Technical Reports in Language Technology
Report No 1, 2008
http://www.njas.helsinki.fi/salama

Multiword Expressions and Machine Translation

Arvi Hurskainen
Institute for Asian and African Studies, Box 59
FIN-00014 University of Helsinki, Finland
arvi.hurskainen@helsinki.fi

Abstract

In this paper we approach the multiword expressions (MWE) from the viewpoint of rule-based machine translation (MT). Rather than trying to find a theoretical definition to the concept, we look at it as a problem of translation result. Because rule-based MT relies on a detailed description of the source language, each translation mistake can be traced and eventually corrected. A large part of translation mistakes can be attributed to inadequate handling of MWEs. The discussion is carried out in the context of SALAMA (Swahili Language manager), which, inter alia, translates text from Swahili to English. Various types of MWEs are discussed and solutions for handling them are introduced. The performance of SALAMA in identifying MWEs was tested. Also the distribution of various types of MWEs in text is presented.

Key Words: multiword expressions, rule-based machine translation, Constraint Grammar, Swahili language, automatic dictionary compilation

Abbreviations

ADC	automatic dictionary compilation
CG	Constraint Grammar
MT	machine translation
MWE	multiword expression
SALAMA	Swahili Language Manager

1 Introduction

The identification of multiword expressions (MWE) and their appropriate handling is necessary in constructing professional tools for language manipulation. MWEs are a problem in the word alignment of parallel corpora, and various strategies for improving the result have been suggested (Ney and Popovic 2004; Schrader 2006; Sharoff et al. 2006; Tiedemann 2004). Machine translation (MT) and automatic dictionary compilation (ADC) are examples of such applications, where MWEs play a major role. The identification of MWEs in running text is a complex problem that requires more than one solution (Mendes et al. 2006; Alonge 2006; Mota et al. 2004). Although some MWEs can be isolated in the tokenizer, and then analysed as a single cluster, most of them cannot.

The discussion below is based on the experience in developing SALAMA, Swahili Language Manager, which since 1985 has grown into a comprehensive language management system (Hurskainen 1992, 1996, 2003, 2004b). Various types of MWEs are discussed and solutions for their handling are demonstrated. Because SALAMA, instead of using parallel corpora for finding translation pairs, uses language analysis and a dictionary in processing, MWEs must be described in the system. So far more than 10,000 MWEs have been included in SALAMA. These were collected from various sources including studies and domain-specific dictionaries (Chuwa 1995; Wamitila 1999, 2001). Many more, especially noun compounds, still need to be described. The work has proceeded to the phase, where all types of MWEs have been described in the way that satisfies the needs of MT and ADC.

In this paper we first discuss the ways of identifying and isolating MWEs. Then we deal with methods of marking MWEs in the way that, in addition to isolating it, each of the members of the MWE is modified so that further processing, such as translation, becomes possible. Then various types of MWEs are discussed and examples are given for each type. Finally we present evaluation results of the performance of SALAMA and give statistics on the distribution of MWEs in prose text.

2 Identification of MWEs

A useful testing environment for identifying a MWE is a rule-based MT system, where, if not otherwise defined, each word constitutes a unit to be translated to another language. In normal cases, translation is based on the lexical gloss of the word in target language and the grammatical information attached to the corresponding surface word in source language. This information provides means for converting the lexical gloss into the correct surface form.

However, there are several types of constructions, where not a single word, but a cluster of words constitutes a meaning unit. There are a number of criteria for defining a MWE. Idioms are considered typical cases of MWEs (Neumann et al. 2004), but there are many other types as well. In MT, where the aim is to produce a correct translation, the solution is often quite practical. If a cluster of words cannot be translated using word-by-word translation, it is treated as a MWE. An example of what this means is the noun compounding in English. For example, *coalition government* can be translated to Finnish by corresponding compounding as *kokoomushallitus* (kokoomus + hallitus). On the other hand, translation to Swahili is not straight forward, because Swahili uses a genitive structure *serikali ya mseto* (government of mixture). Therefore, the identification of some MWEs is task-specific.

