Precision in post-correction of annotated corpus 1

Technical Report · February 2023

DOI: 10.13140/RG.2.2.33247.94888

Arvi Hurskainen
University of Helsinki
59 PUBLICATIONS 199 CITATIONS

Technical Reports on Language Technology
Report No 95 2023
http://www.njas.helsinki.fi/salama

Precision in post-correction of annotated corpus[1]

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

Abstract

It is common that corpus annotation contains mistakes that need to be corrected. There are basically two methods for doing this. One solution is to correct the tagger itself and run the tagging process again. Another solution is to correct only those readings that need correction. In this report I discuss such a case, where the corpus was tagged using a tagging program, and after that several corrections were made either manually or by using correction scripts. As a result, the corpus was considered properly tagged. However, later it was found that there was a commonly occurring verb form, which was not recognised by the tagger in part of the corpus, and it was interpreted as a noun, using a heuristic post-tagger.

It was not feasible to re-analyse the corpus using the corrected analyser, because in that solution the post-processing corrections would be lost. There was a need to correct only those readings, where the mistakes occur and leave other readings intact.

The corpus that I discuss here is Helsinki Corpus of Swahili 2.0 containing 25 million words. The construction of the correction script in this case is not a simple task, because Swahili verbs inflect to both directions, using sequences of prefixes and suffixes. In addition, the noun class system multiplies the number of possible verb forms of each verb.

I will demonstrate phase by phase how the correction script can be constructed.

Key Words: *morphological analysis, corpus annotation.*

1 Introduction

The Language processing tools are generally defective. The best of them perform remarkably well, but it is common that the analyser makes some irritating mistakes. Although we try to correct the analyser, the language is so complex that mistakes cannot be avoided. Often the mistakes are related to defective tokenisation and such words that the analyser does not recognize. Such mistakes cannot perhaps be totally avoided. A more serious case is such a mistake type that is related to the inflection of the words.

This report concerns the correction of a particular verb form in Helsinki Corpus of Swahili 2.0.[2] In a large part of the corpus text, the analyser did not recognize such verb forms that

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

[2] **http://urn.fi/urn:nbn:fi:lb-201608301**

included the narrative case, which is marked with the *-ka-* prefix. It is not known how this bug entered the code in the processing phase. Since the analysed corpus went through heavy post-processing routines, both manually and by using various correction scripts, it was not feasible to re-analyse the corpus using the corrected analyser.

I took another approach, which resembles a medical operation, where a curing ingredient is sent to a certain part of the body without manually touching the target, or even without seeing it. Yet my task was different in that instead of a single operation, the script should handle correctly hundreds of thousands different types of cases.

2 Phases in constructing the correction script

I will describe below the phases of the construction of the correction script. I will use in the demonstration a set of various types of cases. The cases are extracted from the old version of the corpus (1).

(1)

```
a. ikawa    ikawa N      Heur 9/10-SG     { ikawa }   @<P
b. tukalazimika   tukalazimika      N      Heur 9/10-SG     {
tukalazimika }    @<P
c. tukawaeleze    tukawaeleze N      Heur 9/10-SG     {
tukawaeleze }     @SUBJ
d. tukayaboresha  tukayaboresha     N      Heur 9/10-SG     {
tukayaboresha }   @OBJ
e. akamteua akamteua    N      Heur 9/10-SG     { akamteua }
      @SUBJ
f. zikatumika     zikatumika  N      Heur 9/10-SG     { zikatumika
}     @OBJ
g. yakawanufaisha yakawanufaisha    N      Heur 9/10-SG     {
yakawanufaisha }  @<P
h. likaangaliwe   likaangaliwe      N      Heur 9/10-SG     {
likaangaliwe }    @OBJ
i. ukatarajia     ukatarajia  N      Heur 11-SG { ukatarajia }
      @OBJ
j. wakajinunua    wakajinunua N      Heur 1/2-PL      {
wakajinunua }     @OBJ
```

We see that if the verb has a prefix *-ka-* it is not recognized by the analyser. Instead, it is interpreted as a noun, because normally verbs get an analysis, while new nouns are constantly introduced to the language. Because the word is unknown, the wordform is copied as a stem. The noun class is guessed on the basis of the prefix and the corresponding tag is added. The default class pair for new nouns is 9/10, and most unknown words were given that class tag. Only in (i) and (f) the tag is different. The gloss in English is within curly braces. We see that it is the copy of the word. The syntactic tag is `@SUBJ` or `@OBJ`, and in some cases `@<P`. All this was done on a shaky basis.

