

## Adjectives in Swahili to English machine translation<sup>1</sup>

Arvi Hurskainen  
Department of World Cultures, Box 59  
FIN-00014 University of Helsinki, Finland  
[arvi.hurskainen@helsinki.fi](mailto:arvi.hurskainen@helsinki.fi)

### Abstract

In Bantu languages such as Swahili the part-of-speech category 'adjectives' is very limited. Only the most frequently used qualifying concepts are expressed using true adjectives. In machine translation we encounter the need to express all qualifying expressions of the source language in the way that is natural in target language. If the grammatical methods of expressing quality are different in these languages, the direct lexical mapping is not possible. In this paper I discuss such mapping problems in Swahili-to-English machine translation. Also I demonstrate methods for solving the mapping problems

**Key Words:** *machine translation, multiword expressions, disambiguation.*

### 1 Introduction

True adjectives in Swahili constitute only a fraction of adjectival expressions. Examples of original adjectives include *zuri* (good), *baya* (bad), *refu* (long), *fupi* (short), *zima* (healthy). These adjectives get the noun class prefix concerned when used in context. Another class of adjectives is formed of loanwords, usually from Arabic. Such adjectives do not usually take a class prefix.

Then there is a large number of qualifying concepts that need to be expressed in some way. These methods include: (a) the use of a genitive construction, e.g. *mtu wa asili* (aboriginal man); (b) the use of a possessive construction, e.g. *mtu mwenye uwezo* (able man); (c) the use of verb constructions with relative reference.

The problem in translation is that these expressions have also their original meanings. Therefore a careful control mechanism must be created for ensuring that the expressions are interpreted as adjectives only in appropriate contexts. Consider the example in (1).

(1) Huyu ni mtu mwenye uwezo

"<\*huyu>"

"huyu" PRON DEM :hV 1-SG { this } CAP

"<ni>"

"ni" V V-BE NOSUBJ { is }

"<mtu>"

---

<sup>1</sup> The report is issued under licence CC BY-NC

```
"mtu" N 1/2-SG HUM { the } { man }  
"<mwenye_uwezo>"  
  "enye_uwezo" ADJ { able }
```

The same word cluster has another meaning when the context changes. Consider the example in (2).

(2)

Huyu ni mtu mwenye uwezo wa kusoma.

```
"<*huyu>"  
  "huyu" PRON DEM :hV 1-SG { this } CAP  
"<ni>"  
  "ni" V V-BE NOSUBJ { is }  
"<mtu>"  
  "mtu" N 1/2-SG HUM { the } { man }  
"<mwenye>"  
  "enye" POSS-PRON :OTE 1-SG { with }  
"<uwezo>"  
  "uwezo" N 11-SG { ability }  
"<wa>"  
  "wa" GEN-CON 11-SG { of }  
"<kusoma>"  
  "soma" N 15-SG z [soma] { read } SVO
```

## 2 The implementation of true adjectives

The true adjectives can be listed in the morphological lexicon. The inflecting adjectives will be forced to take the appropriate class prefix, and the uninflecting adjectives are listed as such, without the possibility of taking a class prefix. Example (3) illustrates this.

(3)

```
"<chakula>"  
  "chakula" N 7/8-SG { the } { food } MASS  
"<kizuri>"  
  "zuri" ADJ A-INFL 7-SG { good }  
"<na>"  
  "na" CC { and }  
"<bure>"  
  "bure" ADJ A-UNINFL { free }
```

## 3 The implementation of multiword expressions

Adjectival expressions with genitive construction and possessive construction are treated as multiword expressions. Each word cluster with two or more members has to be described as consisting of a single unit. In the implementation discussed here, MWEs are isolated using Constraint Grammar (CG) rules. This is an environment with two advantages compared with direct re-writing. First, rules can be written using both surface features and grammatical tags. Because grammatical tags represent sets, it is possible to

write a rule, which applies for a number of cases. On the other hand, if sets prove too permissive, with the help of surface features over-application can be avoided. Another advantage in using CG rules is that the rule application can be controlled in many ways with the help of context. Below I will demonstrate how MWEs are isolated using the CG grammar parser.