How can MWEs be reliably identified in running text? How can true MWEs be distinguished from similar word sequences that are not MWEs? From the viewpoint of implementation, it is useful to make a distinction between 'frozen clusters' and other types of MWEs. A frozen cluster is a fixed sequence of words that constitutes a meaning unit and cannot be mixed with other constructions in the language. An example of a frozen cluster is *in spite of*. Neither *in* nor *of* in this sequence may be part of any other structure. Frozen clusters that do not need any context test for their validation can be 'glued' together in the tokenizer, which splits the text into meaning units for the morphological analyser. If the real status of a MWE candidate can only be determined on the basis of context, the identification must be done in the phase, where there is access to the linguistic analysis of the text. In other words, it must be performed after morphological analysis, where use can be made of linguistic tags in writing rules for identifying MWEs. Context tests can be used for controlling rule application.

An additional advantage in identifying MWEs after morphological analysis is that various levels of abstraction can be used in writing rules for identifying MWEs (Hurskainen 2004a). This is particularly important in cases, where one or more of the members of the MWE inflect. For example, if a verb is a member in the MWE, it may have thousands of surface forms in Swahili. The rule for identifying the MWE can be constructed using a combination of base forms and linguistic tags, so that all inflected and derived cases will be found.

3 Marking the MWEs

Because a MWE constitutes a meaning unit, each of its members must be identified and treated in the way that the result is a single lexical and semantic unit. The first thing is to find each member of the MWE. The members may be located consecutively, but often a MWE allows such words and punctuation marks in between that are not members of the MWE. Therefore, marking the MWEs reliably requires an environment, where all these restrictions can be taken into account.

In writing rules for marking the MWEs, it is important that each member will be marked, and that each member is treated in the way needed for achieving the desired result. Each of the members of the MWE has originally its own part-of-speech (POS) affiliation, a semantic meaning, and perhaps additional tags indicating inflection and derivation. The MWE has only one POS affiliation, one set of tags pertaining to this affiliation, and one or more semantic interpretations. It also sometimes turns out that the POS affiliation of the MWE is different from any of the POS categories of the members of the MWE.

The identification of MWEs can be implemented in more than one way. One obvious solution is to write substitution rules using regular expressions. This method is suitable in cases, where only the structure of the output needs to be modified, while the meaning of each member of the cluster can be derived from the source language. Some compound nouns of English can be translated to Swahili using this method. In such cases, with a few rules a large number of compound nouns can be handled.

Many types of MWEs, such as idioms, require individual treatment, especially if the meaning of the MWE is non-compositional. These types of MWEs can be described using Constraint Grammar rules (Karlsson 1995; Tapanainen 1996, 1999), especially the REPLACE rule type, which replaces the existing tags with new ones, defined by the rule.

The examples in (1-11) show how the isolation of an idiom takes place. An analysed sentence containing an idiom is in (1).

```
(1) "<*waziri>"
        "waziri" N CAP 9/6-SG { the } { *minister } HUM
        "waziri" N TITLE { *minister } HUM
"<*alipiga>"
        "piga" V CAP 1/2-SG3-SP VFIN { he/she } PAST [piga] { hit } SVO ACT
        "piga" V CAP 1/2-SG3-SP VFIN { he/she } PR:a 5/6-SG-OBJ OBJ { it } [piga] { hit }
SVO ACT
"<moyo>"
```

```
        "moyo" N 3/4-SG { the } { heart }
"<konde>"
        "konde" N 5/6-SG { the } { fist }
        "konde" N 5/6-SG { the } { plantation }
```

Because idioms are constructions that syntactically follow normal linguistic rules, they can be disambiguated using standard disambiguation rules (2).

```
(2) "<*waziri>"
        "waziri" N CAP 9/6-SG { the } { *minister } HUM
"<alipiga>"
        "piga" V 1/2-SG3-SP VFIN { he/she } PAST [piga] { hit } SVO ACT
"<moyo>"
        "moyo" N 3/4-SG { the } { heart }
"<konde>"
        "konde" N 5/6-SG { the } { plantation }
        "konde" N 5/6-SG { the } { fist }
```