When we start to convert these readings into correct form, we first mark the readings in the corpus, so that, when we run the correction script, we only touch these readings and leave other readings intact.

In the first phase we mark the faulty lines by adding '&' in front of them. We also remove the wrong readings and give some new analysis (2).

(2)

```
&ikawa  ika wa  V  @FMAINV
&tukalazimika  tuka lazimika  V  @FMAINV
&tukawaeleze  tuka waeleze  V  @FMAINV
&tukayaboresha  tuka yaboresha  V  @FMAINV
&akamteua  aka mteua  V  @FMAINV
&zikatumika  zika tumika  V  @FMAINV
&yakawanufaisha  yaka wanufaisha  V  @FMAINV
&likaangaliwe  lika angaliwe  V  @FMAINV
&ukatarajia  uka tarajia  V  @FMAINV
&wakajinunua  waka jinunua  V  @FMAINV
```

We see in (2) that we have tried, by adding a single space after -*ka*-, to identify the point, where the verb stem starts. In some cases, it is correct, but not in such verb forms that also include the object prefix. We also know that the words are verbs in finite form. Therefore, readings are given the syntactic tag `@FMAINV`.

In the original corpus, tabs are used for separating sections. Because the use of tabs is not convenient in writing rules, I have replaced tabs with double spaces. They can be returned back to tabs later. Note that between the prefix cluster and the stem there is only one space.

Next we separate the object prefixes from the stem (3).

(3)

```
&ikawa  ika wa  V  @FMAINV
&tukalazimika  tuka lazimika  V  @FMAINV
&tukawaeleze  tuka +wa eleze  V  @FMAINV
&tukayaboresha  tuka +ya boresha  V  @FMAINV
&akamteua  aka +m teua  V  @FMAINV
&zikatumika  zika +tu mika  V  @FMAINV
&yakawanufaisha  yaka +wa nufaisha  V  @FMAINV
&likaangaliwe  lika angaliwe  V  @FMAINV
&ukatarajia  uka tarajia  V  @FMAINV
&wakajinunua  waka +ji nunua  V  @FMAINV
```

Now object prefixes are separated from the stem. Also an anchor '+' is added in front of the prefix, so that further processing becomes easier. However, we see that the word *zikatumika* is wrongly interpreted. It should be *zika tumika*. Mistakes such as this are corrected using suitable scripts.

In the next phase, we mark the verb stems by adding a colon ':' in front of the stem. This is done for making sure that when English glosses are added to the readings, the stem is clearly identified (4).

(4)

```
&ikawa  ika :wa  V  @FMAINV
&tukalazimika  tuka :lazimika  V  @FMAINV
&tukawaeleze  tuka +wa :eleze  V  @FMAINV
&tukayaboresha  tuka +ya :boresha  V  @FMAINV
&akamteua  aka +m :teua  V  @FMAINV
&zikatumika  zika :tumika  V  @FMAINV
&yakawanufaisha  yaka +wa :nufaisha  V  @FMAINV
&likaangaliwe  lika :angaliwe  V  @FMAINV
&ukatarajia  uka :tarajia  V  @FMAINV
&wakajinunua  waka +ji :nunua  V  @FMAINV
```

The stems are now separated from prefixes, but they still have suffixes, which affect the import of English glosses. Therefore, they must be converted into correct form. It is not easy to determine what the correct form in each case is. Basically, the correct form is the form that has no derivational suffixes. There are, however, also such derived forms that have been lexicalised and have a special meaning.

In (5) we modify the verb stems. There are two stems, *:eleze* and *:angaliwe*, which have the final *e* instead the standard vowel *a*. These are subjunctive forms, and the stem must be converted to standard form, with the final *a*. The stem *:angaliwe* has also a passive marker -*w*-, and it must be removed from the stem.

