Challenges of Language Technology of Kiswahili

Arvi Hurskainen

Abstract

The development of language technology of African languages has been relatively slow. This has also advantages, because it has been possible to make use of extensive research and testing that has been carried out elsewhere. By studying different approaches we are getting a clearer picture on whish approach or combination of approaches, would lead to optimal results. What seems optimal to English and other Indo-European languages is not necessarily suitable to Kiswahili and other African languages. This paper presents current trends in text-based language technology of Kiswahili and discusses their strengths and weaknesses.

1.0 Introduction

It is encouraging to see that there are new efforts to develop language technology of Kiswahili. There are also new approaches (de Pauw and de Schryver, 2008) that challenge the old rule-based methods that I have been developing for almost 30 years. Especially in machine translation, statistical methods dominate in the field. It is tempting to use statistical methods especially in machine translation, because they make it possible to use the computer to do the tedious job of sorting out what the text is about. One can also see promising results. For example, if you use Google Translate (GT), to translate the description of Tanzanian Government on the web page, the result is almost faultless. If we continue to test with other types of texts, such as fiction or various domain-specific texts, the results get worse. There are several ways to fail.

I made a small test with an arbitrarily chosen text of 400 words to see what types of mistakes GT makes. The text was an extract from the introductory text in launching the Kiswahili translation of Vatican II documents.

Table 1: Types of mistakes that GT made in translating Swahili text

Type of mistake	No. of errors	%
missing recall	18	4.5
wrong translation	20	5
translation missing	12	3
extra word(s)	7	1.75
wrong part-of-speech	8	2
wrong structure	3	0.75
wrong TAM	1	0.25
wrong SG/PL, person	3	0.75
wrong word order	1	0.25
Total	73	18.25

I also translated the same text with SALAMA. The results are in Table 2.

Table 2: Types of mistakes that SALAMA made in translating Swahili text

Type of mistake	No. of errors	%
missing recall	0	0
wrong translation	0	0
translation missing	0	0
extra word(s)	4	1
wrong part-of-speech	0	0
wrong structure	2	0.50
wrong TAM	0	0
wrong SG/PL, person	0	0
wrong word order	0	0
Total	6	1.50

These results reveal part of differences of the approaches. GT is obviously based mostly on statistical methods. It leaves words without translation or gives wrong translation, and in some cases omits translation altogether. Oddly enough, it sometimes adds translation to words that are not in original text. Disambiguation problems such as resolving the correct part-of-speech and singular/plural distinction are typical problems of statistical MT methods.

SALAMA, which uses rule-based methods, does not encounter unknown word-forms or leave words without translation. This is possible, because the lexicon of SALAMA is being continually updated. New words appearing in texts are added to the lexicon. The question of whether translation of a word is correct or not is problematic. Many words have near-synonyms, and the choice of the best

candidate in each context and domain is very hard to implement in rule-based approaches. Here statistical approaches usually give better results. Extra words in SALAMA translation are restricted to the choice of article (the, a and nil) as well as to double marking of subject and object, which exists in Swahili but not in English. The control of the latter is implemented, but due to imperfect syntactic mapping, doubling may sometimes occur in translated text.

In brief, statistical methods give fluent translation, but they are vulnerable when text contains complex or rare word-forms. Rule-based methods, if properly implemented, produce good coverage. Also verb-forms, even the very complex and rare ones, are recognized and translated. The advantage of rule-based methods is that any problem in translation process can be traced and corrected. The back side of rule-based methods is that, although the translation can be considered `correct,' fluency is not necessarily the best possible.

2.0 Morphological analysis

One of the first tasks in language technology is morphological analysis. Spelling checkers are normally constructed on the basis of these analyzers, especially if the language has complex morphology. A spelling checker does not need to contain much information. It only needs to check whether the word-form exists in language.

There is in the web a Swahili POS tagger (aflat.org), which could be used as a basis for a spelling checker. This tagger was trained using Helsinki Corpus of Swahili (HCS). An example of the performance of the tagger is in (1).

Source text:

> *Kwanza naomba kuchukua nafasi hii kumshukuru sana Mwenyekiti wa Kamati ya Bunge ya Ulinzi na Usalama Mhe. Dkt. John Samwel Malecela na wajumbe wa Kamati yake kwa jinsi ambavyo wamepitia na kuyachambua makadirio ya Bajeti ya Wizara ya Ulinzi na Jeshi la Kujenga Taifa na kuwasilisha taarifa kwenye Bunge lako tukufu.*

(1)

New Sentence	
Kwanza	ADV
naomba	V
kuchukua	V
nafasi	N
hii	PRON
kumshukuru	V
sana	AD-ADJ
Mwenyekiti	N

wa	GEN-CON
Kamati	N
ya	GEN-CON
Bunge	N
ya	GEN-CON
Ulinzi	N
na	CC
Usalama	N

We see that each word is given a POS tag. When encountering an unknown word, it figures out its POS on the basis of the morphological properties of the word and context. It fails to define sentence boundaries properly.

