*Technical Reports on Language Technology*
*Report No 46, 20019*
*http://www.njas.helsinki.fi/salama*

# Multiword expressions in English to Swahili machine translation: Nouns

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

**Abstract**

In machine translation, two types of multiword expressions (MWE) must be considered. Such expressions may occur in source text and target text, or in both. A MWE in source text may be a MWE also in target text. There are also cases, when a MWE in source text corresponds to a single word in target text. However, the majority of cases are such, where a single word in source text corresponds to a MWE in target text. In this report we discuss the last type of cases. Th emphasis is on nouns.

**Key Words:** *multiword expression, multiword nouns, machine translation.*

## 1 Introduction

When we consider translation from one language to another, our assumption is that a word in source language corresponds to a word in target language. This assumption is correct in most cases. Especially in languages, which have been exposed to technological development for a long time, most of the important concepts have been glossed with suitable terms. Therefore, the need to use descriptive translation is rather limited.

The situation is very different with languages such as Swahili, a Bantu language, which, as is typical to all Bantu languages, has a very limited number of lexical adjectives, for example. In addition, many English nouns appearing in text do not have a corresponding gloss in Swahili. This has led to extensive use of descriptions in lexicography. The process of inventing proper lexical glosses has been very poor, or non-existent. The English-Swahili dictionary of TUKI (Taasisi ya Utafiti wa Kiswahili) has thousands of such descriptions instead of proper glosses. But what is even worse, the TUKI dictionary has thousands of English adjectives and adverbs, which have no gloss at all, not even a descriptive one. Therefore, in developing the translation system from English to Swahili, I have been compelled to express the adjectives in some way, most often using the conventional methods of description.

Out of the 8,019 adjectives in the translation dictionary, a total of 3037 use the genitive structure with *-a* plus noun, 1035 use the relative structure *-enye* plus noun, and 2144 use the relative structure of verb. Only some 1800 adjectives have a gloss, which corresponds to a real adjective, inflecting or non-inflecting.

With nouns the situation is less disastrous. Out of the 26532 nouns, which include also many proper names, a total of 1659 uses a description, which includes the particle *-a*, 163 such which use the construction *-enye*, and 137 use the relative construction of the verb.

The methods of handling the correct surface forms of nouns and adjectives are different. Nouns have only singular and plural forms of the noun class pair, to which the noun belongs. We must also take care of the correct forms of the inflecting parts in the description. Adjectives are different in that they do not have any inherent noun class affiliation. They inflect according to their main word, which is a noun.

The correct form of the descriptive nouns can be solved without reference to other parts of the sentence. The English noun form indicates whether the word is in singular or plural, and the inflecting parts can be handled on that basis. In the translation process, the correct noun forms can be handled immediately after the Swahili glosses have been added.

With adjectives we must wait until the inflection tags have been added, based on the environment in each case. Therefore, the adjective forms must be handled in a later phase.

## 2 Multiword nouns

When an English noun does not have a direct gloss is Swahili, the meaning is usually expressed using description. Such descriptions have usually inflecting parts, whereby direct description is not possible. The description must have an abstract form, so that all possible surface forms can be constructed on the basis of the abstract representation.

### 2.1 Constructions with -a or -enye

A very common form of constructing descriptions is the use of *-a* and *-enye* for describing the meaning of the concept. The dash before the word means that a noun class prefix should be attached to the word.

Consider the noun description, derived from the noun conversion dictionary (1).

(1)
```
"<homeland>"
      "homeland" { 9SG 10PL nchi 9SG 10PL -a asili } %NH N SG NOM
```

The noun homeland is glossed as composed of three elements, *nchi*, *-a*, and *asili*. This is a genitive construction meaning *the land of origin* (not a very good description). The noun *nchi* and the genitive connector *-a* inflect according to whether the expression is in singular or plural. And the inflection takes place according to classes 9 and 10.

We see that the analysed English word *homeland* has the tag SG, inherited from the analysis of English. This information is used in selecting the appropriate inflection tags. The result is in (2).