(5)

```
&ikawa  ika :wa  V  @FMAINV
&tukalazimika  tuka :lazimika  V  @FMAINV
&tukawaeleze  tuka +wa :eleza  V  @FMAINV
&tukayaboresha  tuka +ya :boresha  V  @FMAINV
&akamteua  aka +m :teua  V  @FMAINV
&zikatumika  zika :tumika  V  @FMAINV
&yakawanufaisha  yaka +wa :nufaisha  V  @FMAINV
&likaangaliwe  lika :angalia  V  @FMAINV
&ukatarajia  uka :tarajia  V  @FMAINV
&wakajinunua  waka +ji :nunua  V  @FMAINV
```

Now when we have separated verb components and converted the stems into correct form, we can convert these components into tags. First, we convert the subject prefix and the narrative prefix into tags and move them to the correct place (6).

(6)

```
&ikawa  :wa  V  SUB-PREF=9-SG TAM=NARR:ka  @FMAINV
&tukalazimika  :lazimika  V  SUB-PREF=1-PL1 TAM=NARR:ka  @FMAINV
&tukawaeleze  +wa :eleza  V  SUB-PREF=1-PL1 TAM=NARR:ka  @FMAINV
&tukayaboresha  +ya :boresha  V  SUB-PREF=1-PL1 TAM=NARR:ka
@FMAINV
&akamteua  +m :teua  V  SUB-PREF=1-SG3 TAM=NARR:ka  @FMAINV
&zikatumika  :tumika  V  SUB-PREF=10-PL TAM=NARR:ka  @FMAINV
&yakawanufaisha  +wa :nufaisha  V  SUB-PREF=6-PL TAM=NARR:ka
@FMAINV
&likaangaliwe  :angalia  V  SUB-PREF=5-SG TAM=NARR:ka  @FMAINV
```

```
&ukatarajia  :tarajia  V  SUB-PREF=1-SG2 TAM=NARR:ka  @FMAINV
&wakajinunua  +ji :nunua  V  SUB-PREF=1-PL3 TAM=NARR:ka  @FMAINV
```

The subject prefixes and narrative prefixes have now been converted into respective tags and moved to the right. The object prefix still needs to be converted into the tag (7).

(7)

```
&ikawa  :wa  V  SUB-PREF=9-SG TAM=NARR:ka  @FMAINV
&tukalazimika  :lazimika  V  SUB-PREF=1-PL1 TAM=NARR:ka  @FMAINV
&tukawaeleze  :eleza  V  SUB-PREF=1-PL1 TAM=NARR:ka OBJ-
PREF=PL2/PL3  @FMAINV
&tukayaboresha  :boresha  V  SUB-PREF=1-PL1 TAM=NARR:ka OBJ-
PREF=6-PL  @FMAINV
&akamteua  :teua  V  SUB-PREF=1-SG3 TAM=NARR:ka OBJ-PREF=SG3
@FMAINV
&zikatumika  :tumika  V  SUB-PREF=10-PL TAM=NARR:ka  @FMAINV
&yakawanufaisha  :nufaisha  V  SUB-PREF=6-PL TAM=NARR:ka OBJ-
PREF=PL2/PL3  @FMAINV
&likaangaliwe  :angalia  V  SUB-PREF=5-SG TAM=NARR:ka  @FMAINV
&ukatarajia  :tarajia  V  SUB-PREF=1-SG2 TAM=NARR:ka  @FMAINV
&wakajinunua  :nunua  V  SUB-PREF=1-PL3 TAM=NARR:ka OBJ-PREF=REFL
@FMAINV
```

Now all verb prefixes have been converted to tags and moved to their respective places. Note that verb prefix tags have been separated by a single space. This ensures that the tags will be kept in the same slot when the strings will be converted to xml-format.

There are still two verbs with subjunctive form that need to be marked with a tag. These are the verbs *tukawaeleze* and *likaangaliwe* (8).

(8)

```
&ikawa  :wa  V  SUB-PREF=9-SG TAM=NARR:ka  @FMAINV
&tukalazimika  :lazimika  V  SUB-PREF=1-PL1 TAM=NARR:ka  @FMAINV
&tukawaeleze  :eleza  V  SBJN SUB-PREF=1-PL1 TAM=NARR:ka OBJ-
PREF=PL2/PL3  @FMAINV
&tukayaboresha  :boresha  V  SUB-PREF=1-PL1 TAM=NARR:ka OBJ-
PREF=6-PL  @FMAINV
&akamteua  :teua  V  SUB-PREF=1-SG3 TAM=NARR:ka OBJ-PREF=SG3
@FMAINV
&zikatumika  :tumika  V  SUB-PREF=10-PL TAM=NARR:ka  @FMAINV
&yakawanufaisha  :nufaisha  V  SUB-PREF=6-PL TAM=NARR:ka OBJ-
PREF=PL2/PL3  @FMAINV
&likaangaliwe  :angalia  V  SBJN SUB-PREF=5-SG TAM=NARR:ka
@FMAINV
&ukatarajia  :tarajia  V  SUB-PREF=1-SG2 TAM=NARR:ka  @FMAINV
&wakajinunua  :nunua  V  SUB-PREF=1-PL3 TAM=NARR:ka OBJ-PREF=REFL
@FMAINV
```