If we want to develop the tagger further to be useful for various kinds of applications in language technology, we have to add more features. The result has necessarily a lot of ambiguity. An example of such case is in (2). The above sentence was analyzed with SALAMA.

(2)

```
"<<s>>"
        "<s>" { <s> }
"<*kwanza>"
        "kwanza" NUM NUM-INFL ORD { first } CAP
        "kwanza" ADV { first } CAP
"<naomba>"
        "omba" V 1/2-SG1-SP VFIN { *i } PR:a [omba] { ask, ask for, beg for } SVO
        "omba" V 1/2-SG1-SP VFIN { *i } PR:a [omba] { pray } SVO
"<kuchukua>"
        "chukua" V INF { to } [chukua] { take,  withdraw, transport } SVO
        "chukua" N 15-SG [chukua] { take, withdraw, transport } SVO
        "chukua" V INF MOD-CAN [chukua] { take, withdraw, transport } SVO
        "chukua" V INF NO-TO [chukua] { take, withdraw  transport } SVO
"<nafasi>"
        "nafasi" N 9/10-SG { the } { opportunity, space, chance } AR
        "nafasi" N 9/10-PL { the } { opportunity, space, chance } AR
"<hii>"
        "hii" PRON DEM:hV 3/4-PL { these }
        "hii" PRON DEM:hV 9/10-SG { this }
"<kumshukuru>"
        "shukuru" V SBJN 15-SG-SP VFIN { it } 1/2-SG3-OBJ OBJ { him/her }
        [shukuru] { thank, praise } SVO AR
        "shukuru" V SBJN 17-SG-SP VFIN { there } 1/2-SG3-OBJ OBJ { him/her }
        [shukuru] { thank, praise } SVO AR
        "shukuru" V INF { to } 1/2-SG3-OBJ OBJ { him/her } [shukuru] { thank ,
        praise }
```

```
SVO AR
                "shukuru" N 15-SG 1/2-SG3-OBJ OBJ { him/her } [shukuru] { thank, praise }
SVO AR
                "shukuru" V INF MOD-CAN 1/2-SG3-OBJ OBJ { him/her } [shukuru] {thank,
                praise } SVO AR
                "shukuru" V INF NO-TO 1/2-SG3-OBJ OBJ {him/her } [shukuru] {thank,
                praise } SVO AR
                "<sana>"
                "sana" AD-ADJ AR { much, a lot }
                "sana" AD-ADJ AR { very }
"<*mwenyekiti>"
                "mwenyekiti" N 1/2-SG HUM { the } { chairman, chairperson } CAP
                "*mwenyekiti" N TITLE {*chairman } AN HUM
"<wa>"
                "wa" GEN-CON 3/4-SG { of }
                "wa" GEN-CON 11-SG { of }
                "wa" GEN-CON 1/2-SG { of }
                "wa" GEN-CON 1/2-PL { of }
"<*kamati>"
                "kamati" N 5/6-SG { the } { committee } ENG CAP
                "kamati" N 9/10-SG { the } { committee } ENG CAP
                "kamati" N 9/10-PL { the } { committee } ENG CAP
"<ya>"
                "ya" GEN-CON 3/4-PL { of }
                "ya" GEN-CON 9/10-SG { of }
                "ya" GEN-CON 5/6-PL { of }
                "ya" GEN-CON 6-PLSG { of }
"<*bunge>"
                "bunge" N 5/6-SG { the } { parliament } CAP
                "bunge" N 9/10-SG { the } { parliament } CAP
                "bunge" N 9/10-PL { the } { parliament } CAP
                "bunge" N 9/6-SG { the } { *m.*p. , *member of *parliament } MALE HUM
                CAP
"<ya>"          "ya" GEN-CON 3/4-PL { of }
                "ya" GEN-CON 9/10-SG { of }
                "ya" GEN-CON 5/6-PL { of }
                "ya" GEN-CON 6-PLSG { of }
                "<*ulinzi>"
                "ulinzi" N 11-SG { the } { defence } CAP
                "ulinzi" N 11/6-SG { the } DER:zi { guard, defence } CAP
"<na>"
                "na" CC { and }
                "na" AG-PART { by }
                "na" PREP { with }
                "na" NA-POSS { of }
                "na" ADV NOART { past }
                "na" ADV { also }
"<*usalama>" usalama" N 11-SG {the} {safety, security, work of: security officers} AR CAP
```

We see above that each word-form is given every possible interpretation, and context is not considered. This is the basic raw material for developing applications. The concept of ambiguity is intricate. If semantic features are included, ambiguity will be multiplied.

