Technical Reports in Language Technology
Report No 10, 20012
http://www.njas.helsinki.fi/salama

Multiword expressions in English to Swahili machine translation

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

Abstract

There have been several workshops exploring the nature of multiword expressions. This paper discusses the issue from the viewpoint of machine translation and argues that the solutions to be selected depend on the language pair. Examples concern translation from English to Swahili.

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

Abbreviations:
MWE = multiword expressions
MT = machine translation
SL = source language
TL = target language

1 Introduction

There have been a number of workshops on multiword expressions (MWE) in conjunction with world conferences on language technology. At least ACL, COLING and LREC have hosted such workshops. Although the subject has been highlighted from a number of viewpoints, astonishingly little has been written on how MWEs should be treated in machine translation, and especially, how they can be implemented in real life. Below is an analysis of various types of MWEs as they appear in machine translation from Swahili to English, and from English to Swahili. I also show how they have been implemented in Swahili Language Manager (SALAMA).

In machine translation, we encounter three basic types of MWEs, one-to-many, many-to-one, and many-to-many. Any correspondence that fulfils one of these three criteria is a MWE and needs a separate treatment. Therefore, the number of MWEs in a language is not fixed. It varies according to language pair. For example, adjectives often do not pose major problems in MT, because both languages have corresponding adjectives. In Bantu languages, on the other hand, adjectives are a very small category. When translating from English to Swahili, most English adjectives have to be represented, not with corresponding Swahili adjectives, but with various structures that constitute MWEs. What is even worse is that these structures must meet the noun class concordance of the noun referent, wherever in the sentence the noun is.

Another example of the flexibility of what is a MWE are compound nouns and named entities. Many of these are semi-frozen, because they may have a singular and plural form. Therefore, in implementation they should be allowed to have both forms, although in practice some of them occur only in one form.

Consider the following examples (1):

(1)
health service *huduma ya afya* (service of health)
Security Council *Baraza la Usalama* (Council of Security)
East African Community *Jumuiya ya Afrika Mashariki* (Community of Africa Eastern)
Chief Academic Officer *Ofisa Mkuu wa Taaluma* (Officer Chief of Learning)

On the other hand, there are also frozen MWEs that have a fixed form in both languages.

(2)
by all means *kwa udi na uvumba* (by aloe wood and incense)
in all aspects *kwa mapana na marefu* (in breath and length)
income per capita *pato kwa kila mtu* (income for every person)

2 In which processing phase should MWEs be handled?

Rule-based machine translation follows a certain sequence of processing. Every type of MWEs must be handled at some point of processing, but it is not self-evident that all of them should be handled in the same slot of processing sequence. Intuitively one would like to follow the following principle: handle them as early as possible so as to avoid unnecessary rule writing later.

Following this principle, frozen MWEs, at least the frequent ones, could be treated already as part of the basic morphological description. Frozen MWEs could also be handled as part of describing other types of MWEs. I have implemented both solutions in Swahili-to-English MT and compared the work needed in each case. The result is, surprisingly, that it is easier and requires less work to handle also frozen MWEs in conjunction with other types of MWEs. What is even more important is that it is possible to formulate the needed rules through programming. When each type of MWEs in source language is identified with corresponding gloss in target language (two, three, four etc. members; whether the head is a verb, noun, or no head), rules can be written for each type automatically. This method drastically reduces time needed for rule writing. The biggest work load is in compiling the MWE pairs, especially because dictionaries tend to be unaware of the need to list MWEs.

3 Ambiguity in MWEs

Although some MWEs are fully frozen, so that they do not inflect and have only one meaning in both languages, many do inflect and have more than one meaning. The MWEs should be described so that disambiguation is possible in conjunction with normal disambiguation routines. Also this fact supports the solution that MWEs are described

before disambiguation proper. The disambiguation rules for MWEs can be written as part of the normal rules.

4 MWE as a grammatical unit

Because MWEs are a diverse category including anything between fully frozen units and fully inflecting sequences that allow intervening words, it is important that the environment for writing rules allows all needed kinds of constraints. Therefore, the rules can be written with various degrees of frigidity.