In the next phase we add English glosses to the stems (9).

(9)

```
&ikawa  :wa { be }  V  SUB-PREF=9-SG TAM=NARR:ka  @FMAINV
&tukalazimika  :lazimika { be forced }  V  SUB-PREF=1-PL1
TAM=NARR:ka  @FMAINV
&tukawaeleze  :eleza { explain }  V  SBJN SUB-PREF=1-PL1
TAM=NARR:ka OBJ-PREF=PL2/PL3  @FMAINV
&tukayaboresha  :boresha { improve }  V  SUB-PREF=1-PL1
TAM=NARR:ka OBJ-PREF=6-PL  @FMAINV
&akamteua  :teua { appoint }  V  SUB-PREF=1-SG3 TAM=NARR:ka OBJ-
PREF=SG3  @FMAINV
&zikatumika  :tumika { be used }  V  SUB-PREF=10-PL TAM=NARR:ka
@FMAINV
&yakawanufaisha  :nufaisha { profit }  V  SUB-PREF=6-PL
TAM=NARR:ka OBJ-PREF=PL2/PL3  @FMAINV
&likaangaliwe  :angalia { look at }  V  SBJN SUB-PREF=5-SG
TAM=NARR:ka  @FMAINV
&ukatarajia  :tarajia { hope }  V  SUB-PREF=1-SG2 TAM=NARR:ka
@FMAINV
&wakajinunua  :nunua { buy }  V  SUB-PREF=1-PL3 TAM=NARR:ka OBJ-
PREF=REFL  @FMAINV
```

The English tags are surrounded with curly braces. They need to be moved to the right to the appropriate place (10).

(10)

```
&ikawa  :wa  V  SUB-PREF=9-SG TAM=NARR:ka  { be }  @FMAINV
&tukalazimika  :lazimika  V  SUB-PREF=1-PL1 TAM=NARR:ka  { be
forced }  @FMAINV
&tukawaeleze  :eleza  V  SBJN SUB-PREF=1-PL1 TAM=NARR:ka OBJ-
PREF=PL2/PL3  { explain }  @FMAINV
&tukayaboresha  :boresha  V  SUB-PREF=1-PL1 TAM=NARR:ka OBJ-
PREF=6-PL  { improve }  @FMAINV
&akamteua  :teua  V  SUB-PREF=1-SG3 TAM=NARR:ka OBJ-PREF=SG3  {
appoint }  @FMAINV
&zikatumika  :tumika  V  SUB-PREF=10-PL TAM=NARR:ka  { be used }
@FMAINV
&yakawanufaisha  :nufaisha  V  SUB-PREF=6-PL TAM=NARR:ka OBJ-
PREF=PL2/PL3  { profit }  @FMAINV
&likaangaliwe  :angalia  V  SBJN SUB-PREF=5-SG TAM=NARR:ka  { look
at }  @FMAINV
&ukatarajia  :tarajia  V  SUB-PREF=1-SG2 TAM=NARR:ka  { hope }
@FMAINV
&wakajinunua  :nunua  V  SUB-PREF=1-PL3 TAM=NARR:ka OBJ-PREF=REFL
{ buy }  @FMAINV
```

Now when all tags have been added, we can convert the result into the form, where it should be in the corpus (11).