3.0 Disambiguation

The analyzed but not disambiguated text is crowded with many types of information. But there is nothing extra. Perhaps etymological tags could be removed without affecting performance.

Although there is research going on in various fields of Kiswahili language technology, I am not aware of disambiguation schemes of this language, except for the one I have been developing for more than 15 years (Hurskainen, 1996). My experience is that disambiguation is far more difficult to implement than morphological analyzer. There are in fact three separate tasks in disambiguation block. The first one is to isolate various types of multi-word expressions. In Kiswahili, MWEs include idioms where verb is often one member, collocations, multi-word terminology, most adjectives and various kinds of named entities (Hurskainen, 2006, 2007, 2008a).

The article on the subject in Wikipedia states:
"The most promising approach to the challenge of translating MWEs is example based MT, because in this case each MWE can be listed as an example with its translation equivalent in the target language. For rule based MT it would be too difficult to define rules to translate MWEs, due to the magnitude of different kinds of MWEs. Nevertheless, an example based MT system has to apply different rules for the translation of continuous and discontinuous MWEs as it is harder to identify a discontinuous MWE in a sentence where words are inserted between the different components of one MWE." (http://en.wikipedia.org/wiki/Multiword_expression).

Two comments must be made to this statement. First, a list of examples of MWE translations does not work in languages such as Kiswahili, which has a large quantity of verb-forms. The list would be almost endless. Second, the claim, that for rule-based MT it would be too difficult to define rules for translating MWEs, is not true. If there is a good dictionary that contains various kinds of MWEs with glosses in another language, such rules can be written automatically. For Kiswahili, I have written about 13,000 rules for translating MWEs. For English (MT from English to Swahili), there are even more rules. An important note is that not all MWEs are continuous strings. But also for discontinuous MWEs rules can be written.

Although rule writing can be automated, there is a tedious task left in testing that rules work correctly. Sometimes more constraints are needed, and in other times a rule must be relaxed to work properly in all cases.

The disambiguated sentence is in (3):

(3)

```
"<<s>>"
        "<s>" { <s> }
"<*kwanza>"
        "kwanza" ADV { first } CAP @ADVL
"<naomba>"
        "omba" V 1/2-SG1-SP VFIN { *i } PR:a [omba] { ask, ask for, beg for } SVO
        @FMAINVtr+OBJ>
"<kuchukua>"
        "chukua" V 15-SP [chukua] { take, withdraw, transport } SVO
"<nafasi>"
        "nafasi" N 9/10-SG { the } { opportunity, space, chance } AR @OBJ
"<hii>"
        "hii" PRON DEM :hV 9/10-SG { this }
"<kumshukuru>"
        "shukuru" V INF { to } 1/2-SG3-OBJ OBJ NO-OBJ-GLOSS [shukuru] { thank
        , praise } SVO AR @-FMAINV-n
"<sana>"
        "sana" AD-ADJ AR { much , a lot }
"<*mwenyekiti>"
        "mwenyekiti" N 1/2-SG HUM { the } { chairman, chairperson } CAP @OBJ
"<wa>"
        "wa" GEN-CON 1/2-PL { of } @GCON
"<*kamati>"
        "kamati" N 9/10-SG { the } { committee } ENG CAP @<GN
"<ya>"
        "ya" GEN-CON 9/10-SG { of } @GCON
"<*bunge>"
        "bunge" N 9/10-SG { the } { parliament } CAP @<GN
"<ya>"
        "ya" GEN-CON 9/10-SG { of } @GCON
"<*ulinzi>"
        "ulinzi" N 11-SG { the } { defence } CAP @<GN
"<na>"
        "na" CC { and } @CC
"<*usalama>"
        "usalama" N 11-SG { the } { safety, security, work of : security officers } AR
        CAP @<GN
```

We also see that in addition to disambiguation, words have been given syntactic tags, prefixed with @ especially subject and object tags which are important in controlling the translation of subject and object markers in verb.

4.0 Translating from English to Kiswahili

There are good computational grammars of English that can be made use of in MT from English to Kiswahili. For example, en-fdg of Connexor (www.csc.fi) can be used for giving reasonably good disambiguated and tagged analysis, including syntactic and dependency tags. Below I will demonstrate, using SALAMA, how we can convert English text into Kiswahili.