We see that *konde* still has semantic ambiguity. However, this does not matter, because it is part of the idiom and will be rewritten. In the next stage we mark the structure of the idiom and give it a new interpretation. This is done with a rule that makes use of the Constraint Grammar formalism (3).

```
(3) REPLACE (<<MWE-ID { be skilful }) TARGET ("konde")
        (-2 ([piga]))
        (-1 ("moyo")) ;
```

The target is the last member of the MWE and the structure of the MWE will be encoded in it. The code <<MWE-ID means that this word and two consecutive words to the left are members of the multiword expression that is an idiom. In defining the MWE, a base form of the verb ([piga]) is used, so that also the extended verb forms will be found. In addition to the lemma, no specification for the members is used, because each of them has only one POS affiliation. If there would be ambiguity, further restrictions might be necessary. The output is in (4).

```
(4) "<*waziri>"
        "waziri" N CAP 9/6-SG { the } { *minister } HUM
"<alipiga>"
        "piga" V 1/2-SG3-SP VFIN { he/she } PAST [piga] { hit } SVO ACT
"<moyo>"
        "moyo" N 3/4-SG { the } { heart }
"<konde>"
        "konde" <<MWE-ID { be skilful }
```

In the second phase, using the structure description on the last member of the MWE, the other members of the MWE will be modified. Particularly those original tags are removed, which do not have any function in the MWE, and each member of the MWE is provided with its location tag. In the verb, such grammatical tags are retained that are needed later (5).

```
(5) "<*waziri>"
        "waziri" N 9/6-SG { the } { *minister } HUM @SUBJ
```

```
"<alipiga>"
        "piga" V 1/2-SG3-SP VFIN { he/she } PAST SVO ACT  MWE-V>>
@FMAINVtr+OBJ>
"<moyo>"
        "moyo" MWE-ID<>
"<konde>"
        "konde" <<MWE-ID { be skilful }
```

With the help of the MWE codes we can then isolate the MWE as a single unit with all necessary information for further processing (6).

```
(6) ( N 9/6-SG { the } { *minister } HUM @SUBJ )
( V 1/2-SG3-SP VFIN { he/she } PAST SVO ACT MWE-ID { be skilful } @FMAINVtr+OBJ> )
```

This structure can then be further processed using normal processing routines (7).

(7) *The Minister was skilful*

If the MWE is not composed of consecutive words, the members in between that are not members of the MWE must be retained (8).

```
(8) "<*niliwaacha>"
        "acha" V 1/2-SG1-SP VFIN { *i } PAST 1/2-PL3-OBJ OBJ { them } SVO  MWE-V>x>
@FMAINVtr+OBJ>
"<majambazi>"
        "jambazi" N 9/6-PL { the } { armed robber } HUM @OBJ
"<kwenye>"
        "kwenye" MWE-ID<x>
"<taa>"
        "taa" <x<MWE-ID { outwit }
```

We see that there is a noun object in between with original analysis, because the code marking <x<MWE-ID has left it untouched. Yet the order of words is not ideal for translation (9).

```
(9) ( V 1/2-SG1-SP VFIN { *i } PAST 1/2-PL3-OBJ OBJ { them } SVO MWE-V>x>
@FMAINVtr+OBJ> ) ( N 9/6-PL { the } { armed robber } HUM @OBJ ) ( MWE-ID<x> ) (
<x<MWE-ID { outwit } )
```

With a re-writing rule we can modify this and isolate the MWE (10).

```
(10) ( V 1/2-SG1-SP VFIN { *i } PAST 1/2-PL3-OBJ OBJ { them } SVO MWE-V>x>
@FMAINVtr+OBJ>  <x<MWE-ID { outwit } ) ( N 9/6-PL { the } { armed robber } HUM @OBJ
)
```

Now the translation succeeds (11).

(11) *I outwitted the armed robbers*

4 Types of multiword expressions

As it was stated above, rule-based MT requires a rather practical approach to defining MWEs. In fact, one has to consider the optimal solution in order to achieve the desired result. The experience shows that there are at least the following types of constructions that fall within the category of MWEs: idioms, proverbs, multiword named entities, adjectival expressions, adverbs, prepositions, and compound nouns. Below is a brief description of each type.

