Technical Reports in Language Technology
Report No 14, 20014
http://www.njas.helsinki.fi/salama

Representing Swahili adjectives in Machine translation

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

Abstract

In this paper I discuss the problems of translating Swahili adjectives in machine translation from English to Swahili. Various types of adjectives and adjectival expressions are handled. Also methods for handling different types of adjectives are handled.

1. Introduction

Swahili language, and Bantu languages in general, have only a small number of such adjectives that can be classified as true adjectives. In Swahili, true adjectives can be classified in two groups: inflecting and non-inflecting adjectives. In general, old Bantu adjectives inflect and adjectives borrowed from Arabic do not inflect. In translating from world languages into Swahili we encounter a huge number of adjectives, for which there is no established counterpart in Swahili. The most common solution in such cases is to use the corresponding verb with relative structure. For example, the English adjective 'abandoned' can be translated with the structure i-li-yo-achwa or i-li-yo-tupwa. The problem in such structures is that the subject prefix 'i' as well as the relative prefix 'ye' inflects according to the noun class of the word, to which the relative prefix refers. Therefore, we get forms such as mtoto a-li-ye-achwa, watoto wa-li-o-achwa, kiti ki-li-cho-achwa, viti vi-li-vyo-achwa and so on. Also the referent that defines the noun class of the relative prefix can be beyond several words on the left, and not immediately before it.

The problem of representing adjectival expressions is reflected also in dictionaries. For example, English-Swahili Dictionary of TUKI lists abandoned in the following way: abandoned adj 1 (deserted) -lioachwa, -liotupwa, -liotelekezwa, The Swahili glosses do not have the subject prefix, and they have only one form of the relative prefix. The user of the language has to add the subject prefix and change the relative prefix to meet the requirements of the noun class concerned.

When we deal with machine translation, we need a mechanism for controlling these changes. Below I will show how the formation of adjectival relative structures can be implemented in machine translation.

2. Representation in the dictionary

In the rule-based machine translation system, each word is represented in the so-called lexical form in the source language and in the target language. We need a kind of bilingual dictionary, where a lexical word of the source language is represented by its lexical gloss in the target language. In some cases the word has only one gloss in the target language, but in many cases there are more than one gloss, In other words, the

word is polysemous and causes ambiguity in translation. Consider the example in (1), which shows the result of the disambiguated analysis of the string 'abandoned villages'.

(1)
```
"<abandoned>"
      "abandoned" %A> A ABS
"<villages>"
      "village" %<P N NOM PL
```

Using a bilingual electronic dictionary, we add the Swahili glosses to each lexical word. The result is shown in (2).

(2)
```
"<abandoned>"
      "abandoned { -li-achwa , -li-tupwa } A-REL" %A> A ABS
"<villages>"
      "village { 7SG 8PL jiji , 9SG 10PL kaya , 9SG 10PL chengo }"
%<P N NOM PL
```

We see in (2) that the glosses for 'abandoned' are represented with under-defined description -li-achwa and -li-tupwa. The slots for the subject prefix and relative prefix are represented by a dash -, indicating that these slots must be filled later.

The word 'village' has three alternative glosses. These noun glosses are described so that the noun stem is preceded by its noun class codes, one for singular and another for plural.

In order to carry out semantic disambiguation, the result in (2) is transformed into a cascade format, as shown in (3).

(3)
```
"<abandoned>"
      "abandoned" { -li-achwa } A-REL %A> A ABS
      "abandoned" { -li-tupwa } A-REL %A> A ABS
"<villages>"
      "village" { 7SG 8PL jiji } %<P N NOM PL
      "village" { 9SG 10PL kaya } %<P N NOM PL
      "village" { 9SG 10PL chengo } %<P N NOM PL
```

After semantic disambiguation, only one interpretation is left for each token, in this case the first one for both words (4).

(4)
```
"<abandoned>"
      "abandoned" { -li-achwa } A-REL %A> A ABS
"<villages>"
      "village" { 7SG 8PL jiji } %<P N NOM PL
```

On the basis of the sentence structure, we define the noun class, according to which -li-achwa must be inflected. The noun which defines the inflection is 'village + PL', that is,

noun class 8. This is represented by the code A-8 (meaning: adjective of class 8). The result of this phase of processing is in (5).

(5)
```
"<abandoned>"
      "abandoned" { -li-achwa } A-REL %A> A ABS A-8
"<villages>"
      "village" { 8PL jiji } %<P N NOM PL
```

The class code of the adjective (A-8) is split into two codes, one for the subject prefix (SP-8), and another one for the relative prefix (REL-8). These are located to the corresponding places in the adjectival structure -li-achwa, and the dashes are removed. Then the tags are converted into surface form, as shown in (6).