Some MWEs, such as *face to face* (uso kwa uso), are adverbs and can be translated as such without any grammatical modifications. Even long proverbs, such as *Afadhali kenda shika nenda kuliko kumi nenda rudi* (Better to have nine and keep going than to have ten and go and come back), can be treated in the same way. However, many MWEs are idiomatic verbs and need to be treated as verbs. A large number of MWEs in English are compound nouns and they need to preserve the grammatical information of the head noun, so as to provide information for constructing the correct form in TL. Adjectives are a diverse category in Swahili, and many English adjectives are represented with MWEs in Swahili and require special treatment.

5 Adjective constructions

Below are examples of translating various adjective constructions from English to Swahili. The first group contains normal cases, where the adjective of the SL is represented with an adjective in TL. In this group, inflecting adjectives must be separated from non-inflecting ones.

The following examples start from the phase where the English text has been analyzed and disambiguated, and the Swahili glosses with noun class information has been inserted.

5.1 Inflecting adjectives

Inflecting adjectives are processed as follows:

(3)

```
"<The>"
      "the" %DN> DET
"<aged>"
      "aged" { zee } A-INFL %A> A ABS
"<people>"
      "people" { 2PL tu , 11SG mma } %SUBJ N NOM
"<have>"
      "have" { -wa na , AUX } %+FAUXV V PRES
"<retired>"
      "retire" { staafu , ng'atuA , jitoA , uzuA , uzulu } %-
FMAINV EN
"<.>"
      "."
```

The analysis result must be modified for making it suitable for disambiguation (4).

(4)

```
"<The>"
      "the" %DN> DET
"<aged>"
      "aged" { zee } A-INFL %A> A ABS
"<people>"
      "people" { 2PL tu } %SUBJ N NOM
      "people" { 11SG mma } %SUBJ N NOM
"<have>"
      "have" { -wa na } %+FAUXV V PRES
      "have" { AUX } %+FAUXV V PRES
"<retired>"
      "retire" { staafu } %-FMAINV EN
      "retire" { ng'atuA } %-FMAINV EN
      "retire" { jitoA } %-FMAINV EN
"<.>"
      "."
```

The disambiguated text is in (5).

(5)

```
"<The>"
      "the" %DN> DET
"<aged>"
      "aged" { zee } A-INFL %A> A ABS
"<people>"
      "people" { 2PL tu } %SUBJ N NOM
"<have>"
      "have" { AUX } %+FAUXV V PRES
"<retired>"
      "retire" { staafu } %-FMAINV EN
"<.>"
      "."
```

Tags are added for constructing surface forms in Swahili (6).

(6)

```
"<The>"
      "the" %DN> DET
"<aged>"
      "aged" { zee } A-INFL %A> A ABS A-2
"<people>"
      "people" { 2PL tu } %SUBJ N NOM
"<have>"
      "have" { AUX } %+FAUXV V PRES
"<retired>"
      "retire" { staafu } %-FMAINV EN SP-2 SP-2 TAM-me
"<.>"
      "."
```

Tags in the end of appropriate readings are moved to appropriate places (7).

(7)
```
"<The>"
      "the" %DN> DET
"<aged>"
      "aged" { A-2+zee } A-INFL %A> A ABS
"<people>"
      "people" { 2PL tu } %SUBJ N NOM
"<have>"
      "have" { AUX } %+FAUXV V PRES
"<retired>"
      "retire" { SP-2+TAM-me+staafu } %-FMAINV EN
"<.>"
      "."
```

Then the morpheme tags are converted into surface forms (8).

(8)
```
"<The>"
      "the" %DN> DET
"<aged>"
      "aged" { wa+zee } A-INFL %A> A ABS
"<people>"
      "people" { wa+tu } %SUBJ N NOM
"<have>"
      "have" { AUX } %+FAUXV V PRES
"<retired>"
      "retire" { wa+me+staafu } %-FMAINV EN
"<.>"
      "."
```

Finally the words are arranged into the order required by the TL (9).

(9)
```
"<people>" { Watu } N NOM
"<aged>" { wazee } A-INFL A ABS
"<retired>" { wamestaafu } EN
"<.>" "."
```

The clean translated text is in (10).

(10)
Watu wazee wamestaafu.