4.1 Idioms

Swahili uses extensively such idioms that have a verb and one or more modifiers as members of the idiom. For example, for the verb *piga* (to hit), more than 250 idioms have been found. Because most of these idioms are rendered in English as ordinary verbs, these constructions require a special treatment in MT. So far, about 2500 idioms with a verb as a member have been described in SALAMA. How such constructions are handled is described in (1 - 11) above.

4.2 Proverbs

Swahili uses quite frequently proverbs and other fixed expressions that can conveniently be described as MWEs. Although some of them can be translated using normal routines, many of them use ancient words and they do not always obey linguistic rules. And especially if one would like to find a proverbial equivalent in the target language, the treatment of the proverb as a MWE is the correct solution. In (12) is a morphologically analysed proverb.

```
(12) "<*baada_ya>"
        "baada_ya" PREP { after }
"<dhiki>"
        "dhiki" N 9/10-SG { the } { distress }
"<faraja>"
        "faraja" N 9/10-SG { the } { comfort }
```

We see that there is a frozen cluster 'baada_ya', isolated in the tokenizer. The semantic glosses give a fairly good idea of what the expression is about. The proverb is isolated in (13).

```
(13) "<*baada_ya>"
        "baada_ya" PROVERB>>
"<dhiki>"
        "dhiki" PROVERB<>
"<faraja>"
        "faraja" <<PROVERB { *after trouble there is relief }
```

We see that the final translation is attached to the last member of the proverb and no further processing is needed. After some pruning we get (14).

(14) Baada ya dhiki faraja { *After trouble there is relief* }

Sometimes the proverb has variant forms. They can be conveniently taken into account as in rule (15).

```
(15) REPLACE (<<PROVERB { *after trouble there is relief } ) TARGET ("faragha") OR
("faraja") OR ("faraji")
```

```
        (-2 ("baada_ya"))
        (-1 ("dhiki")) ;
```

This realizes all three variants of the proverb (16).

(16) Baada ya dhiki faraja/faragha/faraji { *After trouble there is relief* }

4.3 Multiword named entities

Such named entities that have more than one member and that have a translation should be treated as a single cluster for ensuring correct translation. Such entities can be described alternatively in the tokenizer or after morphological parsing. Which solution is most convenient in each case depends on whether there is variation in form, such as singular vs. plural and upper case vs. lower case alternation. If such alternation occurs, the best place to describe them is after morphological analysis. An example of the analysed named entity is in (17).

```
(17) "<*ofisa>"
        "ofisa" N CAP 9/6-SG { the } { officer } ENG HUM
"<*mkuu>"
        "kuu" CAP ADJ A-INFL 1/2-SG { main }
"<wa>"
        "wa" GEN-CON 1/2-SG { of }
"<*afya>"
        "afya" N CAP 9/10-SG { the } { health }
```

Alternations of this can be isolated with a single rule (18).

```
(18) REPLACE (<<<MW { *chief *health *officer }) TARGET ("afya")
        (-3 ("ofisa"))
        (-2 ("kuu"))
        (-1 (GEN-CON)) ;
```

The rule catches alternative forms of the named entity, including singular and plural as well as cases where one or more of the members is written with a lower case initial (19-20).

```
(19) "<*ofisa>"
        "ofisa" N 9/6-SG { the } ENG HUM MW-N>>>
"<*mkuu>"
        "kuu" MW<>>
"<wa>"
        "wa" MW<<>
"<*afya>"
        "afya" <<<MW { *chief *health *officer }
```

```
(20) "<*maofisa>"
        "ofisa" N 9/6-PL { the } ENG HUM MW-N>>>
"<*wakuu>"
        "kuu" MW<>>
"<wa>"
        "wa" MW<<>
```

```
"<afya>"
        "afya" <<<MW { *chief *health *officer }
```