(6)
```
"<abandoned>"
      "abandoned" { vi+li+vyo+achwa } A-REL %A> A ABS
"<villages>"
      "village" { vi+jiji } %<P N NOM PL
```

The morphemes are separated by a '+' sigh to show the morpheme boundary. Also the noun 'villages' has got its full form. What remains to be done, is to reorder the words. This is shown in (7).

(7)
vijiji
vilivyoachwa

3. More examples in context

Below I will demonstrate how the translation system handles in context adjectival expressions as described above, In the example above, the adjectival expression was the type of -li-VStem. There are also other types of adjectival expressions, such as -na-VStem, -si-VStem, -li- na Noun, -na- na Noun, -si- na Noun. In stead of na there can also be the genitive connector a prefixed by the class marker, e.g. -si- -a kiserikali (non-governmental). Consider the example in (8).

(8)
```
"<Non-renewable>"
      "non-renewable" %A> A ABS CAPINIT
"<resources>"
      "resource" %SUBJ N NOM PL
"<are>"
      "be" %+FMAINV V PRES
"<a>"
      "a" %DN> DET SG
"<measurable>"
      "measurable" %A> A ABS
"<target>"
```

```
      "target" %PCOMPL-S N NOM SG
```

There are two adjectives, for which there is no proper adjective gloss in Swahili. When we add the glosses, the result looks as in (9).

(9)
```
"<Non-renewable>"
      "non-renewable { -si-rekebishika } A-REL" %A> A ABS CAPINIT
"<resources>"
      "resource { 7SGki 8PLvi ingizia , 9SG 10PL akili }" %SUBJ N
NOM PL
"<are>"
      "be { wA , ni , si , AUX , LOC } MONOSLB" %+FMAINV V PRES
"<a>"
      "a" %DN> DET SG
"<measurable>"
      "measurable { -na-pimika } A-REL" %A> A ABS
"<target>"
      "target { 9SG 10PL shabaha , 5SG 6PL lengo , 1SG 2PL lengwa
}" %PCOMPL-S N NOM SG
```

Because some lexical words have more than one interpretation, we put the result into the cascade format (10).

(10)
```
"<Non-renewable>"
      "non-renewable" { -si-rekebishika } A-REL %A> A ABS CAPINIT
"<resources>"
      "resource" { 7SGki 8PLvi ingizia } %SUBJ N NOM PL
      "resource" { 9SG 10PL akili } %SUBJ N NOM PL
"<are>"
      "be" { wA } MONOSLB %+FMAINV V PRES
      "be" { ni } MONOSLB %+FMAINV V PRES
      "be" { si } MONOSLB %+FMAINV V PRES
      "be" { AUX } MONOSLB %+FMAINV V PRES
      "be" { LOC } MONOSLB %+FMAINV V PRES
"<a>"
      "a" %DN> DET SG
"<measurable>"
      "measurable" { -na-pimika } A-REL %A> A ABS
"<target>"
      "target" { 9SG 10PL shabaha } %PCOMPL-S N NOM SG
      "target" { 5SG 6PL lengo } %PCOMPL-S N NOM SG
      "target" { 1SG 2PL lengwa } %PCOMPL-S N NOM SG
```

After disambiguation the result is as in (11).

(11)
```
"<Non-renewable>"
      "non-renewable" { -si-rekebishika } A-REL %A> A ABS CAPINIT
```

```
"<resources>"
      "resource" { 7SGki 8PLvi ingizia } %SUBJ N NOM PL
"<are>"
      "be" { ni } MONOSLB %+FMAINV V PRES
"<a>"
      "a" %DN> DET SG
"<measurable>"
      "measurable" { -na-pimika } A-REL %A> A ABS
"<target>"
      "target" { 9SG 10PL shabaha } %PCOMPL-S N NOM SG
```

Next we add noun class tags to words that are interpreted as adjectives (12).