(2)
```
"<homeland>"
      "homeland" { 9SG nchi 9SG -a asili } %NH N SG NOM
```

Now only the singular tags are left, and the lexical words can be converted into surface form (3).

(3)
```
"<homeland>"
      "homeland" { nchi ya asili } %NH N SG NOM
```

We see that the word *nchi* did not get a prefix at all. This is often, but not always, the case with class 9/10 nouns. The genitive connector was given the prefix *y*.

If we put the English noun *homeland* into plural, we get the abstract form as in (4).

(4)
```
"<homelands>"
      "homeland" { 10PL nchi 10PL -a asili } %NH N PL NOM
```

The reading has the tag PL, and plural tags are selected for inflecting words. The surface form is in (5).

(5)
```
"<homelands>"
      "homeland" { nchi za asili } %NH N PL NOM
```

We see that *nchi* is the same as in singular, but the genitive connector is different, *za*.
We take some additional cases of other noun classes (6).

(6)
```
"<mouse-trap>"
      "mouse-trap" { 3SG 4PL tego 3SG 4PL -a panya } %NH N SG NOM
"<opera-house>"
      "opera-house" { 5SG 6PL jumba 5SG 6PL -a opera } %NH N SG
NOM
"<order-book>"
      "order-book" { 7SG 8PL tabu 7SG 8PL -a maagizo } %NH N SG
NOM
```

All three English words are in singular, and the system selects the singular tags (7).

(7)
```
"<mouse-trap>"
      "mouse-trap" { 3SG tego 3SG -a panya } %NH N SG NOM
"<opera-house>"
      "opera-house" { 5SG jumba 5SG -a opera } %NH N SG NOM
"<order-book>"
      "order-book" { 7SG tabu 7SG -a maagizo } %NH N SG NOM
```

The surface forms are in (8).

(8)
```
"<mouse-trap>"
      "mouse-trap" { mtego wa panya } %NH N SG NOM
"<opera-house>"
```

```
      "opera-house" { jumba la opera } %NH N SG NOM
"<order-book>"
      "order-book" { kitabu cha maagizo } %NH N SG NOM
```

When the English words are in plural, the inflection tags are selected as in (9).

(9)
```
"<mouse-trap>"
      "mouse-trap" { 4PL tego 4PL -a panya } %NH N PL NOM
"<opera-house>"
      "opera-house" { 6PL jumba 6PL -a opera } %NH N PL NOM
"<order-book>"
      "order-book" { 8PL tabu 8PL -a maagizo } %NH N PL NOM
```

The surface forms in plural come as in (10).

(10)
```
"<mouse-trap>"
      "mouse-trap" { mitego ya panya } %NH N PL NOM
"<opera-house>"
      "opera-house" { majumba ya opera } %NH N PL NOM
"<order-book>"
      "order-book" { vitabu vya maagizo } %NH N PL NOM
```

The genitive connector *-a* is sometimes replaced with the relative pronoun *-enye*. The behaviour of this structure is much the same as of *-a*, but *-enye* has more a possessive connotation, while *-a* connects nouns hierarchically. Examples are in (11).

(11)
```
"<optimist>"
      "optimist" { 1SG 2PL -enye msimamo wa kutegemea mazuri } %NH
N SG NOM
"<hotbed>"
      "hotbed" { 5SG 6PL tuta 5SG 6PL -enye mbolea } %NH N SG NOM
"<lugger>"
      "lugger" { 7SG 8PL ombo 7SG 8PL -enye tanga la pembe nne }
%NH N SG NOM
"<lightship>"
      "lightship" { 9SG 10PL meli 9SG 10PL -enye taa za kuongozea
meli nyingine } %NH N SG NOM
```

These examples are converted into singular surface forms as in (12).