In this case the morphological information of the first member is retained, because it is the head of the noun phrase and it determines the subject prefix of the main verb. It also contains information needed in translation (21).

```
(21) "<*maofisa>"
        "ofisa" N 9/6-PL { the } ENG HUM MW-N>>> @SUBJ
"<*wakuu>"
        "kuu" MW<>>
"<wa>"
        "wa" MW<<>
"<*afya>"
        "afya" <<<MW { *chief *health *officer }
"<wanafanya>"
        "fanya" V 1/2-PL3-SP VFIN { they } PR:na SVO  mwe-V> @FMAINVtr-OBJ>
"<kazi>"
        "kazi" <MWE-ID { work }
"<ofisini>"
        "ofisi" N 9/10-SG { the } { office } ENG LOC { in }
```

Translation:
The Chief Health Officers work in the office

4.4 Adjectival expressions

Because Swahili, as well as other Bantu languages, have only a small number of true adjectives, various grammatical structures are used for describing expressions, for which English uses adjectives.

4.4.1 Possessive relative

Sometimes is used a description, which starts with the possessive pronoun *enye* (22).

```
(22) "<*hizi>"
        "hizi" CAP PRON DEM :hV 9/10-PL { these }
"<ni>"
        "ni" V V-BE INIT { they } { are }
        "ni" V V-BE NOSUBJ { are }
"<nyakati>"
        "wakati" N 11/10-PL { the } { time } TIME
"<zenye>"
        "enye" POSS-PRON :OTE 9/10-PL { with }
"<msisimko>"
        "msisimko" N 3/4-SG { the } DER:verb DER:o { thrill }
"<na>"
        "na" CC { and }
"<shughuli>"
        "shughuli" N 9/10-PL { the } { activity }
```

```
"<nyingi>"
        "ingi" ADJ A-INFL 9/10-PL :INGI NOART { many }
        "ingi" ADJ A-INFL 9/10-PL :INGI NOART { much }
```

The adjectival expression is defined using a rule (23).

```
(23) REPLACE (ADJ <<<<MW { hectic }) TARGET ("ingi")
        (-4 ("enye"))
        (-3 ("msisimko"))
        (-2 ("na"))
        (-1 ("shughuli"));
```

MWE isolated:

```
"<*hizi>"
        "hizi" PRON DEM :hV 9/10-PL { these } @SUBJ
"<ni>"
        "ni" V V-BE NOSUBJ { are } @FMAINVintr-def
"<nyakati>"
        "wakati" N 11/10-PL { the } { time } TIME @PCOMPL-S
"<zenye>"
        "enye" MW>>>>
"<msisimko>"
        "msisimko" MW<>>>
"<na>"
        "na" MW<<>>
"<shughuli>"
        "shughuli" MW<<<>
"<nyingi>"
        "ingi" ADJ <<<<MW { hectic }
```

Translation:
These are hectic times

4.4.2 Adjectives with negative connotation

English has a large number of adjectives that negate the meaning of the basic adjective. These are formed by attaching a prefix such as *un-*, *in-*, *non-* or *a-* to the basic adjective. Swahili does not have a corresponding linguistic feature. In case Swahili does not have a proper adjective, the semantic content has to be described. This case is exemplified in (26 - 28).

(24) Original analysis:

```
"<*bahari>"
        "bahari" N CAP 9/10-SG { the } { ocean } PLACE
"<ilikuwa>"
        "wa" V 9/10-SG-SP VFIN { it } PAST INFMARK [wa] AUX-WA { be } SV
MONOSLB
"<si>"
        "si" V V-BE NEG NOSUBJ { not }
"<ya>"
        "ya" GEN-CON 9/10-SG { of }
```

```
"<kuendeka>"
	"endeka" V INF NO-TO [enda] { operate } PREFR SV EXT: STAT :EXT
"<kwa>"
	"kwa" GEN-CON-KWA 15-SG { with }
"<meli>"
	"meli" N 9/10-SG { the } { ship } ENG
```

MWE isolated:

```
"<*bahari>"
	"bahari" N 9/10-SG { the } { ocean } PLACE @SUBJ
"<ilikuwa>"
	"wa" V 9/10-SG-SP VFIN { it } PAST INFMARK [wa] AUX-WA { be } SV
MONOSLB @FMAINVintr
"<si>"
	"si" MW>>>>
"<ya>"
	"ya" MW<>>>
"<kuendeka>"
	"endeka" MW<<>>
"<kwa>"
	"kwa" MW<<<>
"<meli>"
	"meli" ADJ <<<<MW { unnavigable }
```