The English text is analyzed using en-fdg (4). In order to include several translation problems into the same sentence, the sentence is abnormally complicated:

(4)

1	hose	that		OBJ %NH PRON DEM PL
2	my	i	attr:>3	@A> %>N PRON PERS GEN SG1
3	three	three		@QN> %>N NUM CARD
4	good	good	cc:>6	@A> %>N A ABS
5	and	and	cc:>6	@CC %CC CC
6	expensive	expensive	attr:>7	@A> %>N A ABS
7	books	book	subj:>11	@SUBJ %NH N NOM PL
8	that	that	subj:>9	@SUBJ %NH <Rel> PRON
9	pleased	please	mod:>7	@+FMAINV %VA V PAST
10	students	student	obj:>9	@OBJ %NH N NOM PL
11	have	have	v-ch:>12	@+FAUXV %AUX V PRES
12	been	be	v-ch:>13	@-FAUXV %AUX EN
13	found	find	main:>0	@-FMAINV %VP EN
14	.	.		

For our purposes we can modify the output a bit:

(5)

```
"<<s>>"
        "<s>"
"<Those>"
        "that" %OBJ PRON DEM PL
"<my>"
        "i" %A> PRON PERS GEN SG1
"<three>"
        "three" %QN> NUM CARD
"<good>"
        "good" %A> A ABS
"<and>"
        "and" %CC CC
"<expensive>"
        "expensive" %A> A ABS
```

```
"<books>"
        "book" %SUBJ N NOM PL
"<that>"
        "that" %SUBJ <Rel> PRON
"<pleased>"
        "please" %+FMAINV V PAST
"<students>"
        "student" %OBJ N NOM PL
"<have>"
        "have" %+FAUXV V PRES
"<been>"
        "be" %-FAUXV EN
"<found>"
        "find" %-FMAINV EN
"<.>"
        "."
```

We add Kiswahili glosses. Note that some words get more than one gloss. Each noun also gets singular and plural prefix codes.

(6)

```
"<<s>>"
        "<s>"
"<Those>"
        "that" { le } %OBJ PRON DEM PL
"<my>"
        "i" { mimi } %A> PRON PERS GEN SG1
        "i" { angu } %A> PRON PERS GEN SG1
"<three>"
        "three" { tatu } INFL %QN> NUM CARD
"<good>"
        "good" { zuri } %A> A ABS
"<and>"
        "and" { na } %CC CC
"<expensive>"
        "expensive" { ghali } UNINFL %A> A ABS
"<books>"
        "book" { 7SG 8PL tabu } %SUBJ N NOM PL
        "book" { 9SG 10PL buku } %SUBJ N NOM PL
        "book" { 5SG 6PL buku } %SUBJ N NOM PL
"<that>"
        "that" { amba } %SUBJ <Rel> PRON
        "that" { NOGLOSS } %SUBJ <Rel> PRON
"<pleased>"
        "please" { pendezA } %+FMAINV V PAST
        "please" { taanisi } %+FMAINV V PAST
"<students>"
        "student" { 1SG 2PL anafunzi } %OBJ N NOM PL
        "student" { 1SG 2PL kurufunzi } %OBJ N NOM PL
        "student" { 1SG 2PL anachuo } %OBJ N NOM PL
```

```
"<have>"
        "have" { -wa na } %+FAUXV V PRES
        "have" { AUX } %+FAUXV V PRES
"<been>"
        "be" { wA } %-FAUXV EN
        "be" { kuwA } %-FAUXV EN
        "be" { AUX } %-FAUXV EN
"<found>"
        "find" { kutA } %-FMAINV EN
        "find" { patA } %-FMAINV EN
        "find" { okotA } %-FMAINV EN
"<.>"
        "."
```

Singular and plural prefixes in nouns are selected. This is done using contextual information.

(7)

```
"<<s>>"
        "<s>"
"<Those>"
        "that" { le } %OBJ PRON DEM PL
"<my>"
        "i" { mimi } %A> PRON PERS GEN SG1
        "i" { angu } %A> PRON PERS GEN SG1
"<three>"
        "three" { tatu } INFL %QN> NUM CARD
"<good>"
        "good" { zuri } %A> A ABS
"<and>"
        "and" { na } %CC CC
"<expensive>"
        "expensive" { ghali } UNINFL %A> A ABS
"<books>"
        "book" { 8PL tabu } %SUBJ N NOM PL
        "book" { 10PL buku } %SUBJ N NOM PL
        "book" { 6PL buku } %SUBJ N NOM PL
"<that>"
        "that" { amba } %SUBJ <Rel> PRON
        "that" { NOGLOSS } %SUBJ <Rel> PRON
"<pleased>"
        "please" { pendezA } %+FMAINV V PAST
        "please" { taanisi } %+FMAINV V PAST
"<students>"
        "student" { 2PL anafunzi } %OBJ N NOM PL
        "student" { 2PL kurufunzi } %OBJ N NOM PL
        "student" { 2PL anachuo } %OBJ N NOM PL
"<have>"
        "have" { -wa na } %+FAUXV V PRES
```

```
                "have" { AUX } %+FAUXV V PRES
"<been>"
                "be" { wA } %-FAUXV EN
                "be" { kuwA } %-FAUXV EN
                "be" { AUX } %-FAUXV EN
"<found>"
                "find" { kutA } %-FMAINV EN
                "find" { patA } %-FMAINV EN
                "find" { okotA } %-FMAINV EN
"<.>"
                "." 
```