5.2 Non-inflecting adjectives

Non-inflecting adjectives must be kept separate from the rest of adjectives. The tag UNINFL is made use of in writing rules (11).

(11)
```
"<The>"
      "the" %DN> DET
"<clear>"
      "clear" { safi } UNINFL %A> A ABS
"<weather>"
      "weather" { 9SG 10PL hewa } %SUBJ N NOM SG
"<has>"
      "have" { -wa na , AUX } %+FAUXV V PRES SG3
"<been>"
      "be" { wA , kuwA , AUX } %-FAUXV EN
"<lost>"
      "lose" { potezA , sozA } %-FMAINV EN
"<.>"
      "."
```

When the adjective does not inflect, the insertion of the noun class tag is blocked. Otherwise the process is the same as with inflecting adjectives.

5.3 Adjectives represented by genitive construction

This group contains adjectives, where an adjective of SL is represented by a genitive construction in TL (12).

(12)
```
"<The>"
      "the" %DN> DET
"<academic>"
      "academic" { a kitaaluma } A-MW %A> A ABS
"<staff>"
      "staff" { 2PL tumishi , 3SG 4PL kongojo , 9SG 10PL asa ,
11SG kongojo } %SUBJ N NOM
"<has>"
      "have" { -wa na , AUX } %+FAUXV V PRES SG3
"<been>"
      "be" { wA , kuwA , AUX } %-FAUXV EN
"<fired>"
      "fire" { fukuzA kazi , bimbirishA , uzulu } %-FMAINV EN
"<.>"
      "."
```

We see that the adjective academic and the verb fire are MWEs in Swahili. Tags are added at the end of each appropriate reading (13).

(13)
```
"<The>"
      "the" %DN> DET
"<academic>"
      "academic" { a kitaaluma } A-MW %A> A ABS G-2
"<staff>"
      "staff" { 2PL tumishi } %SUBJ N NOM
"<has>"
      "have" { AUX } %+FAUXV V PRES SG3
"<been>"
      "be" { AUX } %-FAUXV EN
"<fired>"
      "fire" { fukuzA kazi } %-FMAINV EN SP-2 SP-2 TAM-me PASS
"<.>"
      "."
```

Morphemes are collected into correct places...(14)

(14)
```
"<The>"
      "the" %DN> DET
"<academic>"
      "academic" { G-2+a kitaaluma } A-MW %A> A ABS
"<staff>"
      "staff" { 2PL tumishi } %SUBJ N NOM
"<has>"
      "have" { AUX } %+FAUXV V PRES SG3
"<been>"
      "be" { AUX } %-FAUXV EN
"<fired>"
      "fire" { SP-2+TAM-me+fukuz+w+A kazi} %-FMAINV EN
"<.>"
      "."
```

and converted into surface forms (15).

(15)
```
"<The>"
      "the" %DN> DET
"<academic>"
      "academic" { w+a kitaaluma } A-MW %A> A ABS
"<staff>"
      "staff" { wa+tumishi } %SUBJ N NOM
"<has>"
      "have" { AUX } %+FAUXV V PRES SG3
"<been>"
      "be" { AUX } %-FAUXV EN
"<fired>"
      "fire" { wa+me+fukuz+w+A kazi} %-FMAINV EN
```

```
"<.>"
      "."
```

Finally the correct word order is implemented (16).

(16)
```
"<staff>" { Watumishi } N NOM
"<academic>" { wa kitaaluma } A-MW A ABS
"<fired>" { wamefukuzwA kazi} %-FMAINV EN
"<.>" "."
```

There is also a construction with *enye* as the genitive connector (17).

(17)
```
"<The>"
      "the" %DN> DET
"<acrylic>"
      "acrylic" { enye nyuzi bandia } A-MW %A> A ABS
"<material>"
      "material" { 7SG 8PL tambaa , 7SG 8PL tambara } %SUBJ N NOM
SG
"<has>"
      "have" { -wa na , AUX } %+FAUXV V PRES SG3
"<been>"
      "be" { wA , kuwA , AUX } %-FAUXV EN
"<lost>"
      "lose" { potezA , sozA } %-FMAINV EN
"<.>"
      "."
```

In this solution, the word *enye* inflects, while the other two members of the MWE do not (18).