(12)

```
"<Non-renewable>"
      "non-renewable" { -si-rekebishika } A-REL %A> A ABS CAPINIT
A-8
"<resources>"
      "resource" { 8PLvi ingizia } %SUBJ N NOM PL
"<are>"
      "be" { ni } MONOSLB %+FMAINV V PRES
"<measurable>"
      "measurable" { -na-pimika } A-REL %A> A ABS A-9
"<target>"
      "target" { 9SG shabaha } %PCOMPL-S N NOM SG
```

Then the Adjective tags A-8 and A-9 are split and transformed into corresponding subject prefix and relative prefix tags. These tags are moved to the appropriate places in the word structure (13)

(13)

```
"<Non-renewable>"
      "non-renewable" { SP-8+siREL-8+rekebishika } A-REL %A> A ABS
CAPINIT
"<resources>"
      "resource" { 8PLvi ingizia } %SUBJ N NOM PL
"<are>"
      "be" { ni } MONOSLB %+FMAINV V PRES
"<measurable>"
      "measurable" { SP-9+na+REL-9+pimika } A-REL %A> A ABS
"<target>"
      "target" { 9SG shabaha } %PCOMPL-S N NOM SG
```

Then the tags are converted to surface form (14).

(14)

```
"<Non-renewable>"
      "non-renewable" { vi+sivyo+rekebishika } A-REL %A> A ABS
CAPINIT
"<resources>"
```

```
      "resource" { vi+ingizia } %SUBJ N NOM PL
"<are>"
      "be" { ni } MONOSLB %+FMAINV V PRES
"<measurable>"
      "measurable" { i+na+yo+pimika } A-REL %A> A ABS
"<target>"
      "target" {  shabaha } %PCOMPL-S N NOM SG
```

When the words are reordered, we het the result as in (15).

(15)
Viingizia
visivyorekebishika
ni
shabaha
inayopimika

Below are more example sentences, for which I show only some phases of processing.

(16)

```
"<Non-toxic>"
      "non-toxic { -si- na sumu } A-REL" %A> A ABS CAPINIT
"<plants>"
      "plant { 3SG 4PL mea }" %SUBJ N NOM PL
"<do>"
      "do { fanyA , AUX , fanzA , tendA } SVO" %+FAUXV V PRES
"<not>"
      "not { NOGLOSS , si }" %ADVL NEG-PART
"<grow>"
      "grow { kuzA , meA , otA , kuA , zidi } SVO " %-FMAINV V INF
"<in>"
      "in { katika , mwaka , kwa , NOGLOSS }" %ADVL PREP
"<waterless>"
      "waterless { -si- na maji } A-REL" %A> A ABS
"<places>"
      "place { 16SG mahali , 5SG 6PL ganjo }" %<P N NOM PL
"<<s>>"
      "<s>"
"<Nongovernmental>"
      "nongovernmental { -si- a kiserikali } A-REL" %A> A ABS
CAPINIT
"<organizations>"
      "organization { 9SG 10PL oganaizesheni , 9SG 10PL
oganaizeisheni , 9SG 10PL oganizeisheni }" %SUBJ N NOM PL
"<are>"
      "be { wA , ni , si , AUX , LOC } MONOSLB" %+FMAINV V PRES
"<nonprofit>"
      "nonprofit { -si- na faida } A-REL" %A> A ABS
"<enterprises>"
      "enterprise { 9SG 10PL shughuli }" %PCOMPL-S N NOM PL
```

After semantic disambiguation and isolation of multiword expressions the result is as in (17)

(17)

```
"<Non-toxic>"
      "non-toxic" { -si- na sumu } A-REL %A> A ABS CAPINIT
"<plants>"
      "plant" { 4PL mea } %SUBJ N NOM PL
"<do>"
      "do" { AUX } SVO %+FAUXV V PRES
"<not>"
      "not" { si } %ADVL NEG-PART
"<grow>"
      "grow" { otA } SVO %-FMAINV V INF
"<in>"
      "in" { katika } %ADVL PREP
"<waterless>"
      "waterless" { -si- na maji } A-REL %A> A ABS
"<places>"
      "place" { 16SG mahali } %<P N NOM PL
"<<s>>"
      "<s>"
"<Nongovernmental>"
      "nongovernmental" MW>
"<organizations>"
      "organization" { 10PL asasi -si- a kiserikali } %SUBJ N NOM
PL <MW REL
"<are>"
      "be" { ni } MONOSLB %+FMAINV V PRES
"<nonprofit>"
      "nonprofit" { -si- na faida } A-REL %A> A ABS
"<enterprises>"
      "enterprise" { 10PL shughuli } %PCOMPL-S N NOM PL
```

Note that 'nongovernmental organization' in (17) is handled as a multiword expression'. The original glosses were removed and a new interpretation was given to this word cluster. Then we add tags for facilitating inflection (18).