(11)

```
ikawa wa    V       SUB-PREF=9-SG TAM=NARR:ka    { be }
      @FMAINV
tukalazimika      lazimika    V      SUB-PREF=1-PL1 TAM=NARR:ka
      { be forced }     @FMAINV
tukawaeleze eleza       V      SBJN SUB-PREF=1-PL1 TAM=NARR:ka
OBJ-PREF=PL2/PL3  { explain } @FMAINV
tukayaboresha     boresha     V      SUB-PREF=1-PL1 TAM=NARR:ka
OBJ-PREF=6-PL     { improve } @FMAINV
akamteua    teua  V      SUB-PREF=1-SG3 TAM=NARR:ka OBJ-PREF=SG3
      { appoint } @FMAINV
zikatumika  tumika      V      SUB-PREF=10-PL TAM=NARR:ka   { be
used }      @FMAINV
yakawanufaisha    nufaisha    V      SUB-PREF=6-PL TAM=NARR:ka
OBJ-PREF=PL2/PL3  { profit }  @FMAINV
likaangaliwe      angalia     V      SBJN SUB-PREF=5-SG
TAM=NARR:ka { look at } @FMAINV
ukatarajia  tarajia     V      SUB-PREF=1-SG2 TAM=NARR:ka   { hope
}     @FMAINV
wakajinunua nunua       V      SUB-PREF=1-PL3 TAM=NARR:ka OBJ-
PREF=REFL   { buy }     @FMAINV
```

The anchors '&' and ':' have been removed and the double spaces have been converted to tabs. This format meets precisely the format of the original corpus.

3 The procedure in correcting the corpus

Now when the correction script is ready, we must consider the optimal method of performing the actual corrections. If the corpus would be a single file, it would be easy to take the corpus as input, run the script, and output it as a corrected file. However, the corpus consists of 454 files, and the file names must be retained precisely in the original form. Therefore, we must run each file separately.

This can be done using two alternative methods. In one method, we handle each file separately, run the script, and output it with the same name to another directory. Using this method, we do not destroy the original file. For each file, the same operation is repeated, and the file names are changed accordingly.

In another method, we first prepare a file, where the command for each operation is listed. Using this method, we only need to copy each command at a time and move it to the prompt and run it. Also a macro can be constructed for making the work easier.

If, after correction, it turns out that the correction script needs improvement, the corrections can be made to the script, and the corpus will be run through the corrected script. It is now easier, because the commands saved to the *.bash_history* file can be copied and used as such. Also here, a macro speeds up the operation.

4 Test of the correction script

Below is a piece of the original corpus before and after running the correction script (12).