Let us take the example of possessive construction in (1). The string *mtu mwenye uwezo* is first analyzed (4).

```
(4)
"<mtu>"
  "mtu" N 1/2-SG HUM { the } { man , :human being , person }
"<mwenye>"
  "enye" POSS-PRON :OTE 1-SG { with , possessor of }
  "enye" POSS-PRON :OTE 18-SG { in }
  "enye" POSS-PRON :OTE 1-SG { who }
  "enye" POSS-PRON :OTE 18-SG { which }
"<uwezo>"
  "uwezo" N 11-SG { ability , capability , capacity }
```

The word cluster *enye uwezo* is marked as a MWE (5).

```
(5)
"<mtu>"
  "mtu" N 1/2-SG HUM { the } { man , :human being , person }
"<mwenye>"
  "enye" POSS-PRON :OTE 1-SG { with , possessor of }
  "enye" POSS-PRON :OTE 18-SG { in }
  "enye" POSS-PRON :OTE 1-SG { who }
  "enye" POSS-PRON :OTE 18-SG { which }
"<uwezo>"
  "uwezo" ADJ <MW { able }
```

We see that the word *uwezo* is marked as a member of MWE. The tag <MW indicates that also the word before it belongs to the MWE cluster. Then the other member of MWE is marked as part of the cluster, and the original meaning analysis of the word is removed (6).

```
(6)
"<mtu>"
  "mtu" N 1/2-SG HUM { the } { man }
"<mwenye>"
  "enye" MW>
"<uwezo>"
  "uwezo" ADJ <MW { able }
```

The output is modified so that the MWE cluster constitutes a single lexical entry (7). Now the cluster *enye\_uwezo* can be handled as a normal adjective in translation.

(7)  
"<mtu>"  
    "mtu" N 1/2-SG HUM { the } { man }  
"<mwenye\_uwezo>"  
    "enye\_uwezo" ADJ { able }

Example (2), repeated here as (8), demonstrates that the word cluster *enye uwezo* cannot be an adjective in all contexts. The rule is not allowed to apply if the cluster is followed immediately by a genitive connector.

(8)  
"<mtu>"  
    "mtu" N 1/2-SG HUM { the } { man }  
"<mwenye>"  
    "enye" POSS-PRON :OTE 1-SG { with }  
"<uwezo>"  
    "uwezo" N 11-SG { ability }  
"<wa>"  
    "wa" GEN-CON 11-SG { of }  
"<kusoma>"  
    "soma" N 15-SG z [soma] { read } SVO

The rule is not either allowed to apply if the cluster is followed by an adjective (9).

(9)  
"<mtu>"  
    "mtu" N 1/2-SG HUM { the } { man }  
"<mwenye>"  
    "enye" POSS-PRON :OTE 1-SG { with }  
"<uwezo>"  
    "uwezo" N 11-SG { ability }  
"<mdogo>"  
    "dogo" ADJ A-INFL 11-SG { small }

The constraint does not, however, affect a longer rule, where the cluster *enye uwezo* is part of construction. Consider the example in (10). The word *kubwa* and two words before it form an adjective cluster.

(10)  
"<mtu>"  
    "mtu" N 1/2-SG HUM { the } { man , human being , person }  
"<mwenye>"  
    "enye" POSS-PRON :OTE 1-SG { with , possessor of }  
    "enye" POSS-PRON :OTE 18-SG { in }  
    "enye" POSS-PRON :OTE 1-SG { who }  
    "enye" POSS-PRON :OTE 18-SG { which }  
"<uwezo>"  
    "uwezo" N 11-SG { ability , capability , capacity }  
"<mkubwa>"

```
"kubwa" ADJ <<MW { puissant }
```

The two other members are marked as part of the cluster (11).

```
(11)
"<mtu>"
  "mtu" N 1/2-SG HUM { the } { man }
"<mwenye>"
  "enye" MW>>
"<uwezo>"
  "uwezo" MW><
"<mkubwa>"
  "kubwa" ADJ <<MW { puissant }
```

The members of the MWE are collected as a single entry (12).

```
(12)
"<mtu>"
  "mtu" N 1/2-SG HUM { the } { man }
"<mwenye_uwezo_mkubwa>"
  "enye_uwezo_kubwa" ADJ { puissant }
```

One must be careful in taking care of the correct order of rule application. The system is constructed so that the longer rule wins. In other words, if both of the clusters *enye uwezo* and *enye uwezo kubwa* are allowed to apply, double application should be prevented. The rule for *enye uwezo* would block the rule application, because the cluster is followed by an adjective. However, the rule for *enye uwezo kubwa* will apply, although there is an adjective. This rule applies only if the adjective is *kubwa*.

