"<*ninakuona>"
      "ona" V 1-SG1-SP VFIN { *i } PR:na 1-SG2-OBJ OBJ { you } z
[ona] { see , feel } SVO HUM-ACT CAP
      "ona" V 1-SG1-SP VFIN { *i } PR:na 15-SG-OBJ OBJ { it } z
[ona] { see , feel } SVO HUM-ACT CAP
      "ona" V 1-SG1-SP VFIN { *i } PR:na 17-SG-OBJ OBJ { then } z
[ona] { see , feel } SVO HUM-ACT CAP
      "ona" V 1-SG1-SP VFIN { *i } PR:na 17-SG-OBJ OBJ { there } z
[ona] { see , feel } SVO HUM-ACT CAP
"<vizuri>"
      "vizuri" <Heur> N 7/8-PL { vizuri }
      "vizuri" <Heur> A 8-PL { vizuri }
      "vizuri" <Heur> ADV { vizuri }
"<.$>"
      ".$" { .$ } **CLB
```

In all positions and in both forms the word *zuri* has three interpretations. In the first sentence, the word must be adjective, because the preceding word is noun and the following word is finite verb.

The second sentence differs from the first one only in that the subject is in singular, and the solution is the same as in the first sentence.

In the third sentence, the word *vizuri* follows the verb. It could be a noun object, but because the object is already encoded in the verb (-ku-), it cannot be an object. It cannot be a postmodifier of the verb either, because it has no preposition. Therefore, the obvious solution is that it is an adverb.

The disambiguated result is in (14).

(14)

```
"<<s>>"
       "<s>" { <s> }
"<*vitu>"
       "kitu" N 7/8-PL { the } { *thing } CAP @SUBJ
"<vizuri>"
       "vizuri" <Heur> A 8-PL { vizuri }
"<vinanipendeza>"
       "pendeza" V 8-PL-SP VFIN NO-SP-GLOSS PR:na 1-SG1-OBJ OBJ {
me } z [pendeza] { please } SVO PREFR CAUS @FMAINVtr-OBJ>
"<.$>"
       ".$" { .$ } **CLB
"<<s>>"
       "<s>" { <s> }
"<*kitu>"
       "kitu" N 7/8-SG { the } { *thing } CAP @SUBJ
"<kizuri>"
       "kizuri" <Heur> A 7-SG { kizuri }
"<kinanipendeza>"
       "pendeza" V 7-SG-SP VFIN NO-SP-GLOSS PR:na 1-SG1-OBJ OBJ {
me } z [pendeza] { please } SVO PREFR CAUS @FMAINVtr-OBJ>
"<.$>"
       ".$" { .$ } **CLB
"<<s>>"
       "<s>" { <s> }
"<*ninakuona>"
       "ona" V 1-SG1-SP VFIN { *i } PR:na 1-SG2-OBJ OBJ NO-OBJ-
GLOSS z [ona] { see } SVO HUM-ACT CAP @FMAINVtr+OBJ>
"<vizuri>"
       "vizuri" <Heur> ADV { vizuri }
"<.$>"
       ".$" { .$ } **CLB
```

We cannot get a translation of the sentences unless we include the unknown words into the lexicon. Yet the disambiguated unknown words help in classifying the unknown words, and this makes the work much easier.

After including the missing words into the lexicon, we get the translation (15).

(15)
The good things please me.
The good thing pleases me.
I see you well.

6 Summary

The report shows that such words that are not included into the morphological lexicon can be handled in two phases. The features of the words give hints to analysis, but they are only hints. On the morphological level, the word is given all possible interpretations in regard to POS and noun class. These ambiguous readings are then disambiguated on the basis of context. The ambiguity can often be resolved with this method, especially if there are not more than one or two unknown word in the sentence. However, for getting proper translation, the unknown words should be included into the lexicon. When the unknown words are classified using the guesser, the inclusion of the words into the lexicon is easy.