Text is then disambiguated:

(8)

```
"<<s>>"
                "<s>"
"<Those>"
                "that" { le } %OBJ PRON DEM PL
"<my>"
                "i" { angu } %A> PRON PERS GEN SG1
"<three>"
                "three" { tatu } INFL %QN> NUM CARD
"<good>"
                "good" { zuri } %A> A ABS
"<and>"
                "and" { na } %CC CC
"<expensive>"
                "expensive" { ghali } UNINFL %A> A ABS
"<books>"
                "bo͝ok" { 8PL tabu } %SUBJ N NOM PL
"<that>"
                "that" { NOGLOSS } %SUBJ <Rel> PRON
"<pleased>"
                "please" { pendezA } %+FMAINV V PAST
"<students>"
                "student" { 2PL anafunzi } %OBJ N NOM PL
"<have>"
                "have" { AUX } %+FAUXV V PRES
"<been>"
                "be" { AUX } %-FAUXV EN
"<found>"
                "find" { kutA } %-FMAINV EN
"<.>"
                "." 
```

Morphological tags of Kiswahili words are added to each relevant word. Note that they are placed in the end of the reading in the order where they will be in the final word.

(9)

```
"<<s>>"
        "<s>"
"<Those>"
        "that" { le } %OBJ PRON DEM PL                  DEM-8
"<my>"
        "i" { angu } %A> PRON PERS GEN SG1              G-8
"<three>"
        "three" { tatu } INFL %QN> NUM CARD             NUM-8
"<good>"
        "good" { zuri } %A> A ABS                       A-8
"<and>"
        "and" { na } %CC CC
"<expensive>"
        "expensive" { ghali } UNINFL %A> A ABS
"<books>"
        "book" { 8PL tabu } %SUBJ N NOM PL
"<that>"
        "that" { NOGLOSS } %SUBJ <Rel> PRON
"<pleased>"
        "please" { pendezA } %+FMAINV V PAST    SP-8 TAM-li REL-8 OP-2
"<students>"
        "student" { 2PL anafunzi } %OBJ N NOM PL
"<have>"
        "have" { AUX } %+FAUXV V PRES
"<been>"
        "be" { AUX } %-FAUXV EN
"<found>"
        "find" { kutA } %-FMAINV EN             SP-8 TAM-me PASS
"<.>"
        "."
```

Morpheme tags are collected and attached to the stem. Plus sign indicates morpheme boundary:

(10)

```
"<<s>>"
        "<s>"
"<Those>"
        "that" { DEM-8+le } %OBJ PRON DEM PL
"<my>"
        "i" { G-8+angu } %A> PRON PERS GEN SG1
"<three>"
```

```
                "three" { NUM-8+tatu } INFL %QN> NUM CARD
"<good>"
                "good" { A-8+zuri } %A> A ABS
"<and>"
                "and" { na } %CC CC
"<expensive>"
                "expensive" { ghali } UNINFL %A> A ABS
"<books>"
                "book" { 8PL tabu } %SUBJ N NOM PL
"<that>"
                "that" { NOGLOSS } %SUBJ <Rel> PRON
"<pleased>"
                "please" { SP-8+TAM-li+REL-8+OP-2+pendezA } %+FMAINV V PAST SP-
8
"<students>"
                "student" { 2PL anafunzi } %OBJ N NOM PL
"<have>"
                "have" { AUX } %+FAUXV V PRES
"<been>"
                "be" { AUX } %-FAUXV EN
"<found>"
                "find" { SP-8+TAM-me+kut+w+A } %-FMAINV EN
"<.>"
                "."
```

Morphological tags are converted to surface form.

(11)

```
"<<s>>"
                "<s>"
"<Those>"
                "that" { vi+le } %OBJ PRON DEM PL
"<my>"
                "i" { vy+angu } %A> PRON PERS GEN SG1
"<three>"
                "three" { vi+tatu } INFL %QN> NUM CARD
"<good>"
                "good" { vi+zuri } %A> A ABS
"<and>"
                "and" { na } %CC CC
"<expensive>"
                "expensive" { ghali } UNINFL %A> A ABS
"<books>"
                "book" { vi+tabu } %SUBJ N NOM PL
"<that>"
                "that" { NOGLOSS } %SUBJ <Rel> PRON
"<pleased>"
                "please" { vi+li+vyo+wa+pendezA } %+FMAINV V PAST SP-8
"<students>"
```

```
                "student" { w+anafunzi } %OBJ N NOM PL
"<have>"
                "have" { AUX } %+FAUXV V PRES
"<been>"
                "be" { AUX } %-FAUXV EN
"<found>"
                "find" { vi+me+kut+w+A } %-FMAINV EN
"<.>"
                "."
```

For the purpose of re-ordering words, text is put to one line format.