Another type of multiword expressions is the genitive structure. Consider the example in (13).

```
(13)
"<ruhusa>"
  "ruhusa" N 9/10-SG { the } { permission }
"<ya>"
  "ya" GEN-CON 9-SG { of }
"<kudumu>"
  "dumu" N 15-SG z [dumu] { last , be permanent } SV
```

This analysed form is disambiguated. In this process also the MWE is isolated (14).

```
(14)
"<ruhusa>"
  "ruhusa" N 9/10-SG { the } { permission }
"<ya_kudumu>"
  "ya_dumu" ADJ { permanent }
```

When this is further processed, we get the translation with correct word order (15).

(15)  
*The permanent permission*

#### 4 The implementation of relative verb structures

There are three types of relative verb structures that can be used for representing adjectives: (a) present tense structures, (b) past tense structures, and (c) negative structures.

##### 4.1 Present tense structures

The first example of present tense relative structures translates into English with gerund forms of verbs. In (16) the transfer is roughly: *the condition of air that changes > changing weather*.

(16)  
"<hali>"  
    "hali" N 9/10-SG { the } { state , condition , essence }  
"<ya>"  
    "ya" GEN-CON 9-SG { of }  
"<hewa>"  
    "hewa" N 9/10-SG { the } { air , weather , atmosphere }  
PLACE  
"<inayobadilika>"  
    "badilika" V 9-SG-SP VFIN { it } PR:na 9-SG-REL { which } z  
[badili] { change , substitute , adapt } SV STAT

The cluster hali ya hewa can be treated as a MQW, as in (17).

(17)  
"<hali\_ya\_hewa>"  
    "hali\_ya\_hewa" N 9/10-SG { the } { weather }  
"<inayobadilika>"  
    "badilika" V 9-SG-SP VFIN { it } PR:na 9-SG-REL { which } z  
[badili] { change , substitute , adapt } SV STAT

Then the verb form inayobadilika will be treated as an adjective (18).

(18)  
"<hali\_ya\_hewa>"  
    "hali\_ya\_hewa" N 9/10-SG { the } { weather }  
"<inayobadilika>"  
    ADJ ADJ-PR-REL { change }

The gloss *change* will be changed to gerund *changing*, and after reordering the constituents the result will be as in (19).

(19)  
*The changing weather*

The interpretation as gerund of present tense relative forms is only one interpretation. If the verb has a stative suffix with the meaning that an action is possible, the translation is different.<sup>2</sup>

If we want to change the example (16) into stative form meaning that the action is possible, we use the true stative form *badilishika* (120).

(20)  
"<hali\_ya\_hewa>"  
    "hali\_ya\_hewa" N 9/10-SG { the } { weather }  
"<inayobadilishika>"  
    ADJ ADJ-PR-REL-STAT { change }

The tag ADJ-PR-REL-STAT indicates that the construction must be treated as stative and translated as in (21).

(21)  
*The changeable weather*

#### 4.2 Past tense structures

Another group of relative verb constructions consists of past tense constructions. These can be divided into two main groups: stative constructions and passive constructions. Also constructions without these features exist. IN (22) is an example of a stative structure.

(22)  
"<hali>"  
    "hali" N 9/10-SG { the } { state , condition , essence }  
"<ya>"  
    "ya" GEN-CON 9-SG { of }  
"<hewa>"  
    "hewa" N 9/10-SG { the } { air , weather , atmosphere }  
PLACE  
"<iliyochafuka>"  
    "chafuka" V 9-SG-SP VFIN { it } PAST 9-SG-REL { which } z  
[chafua] { become dirty } SV STAT PFR

After disambiguation the result is as in (23).

---

<sup>2</sup> In fact also the verb form *badilika* is a stative form of the verb *badili*, but in this case the form is rather neutro-passive, without reference to possibility of action. Therefore it should not be treated as a true stative form.

(23)  
"<hali\_ya\_hewa>"  
    "hali\_ya\_hewa" N 9/10-SG { the } { weather }  
"<iliyochafuka>"  
    "chafuka" ADJ ADJ-REL { turbid }

The final translation is in (24).