Translation:
The ocean was unnavigable

Sometimes more than one construction equals to a single translation in English. An example of this is described in the rule, which has four alternative targets (25).

```
(25) REPLACE (ADJ <<MW { unnatural }) TARGET ("desturi") OR ("asili") OR ("kawaida")
OR ("urahisi")
	(-2 ("si"))
	(-1 ("kwa") OR (GEN-CON));
```

4.4.3 Constructions with relative prefix in verb

A fairly flexible way to express qualification is to use the relative prefix attached to the verb. The prefix can be used only in some tense/aspect forms, which limits its use. The rule in (26) isolates the MWE and makes processing possible.

```
(26) REPLACE (ADJ <<<MW { underprivileged }) TARGET ("msingi") + ADV:ki
	(-3 ("nyimwa") + REL)
	(-2 ("haki"))
	(-1 ("za"));
```

Original analysis:

```
"<mtu>"
	"mtu" N 1/2-SG { the } { man }
"<aliyenyimwa>"
	"nyimwa" V 1/2-SG3-SP VFIN { he/she } PAST 1/2-SG-REL { who } [nyima] { keep
:back } SV EXT: PASS :EXT
```

```
"<haki>"
        "haki" N 9/10-PL { the } { right }
"<za>"
        "za" GEN-CON 9/10-PL { of }
"<kimsingi>"
        "msingi" ADV ADV:ki 3/4-SG { the } { basis } PLACE
```

MWE isolated:

```
"<mtu>"
        "mtu" N 1/2-SG { the } { man } @PAT
"<aliyenyimwa>"
        "nyimwa" MW>>>
"<haki>"
        "haki" MW<>>
"<za>"
        "za" MW<<>
"<kimsingi>"
        "msingi" ADJ <<<MW { underprivileged } @<NADJ
```

Translation:
The underprivileged man

There are also constructions where the relative verb construction alone, without other describing words, forms the adjectival expression (27).

(27) Original analysis:

```
"<siku>"
        "siku" N 9/10-PL { the } { day } TIME
"<zijazo>"
        "ja" V 9/10-PL-SP VFIN { they } [ja] { come } SV GEN-REL 9/10-PL { which }
```

MWE isolated:

```
"<siku>"
        "siku" N 9/10-PL { the } { day } TIME
"<zijazo>"
        "ja" ADJ { coming } NCL-PL @<NADJ
```

Translation:
The coming days

Because the adjectival expression in these cases is constructed with a verb, the verb may have also a past or future tense, depending on context. This does not cause harm, because grammatical information on the verb is deleted as part of the isolation process (28).

(28) Original analysis:

```
"<siku>"
        "siku" N 9/10-PL { the } { day } TIME
"<zitakazokuja>"
        "ja" V 9/10-PL-SP VFIN { they } FUT:taka 9/10-PL-REL { which } [ja] { come } SV
```

MWE isolated:

```
"<siku>"
        "siku" N 9/10-PL { the } { day } TIME
"<zitakazokuja>"
```

```
        "ja" ADJ ADJ-REL { coming } @<NADJ
```

Translation:
The coming days

The Swahili Language Committee in Tanzania has made attempts to replace some grammatically correct but clumsy relative structures with more adjective-like constructions (29).

```
(29) REPLACE (ADJ <MW { unstable , unsteady }) TARGET ("imara")
        (-1 ("si")) (NOT 0 MW);
```

Original analysis:

```
"<daraja>"
        "daraja" N 5/6-SG { the } { bridge }
"<si>"
        "si" V-BE NEG { is not }
"<imara>"
        "imara" ADJ A-INFL 5/6-SG { strong }
```

MWE isolated:

```
"<daraja>"
        "daraja" N 5/6-SG { the } { bridge } @SUBJ
"<si>"
        "si" MW>
"<imara>"
        "imara" ADJ <MW { unstable } @<NADJ
```

Translation:
The unstable bridge

4.5 Adverbs

Although Swahili can derive adverbs by attaching the prefix *ki-* to a noun or adjective, it also uses the preposition *kwa* for this purpose (30).

```
(30) REPLACE ( ADV <MW { perfectly } ) TARGET ("ukamilifu")
        (-1 ("kwa")) ;
```