(12)
```
"<optimist>"
      "optimist" { mwenye msimamo wa kutegemea mazuri } %NH N SG
NOM
"<hotbed>"
      "hotbed" { tuta lenye mbolea } %NH N SG NOM
```

```
"<lugger>"
      "lugger" { chombo chenye tanga la pembe nne } %NH N SG NOM
"<lightship>"
      "lightship" { meli yenye taa za kuongozea meli nyingine }
%NH N SG NOM
```

Correspondingly, the plural forms are as in (13).

(13)
```
"<optimists>"
      "optimist" { 2PL -enye msimamo wa kutegemea mazuri } %NH N
PL NOM
"<hotbeds>"
      "hotbed" { 6PL tuta 6PL -enye mbolea } %NH N PL NOM
"<luggers>"
      "lugger" { 8PL ombo 8PL -enye tanga la pembe nne } %NH N PL
NOM
"<lightships>"
      "lightship" { 10PL meli 10PL -enye taa za kuongozea meli
nyingine } %NH N PL NOM
```

And the surface forms in plural are as in (14).

(14)
```
"<optimists>"
      "optimist" { wenye msimamo wa kutegemea mazuri } %NH N PL
NOM
"<hotbeds>"
      "hotbed" { matuta yenye mbolea } %NH N PL NOM
"<luggers>"
      "lugger" { vyombo vyenye tanga la pembe nne } %NH N PL NOM
"<lightships>"
      "lightship" { meli zenye taa za kuongozea meli nyingine }
%NH N PL NOM
```

### 2.2 Method using the relative verb construction

In the formulation of noun descriptions, Swahili makes use of the possibility to use the relative marking in verb. The relative marker comes after the TAM marker, or to the end of the verb in the cases, whereby time marking is not needed. Each noun class has its own relative marker, although two or three noun classes may have the same surface form.

In all, there are three methods of using the relative marker. We will discuss each of them.

#### 2.2.1 Relative marker as a prefix

When the relative marker is before the verb stem, we call it a prefix. Examples of this structure are in (15).

(15)

```
"<headhunter>"
      "headhunter" { 1SG 2PL tu 1SG 2PL -na-tafuta mabingwa } %NH
N SG NOM
"<halberdier>"
      "halberdier" { 9SG 10PL askari 1SG 2PL -na-tumia mkukishoka
} HUM %NH N SG NOM
"<gigolo>"
      "gigolo" { 1SG 2PL anamume 1SG 2PL -na-kodishwa na 1SG 2PL
anamke } %NH N SG NOM
"<reticulation>"
      "reticulation" { 3SG 4PL fumo 3SG 4PL -na-fanana na wavu }
%NH N SG NOM
"<epicycle>"
      "epicycle" { 5SG 6PL duara 5SG 6PL -na-zunguka nje ya duara
} %NH N SG NOM
"<pocketful>"
      "pocketful" { 7SGki 8PLvi asi 7SG 8PL -na-jaa mfuko } %NH N
SG NOM
"<convenience>"
      "convenience" { 9SG 10PL hali 9SG 10PL -na-faa } %NH N SG
NOM
```

When we remove the unsuitable class tags, we get the result with singular tags as in (16).

(16)

```
"<headhunter>"
      "headhunter" { 1SG tu 1SG -na-tafuta mabingwa } %NH N SG NOM
"<halberdier>"
      "halberdier" { 9SG askari 1SG -na-tumia mkukishoka } HUM %NH
N SG NOM
"<gigolo>"
      "gigolo" { 1SG anamume 1SG -na-kodishwa na 1SG anamke } %NH
N SG NOM
"<reticulation>"
      "reticulation" { 3SG fumo 3SG -na-fanana na wavu } %NH N SG
NOM
"<epicycle>"
      "epicycle" { 5SG duara 5SG -na-zunguka nje ya duara } %NH N
SG NOM
"<pocketful>"
      "pocketful" { 7SGki asi 7SG -na-jaa mfuko } %NH N SG NOM
"<convenience>"
      "convenience" { 9SG hali 9SG -na-faa } %NH N SG NOM
```

We see in (16) above that in most cases the noun class of the relative construction is the same as of its head. An exception is *halberdier*, which has two noun classes. The noun *askari* is in class 9 but the verb inflects according to class 1. This is due to the fact that many borrowed nouns meaning a human being inflect according to the class pair 9/10,

but because they are humans, verbs inflect according to the class of humans, that is classes 1 and 2.