(12)

```
Kwa_hiyo    kwa_hiyo    ADV   _     { therefore }     @ADVL
,     ,     COMMA _     { , }
tukaangalie tukaangalie N      Heur 9/10-SG     { tukaangalie }
      @<P
kama  kama  ADV   _     { as }      @CS
hizi  hizi  PRON   DEM 10-PL  { these }   @<NDEM
fedha fedha N      9/10-PL    { money }   @SUBJ MASS
hazitoshi   tosha V      TAM=NEG-a SUB-PREF=10-PL [tosha]    {
suffice }   @FMAINVtr-OBJ>    CAUS SVO VFIN
,     ,     COMMA _     { , }
ni    ni    V      V-BE { are }      @FMAINVintr-def
vyema vyema ADV   _     { well }    @ADVL
zikaongezwa zikaongezwa N      Heur 9/10-SG     { zikaongezwa }
      @SUBJ
.     .     _     _     { . }
Anaweza     weza  V      SUB-PREF=1-SG3 TAM=PR:na [weza]    { can
}     @FMAINVtr-OBJ>    SVO VFIN
kuwatuma    tuma  V      NO-TO OBJ-PREF=2-PL3 [tuma]  { send }
      @-FMAINV-n  SVO INF
wataalam    mtaalam     N      1/2-PL     { expert }  @OBJ
wake  ake   PRON   POSS 2-PL SG3    { his }     @GCON
wakamletea  wakamletea  N      Heur 1/2-PL      { wakamletea }
      @<P
pale  pale  ADV   _     { there }   @FMAINVintr-loc   LOC-16
na    na    CC    _     { and }     @CC
kwa_sababu  kwa_sababu  CONJ  _     { because } @CS
anawaamini  amini V      SUB-PREF=1-SG3 TAM=PR:na OBJ-PREF=2-PL3
[amini]     { believe } @FMAINVtr+OBJ>    SVO VFIN
wakafanya   wakafanya   N      Heur 1/2-PL      { wakafanya }
      @OBJ
hivyo hivyo ADV   _     { so }      @ADVL
.     .     _     _     { . }
Kwa_hiyo    kwa_hiyo    ADV   _     { therefore }     @ADVL
,     ,     COMMA _     { , }
ningeomba   omba  V      SUB-PREF=1-SG1 TAM=COND:nge [omba] { ask
}     @FMAINVtr+OBJ>    SVO VFIN
hilo  hilo  PRON   DEM 5-SG    { this }    @ADVL
likaangaliwe      likaangaliwe      N      Heur 9/10-SG     {
likaangaliwe }    @OBJ
lisije      ja    V      TAM=SBJN SUB-PREF=5-SG NEG [ja]    { come
}     @FMAINVintr MONOSLB SV VFIN
likawa      likawa      N      Heur 9/10-SG     { likawa }  @<P
tunasema    sema  V      SUB-PREF=2-PL1 TAM=PR:na [sema]    { say
}     @FMAINVtr+OBJ>    SVO VFIN
njaa  njaa  N      9/10-SG    { hunger }  @SUBJ
hii   hii   PRON   DEM 9-SG    { this }    @<NDEM
ni    ni    V      V-BE NOSUBJ      { is }      @FMAINVintr-def
kwa   kwa   PREP  _     { for }     @ADVL
maeneo      eneo  N      5/6-PL     { area }    @<P
fulani      fulani      ADJ    A-UNINFL   { certain } @<NADJ
tu    tu    ADV   _     { only }    @FMAINVintr-def
```

```
na    na    CC    _     { and }     @CC
maeneo      eneo  N      5/6-PL     { area }    @<P
fulani      fulani      ADJ    A-UNINFL    { certain } @<NADJ
hamna hamna V      SUB-PREF=18-SG NEG C:na      { there is not }
      @FMAINVintr VFIN
njaa  njaa  N      9/10-SG    { hunger }  @OBJ
.     .     _     _     { . }
Kwa_hiyo    kwa_hiyo    ADV   _     { therefore }     @ADVL
,     ,     COMMA _     { , }
wasije      ja    V      TAM=SBJN SUB-PREF=2-PL3 NEG [ja]    { come
}     @FMAINVintr MONOSLB SV VFIN
wakarudi    wakarudi    N      Heur 1/2-PL      { wakarudi }
      @SUBJ
hapa  hapa  ADV   _     { here }    @FMAINVintr-loc   LOC-16
wakasema    wakasema    N      Heur 1/2-PL      { wakasema }
      @SUBJ
,     ,     COMMA _     { , }
mlisema     sema  V      SUB-PREF=2-PL2 TAM=PAST [sema]      { say
}     @FMAINVtr-OBJ>    SVO VFIN
fanyeni_haraka    fanya_haraka      V      IMP [fanya] IMP-PL2
      { hurry }   @FMAINVtr-OBJ>    SVO VFIN
,     ,     COMMA _     { , }
aah   aah   N      Heur 9/10-SG     { aah }     @OBJ
"<!>" "!"   { ! } <Heur>
```