(18)
```
"<The>"
      "the" %DN> DET
"<acrylic>"
      "acrylic" { G-7+enye nyuzi bandia } A-MW %A> A ABS
"<material>"
      "material" { 7SG tambaa } %SUBJ N NOM SG
"<has>"
      "have" { AUX } %+FAUXV V PRES SG3
"<been>"
      "be" { AUX } %-FAUXV EN
"<lost>"
      "lose" { SP-7+TAM-me+potez+w+A } %-FMAINV EN
"<.>"
      "."
```

Finally we get the translation (19).

(19)
```
"<material>" { Kitambaa } N NOM SG
"<acrylic>" { chenye nyuzi bandia } A-MW A ABS
"<lost>" { kimepotezwa } EN
"<.>" "."
```

5.4 Adjective represented by verb relative structure

The fourth group includes adjectives, where the adjective in SL is represented by a verb relative structure in TL. The structure in TL may also include adverbs for more precise expression. There are a number of ways to express relative function. Here I describe two of them, one that refers to present tense and the one that refers to past tense.

An example of present tense relative is in (20). We see that the slots for subject prefix and relative prefix are left open and marked with '-', to be replaced with appropriate prefixes later. In between is the tense marker *na*, which does not change and can be written in its surface form.

(20)
```
"<The>"
      "the" %DN> DET
"<administering>"
      "administering" { -na-tawalA } A-REL %A> A ABS
"<staff>"
      "staff" { 2PL tumishi , 3SG 4PL kongojo , 9SG 10PL asa ,
11SG kongojo } %SUBJ N NOM
"<is>"
      "be" { wA , kuwA , AUX } %+FMAINV V PRES SG3
"<on>"
      "on" { katika , kwa } %ADVL PREP
"<leave>"
      "leave" { 9SG 10PL likizo , 9SG 10PL livu , 5SG 6PL likizo }
%<P N NOM SG
"<.>"
      "."
```

When morpheme tags are added and moved to the appropriate places, the result looks as in (21).

(21)
```
"<The>"
      "the" %DN> DET
"<administering>"
      "administering" { SP-2+naREL-2+tawalA } A-REL %A> A ABS
"<staff>"
      "staff" { 2PL tumishi } %SUBJ N NOM
"<is>"
      "be" { SP-2+TAM-na+kuwA } %+FMAINV V PRES SG3
"<on>"
```

```
      "on" { katika } %ADVL PREP
"<leave>"
      "leave" { 9SG likizo } %<P N NOM SG
"<.>"
      "."
```

In case the relative refers to past action, the TAM marker is *li* and the verb end has a passive marker, in this case the neutro-passive marker *ik* (22).

(22)
```
"<The>"
      "the" %DN> DET
"<completed>"
      "completed" { -li-malizikA } A-REL %A> A ABS
"<work>"
      "work" { 9SG 10PL kazi , 9SG 10PL ajira , 9SG 10PL amali }
%SUBJ N NOM SG
"<has>"
      "have" { -wa na , AUX } %+FAUXV V PRES SG3
"<been>"
      "be" { wA , kuwA , AUX } %-FAUXV EN
"<lost>"
      "lose" { potezA , sozA } %-FMAINV EN
"<.>"
      "."
```

When morphemes have been added and moved to their places, the result is as in (23).

(23)
```
"<The>"
      "the" %DN> DET
"<completed>"
      "completed" { SP-9+liREL-9+malizikA } A-REL %A> A ABS
"<work>"
      "work" { 9SG kazi } %SUBJ N NOM SG
"<has>"
      "have" { AUX } %+FAUXV V PRES SG3
"<been>"
      "be" { AUX } %-FAUXV EN
"<lost>"
      "lose" { SP-9+TAM-me+potez+w+A} %-FMAINV EN
"<.>"
      "."
```

After further cleaning, the sentence looks as in (24).

(24)
```
"<work>" { Kazi } N NOM SG
"<competence-based>" { iliyo na msingi katika ustadi } A-REL A ABS
"<lost>" { imepotezwa } EN
```

```
"<.>" "."
```

5.5 Adjective represented by description

The fifth group contains such adjectives in SL, for which there is no convenient way of expressing the meaning in TL. Instead, one has to resort to some kind of description in translation (25).