(18)

```
"<Non-toxic>"
      "non-toxic" { -si- na sumu } A-REL %A> A ABS CAPINIT A-4
"<plants>"
      "plant" { 4PL mea } %SUBJ N NOM PL
"<do>"
      "do" { AUX } SVO %+FAUXV V PRES SP-NEG-4
"<not>"
      "not" { si } %ADVL NEG-PART
"<grow>"
```

```
      "grow" { otA } SVO %-FMAINV V INF SP-4 SP-NEG-2 SP-NEG-4
TAM-0
"<in>"
      "in" { katika } %ADVL PREP
"<waterless>"
      "waterless" { -si- na maji } A-REL %A> A ABS A-16
"<places>"
      "place" { 16SG mahali } %<P N NOM PL
"<<s>>"
      "<s>"
"<Nongovernmental>"
      "nongovernmental" MW>
"<organizations>"
      "organization" { 10PL asasi -si- a kiserikali } %SUBJ N NOM
PL <MW REL G-10 SP-10 REL-10
"<are>"
      "be" { ni } MONOSLB %+FMAINV V PRES
"<nonprofit>"
      "nonprofit" { -si- na faida } A-REL %A> A ABS A-10
"<enterprises>"
      "enterprise" { 10PL shughuli } %PCOMPL-S N NOM PL
```

When morpheme tags are moved to appropriate places, the result is as in (19).

(19)

```
"<Non-toxic>"
      "non-toxic" { SP-4+siREL-4+ na sumu } A-REL %A> A ABS
CAPINIT
"<plants>"
      "plant" { 4PL mea } %SUBJ N NOM PL
"<do>"
      "do" { AUX } SVO %+FAUXV V PRES
"<not>"
      "not" { si } %ADVL NEG-PART
"<grow>"
      "grow" { SP-NEG-4+TAM-0+ot+i } SVO %-FMAINV V INF
"<in>"
      "in" { katika } %ADVL PREP
"<waterless>"
      "waterless" { SP-16siREL-16 na maji } A-REL %A> A ABS
"<places>"
      "place" { 16SG mahali } %<P N NOM PL
"<<s>>"
      "<s>"
"<Nongovernmental>"
      "nongovernmental" MW>
"<organizations>"
      "organization" { 10PL asasi SP-10+-si-REL-10+ G-10+a
kiserikali } %SUBJ N NOM PL <MW REL
"<are>"
      "be" { ni } MONOSLB %+FMAINV V PRES
```

```
"<nonprofit>"
      "nonprofit" { SP-10+siREL-10 na faida } A-REL %A> A ABS
"<enterprises>"
      "enterprise" { 10PL shughuli } %PCOMPL-S N NOM PL
```

These noun class tags are converted to surface form (20).

(20)

```
"<Non-toxic>"
      "non-toxic" { i+siyo+ na sumu } A-REL %A> A ABS CAPINIT
"<plants>"
      "plant" { mi+mea } %SUBJ N NOM PL
"<do>"
      "do" { AUX } SVO %+FAUXV V PRES SP
"<not>"
      "not" { si } %ADVL NEG-PART
"<grow>"
      "grow" { ha+i+ot+i } SVO %-FMAINV V INF
"<in>"
      "in" { katika } %ADVL PREP
"<waterless>"
      "waterless" { pa+si+po na maji } A-REL %A> A ABS
"<places>"
      "place" {  mahali } %<P N NOM PL
"<<s>>"
      "<s>"
"<Nongovernmental>"
      "nongovernmental" MW>
"<organizations>"
      "organization" { asasi zi+si+zo za kiserikali } %SUBJ N NOM
PL <MW REL
"<are>"
      "be" { ni } MONOSLB %+FMAINV V PRES
"<nonprofit>"
      "nonprofit" { zi+sizo na faida } A-REL %A> A ABS
"<enterprises>"
      "enterprise" {  shughuli } %PCOMPL-S N NOM PL
```

When the reordering of words is performed, the result is as in (21).

(21)
Mimea
isiyo na sumu
haioti
katika
mahali
pasipo na maji
Asasi zisizo za kiserikali
ni

shughuli
zisizo na faida

4. Conclusion

In this paper I have shown that although it is very difficult to describe correctly into a dictionary such adjectival expressions, which are formed using relative verb structures, it is possible to handle them precisely in machine translation. Therefore, the scarcity of true adjectives in languages such as Swahili is not an obstacle for machine translation. As long as true adjectives are not coined for all commonly occurring adjectives, structures described above can be used. And in fact it is obligatory, because it must be possible to translate any normal text into Swahili.