After running the correction script, the result is as in (13).

```
Kwa_hiyo    kwa_hiyo    ADV   _     { therefore }     @ADVL
,     ,     COMMA _     { , }
tukaangalie angalia     V      SBJN SUB-PREF=1-PL1 TAM=NARR:ka
      { look at } @FMAINV
kama  kama  ADV   _     { as }      @CS
hizi  hizi  PRON   DEM 10-PL  { these }   @<NDEM
fedha fedha N      9/10-PL    { money }   @SUBJ MASS
hazitoshi   tosha V      TAM=NEG-a SUB-PREF=10-PL [tosha]    {
suffice }   @FMAINVtr-OBJ>    CAUS SVO VFIN
,     ,     COMMA _     { , }
ni    ni    V      V-BE { are }     @FMAINVintr-def
vyema vyema ADV   _     { well }    @ADVL
zikaongezwa ongeza      V      SUB-PREF=10-PL TAM=NARR:ka   { add
}     @FMAINV
.     .     _     _     { . }
Anaweza     weza  V      SUB-PREF=1-SG3 TAM=PR:na [weza]     { can
}     @FMAINVtr-OBJ>    SVO VFIN
kuwatuma    tuma  V      NO-TO OBJ-PREF=2-PL3 [tuma]  { send }
      @-FMAINV-n  SVO INF
wataalam    mtaalam     N      1/2-PL      { expert }  @OBJ
wake  ake   PRON   POSS 2-PL SG3     { his }     @GCON
wakamletea  leta  V      SUB-PREF=1-PL3 TAM=NARR:ka OBJ-PREF=SG3
      { bring }   @FMAINV
pale  pale  ADV   _     { there }   @FMAINVintr-loc   LOC-16
```

```
na      na      CC      _       { and }       @CC
kwa_sababu  kwa_sababu  CONJ  _       { because } @CS
anawaamini  amini V       SUB-PREF=1-SG3 TAM=PR:na OBJ-PREF=2-PL3
[amini]     { believe } @FMAINVtr+OBJ>     SVO VFIN
wakafanya   fanya       V       SUB-PREF=1-PL3 TAM=NARR:ka   { do }
      @FMAINV
hivyo hivyo ADV   _       { so }        @ADVL
.      .      _      _       { . }
Kwa_hiyo    kwa_hiyo    ADV     _       { therefore }      @ADVL
,      ,      COMMA _       { , }
ningeomba   omba   V       SUB-PREF=1-SG1 TAM=COND:nge [omba] { ask
}      @FMAINVtr+OBJ>     SVO VFIN
hilo  hilo   PRON   DEM 5-SG    { this }     @ADVL
likaangaliwe      angalia      V       SBJN SUB-PREF=5-SG
TAM=NARR:ka { look at } @FMAINV
lisije      ja     V       TAM=SBJN SUB-PREF=5-SG NEG [ja]     { come
}      @FMAINVintr MONOSLB SV VFIN
likawa      wa     V       SUB-PREF=5-SG TAM=NARR:ka     { be }
      @FMAINV
tunasema    sema   V       SUB-PREF=2-PL1 TAM=PR:na [sema]     { say
}      @FMAINVtr+OBJ>     SVO VFIN
njaa  njaa   N       9/10-SG     { hunger }  @SUBJ
hii   hii    PRON   DEM 9-SG    { this }     @<NDEM
ni     ni     V       V-BE NOSUBJ      { is }       @FMAINVintr-def
kwa   kwa    PREP   _       { for }       @ADVL
maeneo      eneo  N       5/6-PL      { area }     @<P
fulani      fulani      ADJ     A-UNINFL    { certain } @<NADJ
tu     tu     ADV     _       { only }     @FMAINVintr-def
na     na     CC      _       { and }       @CC
maeneo      eneo  N       5/6-PL      { area }     @<P
fulani      fulani      ADJ     A-UNINFL    { certain } @<NADJ
hamna hamna V       SUB-PREF=18-SG NEG C:na       { there is not }
      @FMAINVintr VFIN
njaa  njaa   N       9/10-SG     { hunger }  @OBJ
.      .      _      _       { . }
Kwa_hiyo    kwa_hiyo    ADV     _       { therefore }      @ADVL
,      ,      COMMA _       { , }
wasije      ja     V       TAM=SBJN SUB-PREF=2-PL3 NEG [ja]    { come
}      @FMAINVintr MONOSLB SV VFIN
wakarudi    rudi   V       SUB-PREF=1-PL3 TAM=NARR:ka    { return }
      @FMAINV
hapa  hapa   ADV     _       { here }     @FMAINVintr-loc   LOC-16
wakasema    sema   V       SUB-PREF=1-PL3 TAM=NARR:ka    { say }
      @FMAINV
,      ,      COMMA _       { , }
mlisema     sema   V       SUB-PREF=2-PL2 TAM=PAST [sema]      { say
}      @FMAINVtr-OBJ>     SVO VFIN
fanyeni_haraka    fanya_haraka      V       IMP [fanya] IMP-PL2
      { hurry }   @FMAINVtr-OBJ>     SVO VFIN
,      ,      COMMA _       { , }
aah   aah    N       Heur 9/10-SG     { aah }      @OBJ
```

```
"<!>" "!"   { ! } <Heur>
```

All verbs with narrative form were converted to the form, where they should be in the corpus.

5 Conclusion

In this report I have shown a method for correcting re-occurring analysis mistakes in an annotated corpus. The correction measures are targeted precisely to the words with faulty analysis, and other words are left intact. The analysis thus achieved may de defective in such cases, where one or more morphemes have more than one interpretation. In such cases all alternatives are presented, whereby the analysis is under-specified. The approach described can be applied to any kinds of corpus correction needs.