The surface forms of the subject prefix and relative prefix are in (17).

(17)
```
"<headhunter>"
      "headhunter" { 1SG tu anayetafuta mabingwa } %NH N SG NOM
"<halberdier>"
      "halberdier" { 9SG askari anayetumia mkukishoka } HUM %NH N
SG NOM
"<reticulation>"
      "reticulation" { 3SG fumo unaofanana na wavu } %NH N SG NOM
"<epicycle>"
      "epicycle" { 5SG duara linalozunguka nje ya duara } %NH N SG
NOM
"<pocketful>"
      "pocketful" { 7SGki asi kinachojaa mfuko } %NH N SG NOM
"<convenience>"
      "convenience" { 9SG hali inayofaa } %NH N SG NOM
```

When we finalise also the form of the head nouns, we get the result as in (18).

(18)
```
"<headhunter>"
      "headhunter" { mtu anayetafuta mabingwa } %NH N SG NOM
"<halberdier>"
      "halberdier" { askari anayetumia mkukishoka } HUM %NH N SG
NOM
"<reticulation>"
      "reticulation" { mfumo unaofanana na wavu } %NH N SG NOM
"<epicycle>"
      "epicycle" { duara linalozunguka nje ya duara } %NH N SG NOM
"<pocketful>"
      "pocketful" { kiasi kinachojaa mfuko } %NH N SG NOM
"<convenience>"
      "convenience" { hali inayofaa } %NH N SG NOM
```

The surface forms of the plural forms of the English words are in (19).

(19)
```
"<headhunters>"
      "headhunter" { watu wanaotafuta mabingwa } %NH N PL NOM
"<halberdiers>"
      "halberdier" { askari wanaotumia mkukishoka } HUM %NH N PL
NOM
"<gigolos>"
      "gigolo" { wanaume wanaokodishwa na wanawake } %NH N PL NOM
"<reticulations>"
      "reticulation" { mifumo inayofanana na wavu } %NH N PL NOM
"<epicycles>"
```

```
      "epicycle" { maduara yanayozunguka nje ya duara } %NH N PL
NOM
"<pocketfuls>"
      "pocketful" { viasi vinavyojaa mfuko } %NH N PL NOM
"<conveniences>"
      "convenience" { hali zinazofaa } %NH N PL NOM
```

In all examples above, the verb has the TAM marker *-na-*, meaning present tense. A similar structure would be possible also with the TAM markers *-li-* (past tense) and *-si-* (negative), and *-taka-* (future tense). However, they are not feasible, except for *-si-*, in describing nouns on the lexical level, and they do not appear in noun descriptions.

Below are examples of the use of the negative marker *-li-* in constructing relative structures of the verb (20).

(20)

```
"<non-smoker>"
      "non-smoker" { 1SG 2PL tu 1SG 2PL -si-vuta sigara } HUM %NH
N SG NOM "<firebrick>"
      "firebrick" { 5SG 6PL tofali 5SG 6PL -si-athiriwa na moto }
%NH N SG NOM
"<banality>"
      "banality" { 9SG 10PL hali 9SG 10PL -si-vutia } %NH N SG NOM
"<ex-officer>"
      "ex-officer" { 9SG 10PL afisa 1SG 2PL -si-ajiriwa } HUM %<P
N SG NOM
```

When we process these descriptions further, we get the surface singular forms as in (21).