(12)

```
( "<<s>>" ) ( "<s>" ) ( "<Those>" { vi+le } PRON DEM PL ) ( "<my>" { vy+angu } PRON
PERS GEN SG1 ) ( "<three>" { vi+tatu } INFL NUM CARD ) ( "<good>" { vi+zuri } A ABS )
( "<and>" { na } CC ) ( "<expensive>" { ghali } UNINFL A ABS ) ( "<books>" { vi+tabu } N
NOM PL ) ( "<that>" { NOGLOSS } <Rel> PRON ) ( "<pleased>" { vi+li+vyo+wa+pendezA }
V PAST SP-8 ) ( "<students>" { w+anafunzi } N NOM PL ) ( "<have>" { AUX } V PRES )
("<been>" { AUX } EN ) ( "<found>" { vi+me+kut+w+A } EN ) ( "<.>" "."
```

Reordering rules are applied.

(13)

```
( "<<s>>" ) ( "<s>" ) ( "<books>" { vi+tabu } N NOM PL ) ( "<my>" { vy+angu } PRON
PERS GEN SG1 ) ( "<good>" { vi+zuri } A ABS ) ( "<and>" { na } CC ) ( "<expensive>"
{ ghali } UNINFL A ABS ) ( "<Those>" { vi+le } PRON DEM PL ) ( "<three>" { vi+tatu }
INFL NUM CARD ) ( "<that>" { NOGLOSS } <Rel> PRON ) ( "<pleased>"
{ vi+li+vyo+wa+pendezA } V PAST SP-8 ) ( "<students>" { w+anafunzi } N NOM PL )
( "<have>" { AUX } V PRES ) ( "<been>" { AUX } EN ) ( "<found>" { vi+me+kut+w+A } EN
) ( "<.>" "."
```

Text is converted back to word-per-line format.

(14)

```
"<books>" { Vi+tabu } N NOM PL
"<my>" { vy+angu } PRON PERS GEN SG1
"<good>" { vi+zuri } A ABS
"<and>" { na } CC
"<expensive>" { ghali } UNINFL A ABS
"<Those>" { vi+le } PRON DEM PL
"<three>" { vi+tatu } INFL NUM CARD
"<that>" { NOGLOSS } <Rel> PRON
"<pleased>" { vi+li+vyo+wa+pendezA } V PAST SP-8
"<students>" { w+anafunzi } N NOM PL
"<have>" { AUX } V PRES
"<been>" { AUX } EN
"<found>" { vi+me+kut+w+A } EN
"<.>"           "."
```

Finally translated text is cleaned.

(15)
Vitabu vyangu vizuri na ghali vile vitatu vilivyowapendeza wanafunzi vimekutwa.

We shall see how GT translates the example sentence (16).

(16)
Wale wangu watatu nzuri na ghali vitabu ambavyo wanafunzi lilimpendeza hayajaonekana. (GT)

The translation is not comprehensible. In addition to failing to construct concordance and word order, it has serious problems in constructing verb forms. Both verbs are utterly wrong.

I do not deny that the example sentence is a bit complex and sounds artificial. Yet it contains only structures that are frequently used in normal language. Production of correct Kiswahili is difficult and it challenges statistical methods.

5.0 Statistical vs. Rule-based Approach

Each approach has its proponents and the developers seem to be addicted to their own approach. Statistical approaches in MT are attracting, because they promise reasonable results in a short period of time, and language independent methods and utilities can be used in training the system. The most labour-intensive part of the work is the compilation of a parallel corpus. Such one has been compiled also for Swahili - English (de Pauw, Wagacha and de Schryver, 2009). But it is very hard to compile such a corpus that has examples for each word-form.

According to general opinion, rule-based approaches produce good word coverage, but they require a lot of time and expertise in compiling the whole system. But the needed time is relative. If there is a good dictionary in digital form and the developer knows the grammar of the language, it is possible to construct a morphological analyzer within a few days. To make it fully faultless requires much more time, of course.