(25)
```
"<The>"
      "the" %DN> DET
"<competence-based>"
      "competence-based" { -li- na msingi katika ustadi } A-REL
%A> A ABS
"<work>"
      "work" { 9SG 10PL kazi , 9SG 10PL ajira , 9SG 10PL amali }
%SUBJ N NOM SG
"<has>"
      "have" { -wa na , AUX } %+FAUXV V PRES SG3
"<been>"
      "be" { wA , kuwA , AUX } %-FAUXV EN
"<lost>"
      "lose" { potezA , sozA } %-FMAINV EN
"<.>"
      "."
```

In this example, *competence-based* is translated with the structure 'which has its base in competence'. The verb particle *li* takes the subject prefix and relative prefix from the head noun. A further processing phase is in (26).

(26)
```
"<The>"
      "the" %DN> DET
"<competence-based>"
      "competence-based" { SP-9+liREL-9+ na msingi katika ustadi }
A-REL %A> A ABS
"<work>"
      "work" { 9SG kazi } %SUBJ N NOM SG
"<has>"
      "have" { AUX } %+FAUXV V PRES SG3
"<been>"
      "be" { AUX } %-FAUXV EN
"<lost>"
      "lose" { SP-9+TAM-me+potez+w+A } %-FMAINV EN
"<.>"
      "."
```

6 Types of MWEs

MWEs in MT can be grouped roughly on the basis of two principles. The first grouping principle is mechanical: one-to-many, many-to-one, and many-to-many. The first type does not need any isolation measures. A single word in SL has one or more glosses in TL, and at least one of them is a MWE. It is the question of disambiguation to select the correct one depending of the context.

The other two of these three types, many-to-one and many-to-many, require the isolation and treatment of the MWEs in the SL.

7 Summary of how different types of MWEs should be implemented

The examples above have been from English to Swahili MT. The same basic principles apply also to Swahili to English MT.

One-to-many:
1. Describe the word in SL and disambiguate it.
2. Add the gloss or glosses in TL, and if needed, disambiguate.
3. Process the expression to surface form in TL.

No special MWE isolation measures needed.

Many-to-one and many-to-many:
1. Describe the word in SL.
2. Isolate the MWE and give it the appropriate gloss in TL.
3. Perform the basic disambiguation.
4. Perform also the disambiguation of the MWE, if needed.
5. Process the expression to surface form in TL.

8 Frozen MWEs

Finally, a small sub-class of MWEs is the so-called frozen constructions. They do not inflect or allow other words in between.

Below is a frozen Swahili MWE analyzed word by word (27).

(27)

```
"<ana>"
      "ana" V 1/2-SG-SP VFIN { he } C:na { have } SVO
      "ana" V 1/2-SG-SP VFIN { she } C:na { have } SVO
"<kwa>"
      "kwa" PREP { with }
      "kwa" PREP { for }
      "kwa" PREP { to }
      "kwa" PREP { by }
      "kwa" PREP { on }
      "kwa" PREP { in }
```

```
      "kwa" PREP { from }
      "kwa" PREP { at }
      "kwa" GEN-CON-KWA 15-SG { of }
      "kwa" GEN-CON-KWA 17-SG { of }
"<ana>"
      "ana" V 1/2-SG-SP VFIN { he } C:na { have } SVO
      "ana" V 1/2-SG-SP VFIN { she } C:na { have } SVO
```

When the MWE is isolated, the result looks like in (28). Note that all information is on the last member of the MWE.

```
(28)
"<ana>"
      "ana" MW>>
"<kwa>"
      "kwa" MW<>
"<ana>"
      "ana" ADV <<MW { face to face } @ADVL
```

9 Conclusion

In order for machine translation to succeed, handling multiword expressions is a crucial issue. What must be treated as a multiword expression depends on the language pair in question. Each type of MWE should be handled in such a way that variation such as inflection and intervening words become possible. On the other hand, 'frozen' MWEs should be isolated as such and handled as one fixed unit.