(24)  
*The turbid weather*

Some adjectival expressions are formed with passive structures (25)

(25)  
"<hali>"  
    "hali" N 9/10-SG { the } { state , condition , essence }  
"<ya>"  
    "ya" GEN-CON 9-SG { of }  
"<hewa>"  
    "hewa" N 9/10-SG { the } { air , weather , atmosphere }  
PLACE  
"<iliyoahidiwa>"  
    "ahidiwa" V 9-SG-SP VFIN { it } PAST 9-SG-REL { which } z  
[ahidi] { promise , engage } SVO PASS

After disambiguation we get the result as in (26).

(26)  
"<hali\_ya\_hewa>"  
    "hali\_ya\_hewa" N 9/10-SG { the } { weather }  
"<iliyoahidiwa>"  
    ADJ ADJ-PAST-REL { promise }

When the adjective gloss is converted into the needed surface form, the result looks like in (27).

(27)  
*The promised weather*

### 4.3 Negative structures

The third type of relative structures with adjectival interpretation is the negative form. This form does not have a tense. To demonstrate this we use the above example (13) and change it into negative form (28).

(28)  
"<hali>"

```
"hali" N 9/10-SG { the } { state , condition , essence }
"<ya>"
"ya" GEN-CON 9-SG { of }
"<hewa>"
"hewa" N 9/10-SG { the } { air , weather , atmosphere }
PLACE
"<isiyobadilika>"
"badilika" V 9-SG-SP VFIN { it } NEG-REL 9-SG-REL { which }
z [badili] { change , substitute , adapt } SVO
```

After disambiguation we get the form as in (29).

```
(29)
"<hali_ya_hewa>"
"hali_ya_hewa" N 9/10-SG { the } { weather }
"<isiyobadilika>"
"badilika" ADJ ADJ-REL-NEG { invariable }
```

After further processing we get the final translation (30).

```
(30)
The invariable weather
```

## 5 Constraining inappropriate application of rules

We see above that in Swahili there is a number of methods for expressing meanings that are rendered with adjectives in English. I have shown examples of implementing various types of adjectival expressions. The major problem in implementation is that each structure has also the original meaning, and only in some contexts these structures have an adjectival meaning. We have seen above (2) that the possessive structure should not be followed immediately by a genitive connector or adjective.

Another major constraining factor is that each rule type should be constrained only to certain extended verb stems. Such extended forms in present tense include stative and non-stative, each type with distinct meaning type (18 and 20). Extended verb stems in past tense tend to either in stative or passive (23 and 26).

A further constraining factor is that each type appears in a distinct subset of verbs. Each subset was extracted from the English-Swahili dictionary, constructed for the English to Swahili translation system.

## 6 Alternative methods for writing rules for relative verb forms

Because global rules, which would apply to all verbs, would be too permissive, one has to consider methods for implementing needed constraints, so that each rule would apply only to appropriate cases.

Careful rule writing would mean that for each verb stem a rule is written with appropriate constraints. This method was used for negative relative forms. As a result there are hundreds of rules for this type of cases.

For implementing present tense and past tense structures, more abstract rules were written. Verb stems with each meaning type were collected as a set, and rules were written using the given set name. With this method it was possible to reduce the number of rules into four. These handle several hundreds of verb stems. This method works if constraints for each member of the set are the same. This seems to work.

The operation takes place in two phases. First an appropriate tag (ADL-REL, ADJ-REL-STAT, ADJ-PR-REL or ADJ-PR-REL-STAT) are added to the analysis of the verb. In the second phase, on the basis of this tag the surface form of the verb is rewritten as the appropriate adjective.

Now when the verb has been tagged as adjective, the word order rules treat it as a proper adjective.

There is still one pitfall to be avoided. The rules should be allowed to apply only to such relative constructions, where the subject prefix and relative prefix refer to the same target, i.e. the subject. Swahili allows also structures, where the relative prefix refers to the object. If this is the case, the rule application should be blocked.

## **7 Conclusion**

In this report I have discussed problems encountered in translating such constructions to English, which in the target language are expressed with adjectives. There is a small number of true adjectives, which can be translated as such. There are also multiword expressions, which must be isolated and controlled so that only the appropriate word clusters are translated as adjectives. A large number of adjectival expressions consist of verbs with relative structure. I have discussed the methods of controlling the correct rule application. When the verbs, originally analysed as finite verb forms, have been rewritten as adjectives, they can be treated as normal adjectives in the further translation process.