(21)

```
"<non-smoker>"
      "non-smoker" { mtu asiyevuta sigara } HUM %A> N SG NOM
"<firebrick>"
      "firebrick" { tofali lisiloathiriwa na moto } %A> N SG NOM
"<banality>"
      "banality" { hali isiyovutia } %NH N SG NOM
"<ex-officer>"
      "ex-officer" { afisa asiyeajiriwa } HUM %<P N SG NOM
```

The surface forms in plural are in (22)

(22)

```
"<non-smokers>"
      "non-smoker" { 2PL tu wasiovuta sigara } %NH N PL NOM
"<firebricks>"
      "firebrick" { 6PL tofali yasiyoathiriwa na moto } %NH N PL
NOM
"<banalities>"
      "banality" { 10PL hali zisizovutia } %NH N PL NOM
"<ex-officers>"
```

```
      "ex-officer" { 10PL afisa wasioajiriwa } %<NOM Heur N PL NOM
```

### 2.2.2 Relative marker as a suffix

Relative structures in describing nouns can also be used so that, instead of being a verb prefix, the marker is a suffix. This takes place in two ways. One way is to put the relative marker after the verb substitute *li* (affirmative) or *si* (negative). Examples are in (23).

(23)

```
"<holiday-maker>"
      "holiday-maker" { 1SG 2PL tu 1SG 2PL -li- likizoni } %A> N
SG NOM
"<catchment>"
      "catchment" { 5SG 6PL eneo 5SG 6PL -li- chanzo cha maji }
%A> N SG NOM
"<minimum>"
      "minimum" { 9SG 10PL kadiri 9SG 10PL -li- dogo } %A> N SG
NOM
"<roadster>"
      "roadster" { 9SG 10PL gari ndogo 9SG 10PL -li- wazi } %NH N
SG NOM
"<extremist>"
      "extremist" { 1SG 2PL tu 1SG 2PL -si- na kadiri } %NH N SG
NOM
"<misfit>"
      "misfit" { 5SG 6PL vazi 5SG 6PL -si- sawa na kimo } %NH N SG
NOM
"<vagary>"
      "vagary" { 5SG 6PL tukio 5SG 6PL -si- 5SG 6PL -a kawaida }
%NH N SG NOM
"<divan>"
      "divan" { 7SG 8PL ti 7SG 8PL -si- na egemeo } %NH N SG NOM
"<bungalow>"
      "bungalow" { 9SG 10PL nyumba 9SG 10PL -si- na ghorofa } %NH
N SG NOM
```

The surface form of these nouns in singular are in (24).

(24)

```
"<holiday-maker>"
      "holiday-maker" { mtu aliye likizoni } %A> N SG NOM
"<catchment>"
      "catchment" { eneo lililo chanzo cha maji } %A> N SG NOM
"<minimum>"
      "minimum" { kadiri iliyo dogo } %A> N SG NOM
"<roadster>"
      "roadster" { gari ndogo iliyo wazi } %A> N SG NOM
"<extremist>"
      "extremist" { mtu asiye na kadiri } %A> N SG NOM
"<misfit>"
```

```
      "misfit" { vazi lisilo sawa na kimo } %A> N SG NOM
"<vagary>"
      "vagary" { tukio lisilo la kawaida } %A> N SG NOM
"<divan>"
      "divan" { kiti kisicho na egemeo } %A> N SG NOM
"<bungalow>"
      "bungalow" { nyumba isiyo na ghorofa } %NH N SG NOM
```

The plural forms of the nouns are in (25).

(25)

```
"<holiday-makers>"
      "holiday-maker" { watu walio likizoni } %NH N PL NOM
"<catchments>"
      "catchment" { maeneo yaliyo chanzo cha maji } %NH N PL NOM
"<minimums>"
      "minimum" { kadiri zilizo dogo } %NH N PL NOM
"<roadsters>"
      "roadster" { gari ndogo zilizo wazi } %NH N PL NOM
"<extremists>"
      "extremist" { watu wasio na kadiri } %NH N PL NOM
"<misfits>"
      "misfit" { mavazi yasiyo sawa na kimo } %NH N PL NOM
"<vagaries>"
      "vagary" { matukio yasiyo ya kawaida } %NH N PL NOM
"<divans>"
      "divan" { viti visivyo na egemeo } %A> N PL NOM
"<bungalows>"
      "bungalow" { nyumba zisizo na ghorofa } %NH N PL NOM
```

Another method of using the relative marker as a suffix is to put the relative marker at the end of the verb. In most cases this is an alternative to putting the relative marker as a prefix after the TAM marker *-na-*. When the suffix alternative is used, the verb has no reference to time, and the verb does not have a TAM marker. Examples of this usage are in (26).