Perhaps the most important factor that supports the development of rule-based approach is that it produces comprehensive language resources, which can be made use of in developing many kinds of applications. It produces a dictionary that is more comprehensive than any of printed dictionaries, including full

morphological information. Applications already developed for Kiswahili on the basis of SALAMA include: extraction of domain-specific terms from a corpus (Sewangi, 2001), word sense disambiguation using machine learning (Ng'ang'a, 2005), the dictionary compiler based on a corpus, intelligent language learning system, and use of SALAMA in translating Bible to other Bantu languages (Hurskainen, 2008b, 2009a, 2009b, 2010, 2012a, 2012b). One should also add the evaluation of Kiswahili dictionaries. Evaluations have been carried out using two separate methods (Hurskainen, 1994, 2002, 2004; de Pauw, de Schryver and Wagacha, 2009).

6.0 Reflections written on SALAMA

A few comments are in place on what has been commented on SALAMA and its early phases. Akinyi and Matu have given wrong information when they write,

> "Also available are two spell checkers of Kiswahili. One developed by Lingsoft and the other was spearheaded by Professor Arvi Hurskainen, University of Helsinki to help editors who choose to write in Kiswahili" (Akinyi and Matu, 2011).

The spelling checker of Lingsoft, now integrated into Office, is the same one which was earlier distributed on CD by Lingsoft. The only difference is that the CD version includes also a hyphenation module, while the Office version does not. The spelling checker is entirely based on the morphological analyzer developed by Hurskainen.

SALAMA has been used in a number of research projects and its usefulness is widely recognized. Here is one example:

> "The isolated Swahili morphemes can more easily be linked to their English counterparts, since there will be more linguistic evidence in the parallel corpus, linking for example *ni* to I and *m* to him. To perform this kind of morphological analysis, we developed a machine learning system trained and evaluated on the Helsinki corpus of Swahili (Hurskainen, 2004). Experimental results show that the data-driven approach achieves state-of-the-art performance in a direct comparison with a rule-based method, with the added advantage of being robust to word forms for previously unseen lemmas" (de Pauw, Wagacha and de Schryver, 2009).

One can add that the version of SALAMA used in tagging Helsinki Corpus of Swahili (HCS) is now about 10 years old, and a lot of development has taken place since then. The question on how well the tagger guesses the morphological properties of unseen lemmas depends on how fine-grained the algorithm is. When the morphological analyzer becomes mature, unseen real words are seldom encountered. It is a good strategy to update the analyzer and

include those words into the system, and at the same time include full information, including glosses.

And a further extract:

> "We ... showed how a robust memory-based lemmatiser can be constructed on the basis of automatically annotated data. This research showed how previous rule based efforts can go hand in hand with a data-driven approach and help construct a more accurate lemmatiser that is inherently capable of analysing previously unseen word forms, even when the underlying lemma is unknown. The lemmatiser is currently being used as a pre-processing module in the context of machine translation for the language pair English—Swahili" (de Pauw and de Schryver, 2008).

Efforts to develop speech technology for Swahili include the text to speech system, developed by Ngugi, Okelo-Odongo and Wagacha (2005). Gelas, Besacier and Pellegrino (2012) have recently listed advances in Kiswahili localization.

"Morcover, Swahili being an impactful "vehicular" language of East Africa explains why many of mainstream IT services are already proposing localization in this language. Among others:

- Microsoft launched Swahili version of Microsoft Office and Windows in 2005.
- Wikipedia reached 23k articles in December 2011 (80th on 283 languages) after a launch in 2003. It is the first Bantu language and is second after Yoruba (30k articles) in the Niger-Congo family.
- Facebook Swahili version was launched in 2009 and was made by a group of scholars with the firm permission.
- Google also offers many of its services in Swahili: Google search interface in 2004, Google Translate since 2009, Text to speech, Gmail, Google Chrome and Google Maps in 2010, but not yet Voice Search ASR.
- there are initiatives for Swahili promotion over the web. This includes the following websites: the *Kamusi project* (the internet living Swahili dictionary) or *the one-stop Swahili portal goswahili.org* regrouping many resources on the language. It is also to be mentioned *the Kiswahili Linux Localization Project* (klnX) who made great efforts to localize free and open source software to the Swahili language (H. Gelas, L. Besacier and F. Pellegrino1, 2012).

7.0 Conclusion

Although the language technology of Kiswahili and other African languages has been developing slowly, important basic research and testing has been done. What worries me, however, is that the research and development has largely been in the hands of some individuals. There is no centre of language technology in any of the universities of Kiswahili-speaking countries. Also credible funds allocated to language technology are missing. While European Union has used hundreds of millions of Euros for developing LT for its languages, I am not aware of corresponding funds for Kiswahili and other African Union languages.