(26)

```
"<clapper>"
      "clapper" { 1SG 2PL tu 1SG 2PL -fanya- sauti ya mpasuko }
%A> N SG NOM
"<cottager>"
      "cottager" { 1SG 2PL tu 1SG 2PL -ishi- kwenye nyumba ndogo }
%A> N SG NOM
"<dilettant>"
      "dilettant" { 1SG 2PL -jifunza- kitu kijuujuu } HUM %A> Heur
N SG NOM
"<cropper>"
      "cropper" { 3SG 4PL mea 3SG 4PL -toa- mazao mazuri } %A> N
SG NOM
"<dressing-gown>"
```

```
      "dressing-gown" { 5SG 6PL vazi 5SG 6PL -valiwa- wakati 1SG
tu anapopumzika } %A> N SG NOM
"<love-match>"
      "love-match" { 9SG 10PL ndoa 9SG 10PL -tokana- na mapenzi tu
} %A> Heur N SG NOM
"<zoo>"
      "zoo" { 16SG mahali 16SG -fugwa- wanyama pori } %NH N SG NOM
```

An interesting example is the last one *zoo*, which has the locative class 16. The description means literally *the place where wild animals are taken care of.*

The singular surface forms are in (27).

(27)

```
"<clapper>"
      "clapper" { mtu afanyaye sauti ya mpasuko } %A> N SG NOM
"<cottager>"
      "cottager" { mtu aishiye kwenye nyumba ndogo } %A> N SG NOM
"<dilettant>"
      "dilettant" { ajifunzaye kitu kijuujuu } HUM %A> Heur N SG
NOM
"<cropper>"
      "cropper" { mmea utoao mazao mazuri } %A> N SG NOM
"<dressing-gown>"
      "dressing-gown" { vazi livaliwalo wakati mtu anapopumzika }
%A> N SG NOM
"<love-match>"
      "love-match" { ndoa itokanayo na mapenzi tu } %A> Heur N SG
NOM
"<zoo>"
      "zoo" { mahali pafugwapo wanyama pori } %NH N SG NOM
```

The plural forms are in (28).

(28)

```
"<clappers>"
      "clapper" { watu wafanyao sauti ya mpasuko } %NH N PL NOM
"<cottagers>"
      "cottager" { watu waishio kwenye nyumba ndogo } %NH N PL NOM
"<dilettants>"
      "dilettant" { wajifunzao kitu kijuujuu } HUM %A> Heur N PL
NOM
"<croppers>"
      "cropper" { mimea itoayo mazao mazuri } %NH N PL NOM
"<dressing-gowns>"
      "dressing-gown" { mavazi yavaliwayo wakati mtu anapopumzika
} %NH N PL NOM
"<love-matches>"
      "love-match" { ndoa zitokanazo na mapenzi tu } %A> Heur N SG
NOM
"<zoos>"
      "zoo" { mahali pafugwapo wanyama pori } %NH N PL NOM
```

Not that the last example *zoos* has only a singular description, and plural form does not exist for the noun *mahali*.

## 3 Conclusion

We see in the above examples that all surface forms of the complex noun descriptions can be produced on the basis of the abstract form of these descriptions in conversion dictionary. This translation method is far from ideal, but as long as the proper lexical glosses are missing from the language, there are hardly other alternatives. With adjectives the situation is somewhat easier, because there are established methods of constructing adjectival expressions, when proper adjectives are missing.