References

Akinyi J. J. and Matu P. M. (2011). "Challenges that face Kiswahili Usage in ICT in NEPAD Secondary Schools in Kenya." *International Journal of Academic Research in Business and Social Sciences*, October, 2011, Vol. 1, No. 3. ISSN: 2222-6990.

de Pauw G. and de Schryver G-M. (2008). "African Language Technology: The Data-Driven Perspective." LULCL II 2008 Proceedings of the Second Colloquium on Lesser Used Languages and Computer Linguistics Bozen-Bolzano, 13th-14th November 2008, pp. 79-94, Verena Lyding (Ed.)

de Pauw G., de Schryver G-M. & Wagacha P. W. (2009). *A Corpus-based Survey of Four Electronic Swahili–English Bilingual Dictionaries*. Lexikos 19 (AFRILEX-reeks/series 19: 2009): 340-352.

de Pauw G., Wagacha P. W. and de Schryver G-M. (2009). "The SAWA Corpus: a Parallel Corpus English - Swahili." Proceedings of the EACL 2009 Workshop on Language Technologies for African Languages – AfLaT 2009, pages 9–16, Athens, Greece, 31 March, 2009.

Gelas H., Besacier L. and Pellegrino F. (2012). "Development of Swahili resources for an automatic speech recognition system." http://www.ddl.ish-lyon.cnrs.fr/fulltext/Gelas/Gelas_2012_SLTU.pdf

Hurskainen A. (1994). *Kamusi ya Kiswahili Sanifu* in test: A computer system for analyzing dictionaries and for retrieving lexical data. Afrikanistische Arbeitspapiere 37: 169 179.

Hurskainen A. (1996). "Disambiguation of morphological analysis in Bantu languages." In *Proceedings of COLING* 96, The 16th International Conference on Computational Linguistics, Copenhagen 5 9.8. 1996. Pp. 568 573.

Hurskainen A. (2002). "Tathmini ya Kamusi Tano za Kiswahili (Computer Evaluation of Five Swahili Dictionaries)." *Nordic Journal of African Studies* 11(2): 283-300.

Hurskainen A. (2006). "Constraint Grammar in unconventional use: Handling complex Swahili idioms and proverbs." In A Man of Measure. Festschrift in Honour of Fred Karlsson on his 60th Birthday. Michael Suominen, Antti Arppe, Anu Airola, Orvokki Heinämäki, Matti Miestamo, Urho Määttä, Jussi Niemi, Kari K. Pitkänen and Kaius Sinnemäki (Editors). A Special Supplement to SKY Journal of linguistics. Turku: The Linguistic Association of Finland. Pp. 397- 406.

Hurskainen A. (2007). "Constraint Grammar in Unconventional Use: Describing Multiword Expressions." A speech in the Workshop on Constraint Grammar, in conjunction with the NODALIDA-2007 conference, held on May 23rd – 25,th 2007, Tartu (Estonia).

Hurskainen A. (2008a). Multiword Expressions and Machine Translation. *Technical Reports in Language Technology. Report No. 1, 2008* http://www.njas.helsinki.fi/salama

Hurskainen A. (2008b). SALAMA Dictionary Compiler - A Method for Corpus-Based Dictionary Compilation. *Technical Reports in Language Technology. Report No 2, 2008* http://www.njas.helsinki.fi/salama

Hurskainen A. (2009a). Intelligent Computer-Assisted Language Learning: Implementation to Swahili. *Technical Reports in Language Technology. Report No. 2, 2009* http://www.njas.helsinki.fi/salama

Hurskainen, A. (2004). "Computational testing of five Swahili dictionaries." A paper read in the 20th Scandinavian Conference of Linguistics, Helsinki,7-9.1.2004. http://www.ling.helsinki.fi/kielitiede/20scl/proceedings.shtml.

Hurskainen, A. (2009b). Machine translation of the Bible. *Technical Reports in Language Technology, Report No. 5, 2009* http://www.njas.helsinki.fi/salama

Hurskainen, A. (2010). Language learning system using language analysis and disambiguation. *Technical Reports in Language Technology, Report No. 9, 2010* http://www.njas.helsinki.fi/salama

Hurskainen, A. (2012a). Compounding in English to Swahili machine translation. *Technical Reports in Language Technology, Report No. 11, 2012* http://www.njas.helsinki.fi/salama

Hurskainen, A. (2012b). Multiword expressions in English to Swahili machine translation. *Technical Reports in Language Technology, Report No. 10, 2012* http://www.njas.helsinki.fi/salama

Ng'ang'a W. (2005). Word Sence Disambiguation of Swahili: Extending Swahili Language Technology with Machine Learning. Publications of the Department of General Linguistics 39, University of Helsinki.

Ngugi K., et al., (2005). "Swahili Text-to-Speech System." *African Journal of Science and Technology* (AJST) Science and Engineering Series Vol. 6, No. 1, pp. 80 – 89.

Sewangi, S. (2001). "Computer-Assisted Extraction of Terms in Specific Domain: The case of Kiswahili." Ph.D. Thesis. Publication of Institute for Asian and African Studies, University of Helsinki.