Original analysis:

```
"<kwa>"
        "kwa" PREP { with }
"<ukamilifu>"
        "ukamilifu" N 11-SG { the } DER:fu { perfection }
```

MWE isolated:

```
"<kwa>"
        "kwa" MW>
"<ukamilifu>"
        "ukamilifu" ADV <MW { perfectly } @ADVL
```

Other types of adverbs include frozen structures, such as shown in (31).

```
(31) REPLACE (ADV <<MW { hand in hand }) TARGET ("bega")
```

```
        (-2 ("kwa"))
        (-1 ("bega"));
```

Original analysis:

```
"<bega>"
        "bega" N 5/6-SG { the } { shoulder }
"<kwa>"
        "kwa" PREP { to }
"<bega>"
        "bega" N 5/6-SG { the } { shoulder }
```

MWE isolated:

```
"<bega>"
        "bega" MW>>
"<kwa>"
        "kwa" MW<>
"<bega>"
        "bega" ADV <<MW { hand in hand } @ADVL
```

4.6 Prepositions

It is common in Swahili to construct prepositions using a construction, where a genitive particle follows a noun. These are implemented in the tokenizer, as seen in (32).

```
(32) "<mbele_ya>"
        "mbele_ya" PREP { in front of }

"<kwa_niaba_ya>"
        "kwa_niaba_ya" PREP { on behalf of }
```

4.7 Compound nouns and multiword terms

Noun compounding in the way it is done in English is quite a new phenomenon in Swahili. Such compounds have been coined mainly for domain-specific vocabularies. It should be noted, however, that the word order is different, because in Swahili the head of the compound comes first (33).

(33) Original analysis:

```
"<uchambuzi>"
        "uchambuzi" N 11-SG { the } DER:verb DER:zi { analysis }
"<msamiati>"
        "msamiati" N 3/4-SG { the } { vocabulary }
```

MWE isolated:

```
"<uchambuzi>"
        "uchambuzi" N 11-SG DER:zi MW-N>
"<msamiati>"
        "msamiati" <MW { lexicology }
```

One advantage of describing the MWEs after morphological analysis is that the description may succeed although none of the members is recognized by the morphological dictionary. The heuristic guesser gives the correct analysis to the head of the compound, and the rule does the rest (34).

```
(34) REPLACE (<MW { contrastive substitution }) TARGET ("luanuzi")
        (-1 ("ubadilishano")) ;
```

Original analysis:

```
"<ubadilishano>"
        "ubadilishano" <Heur> N 11-SG
"<luanuzi>"
        "luanuzi" <Heur> N 9/10-SG
```

MWE isolated:

```
"<ubadilishano>"
        "ubadilishano" <Heur> N 11-SG MW-N>
"<luanuzi>"
        "luanuzi" <MW { contrastive substitution }
```

The compound noun may have an adjectival modifier (35).

```
(35) REPLACE (<<MW { luminous flux density }) TARGET ("ng'aavu")
        (-2 ("msongamano"))
        (-1 ("mbubujiko")) ;
```

Original analysis:

```
"<msongamano>"
        "msongamano" N 3/4-SG { the } DER:verb { crowding }
"<mbubujiko>"
        "mbubujiko" N 3/4-SG { the } DER:verb { bubbling up }
"<ng'aavu>"
        "ng'aavu" ADJ A-INFL 9/10-SG { glittering }
```

MWE isolated:

```
"<msongamano>"
        "msongamano" N 3/4-SG { the } < MW-N>>
"<mbubujiko>"
        "mbubujiko" MW<>
"<ng'aavu>"
        "ng'aavu" <<MW { luminous flux density }
```

A common method of constructing compounds in Swahili is to use the genitive structure (36).

```
(36) REPLACE (<<MW { *revolutionary *government }) TARGET ("mapinduzi")
        (-2 ("serikali"))
        (-1 ("ya")) ;
```

Original analysis:

```
"<serikali>"
        "serikali" N 9/10-SG { the } { government }
"<ya>"
        "ya" GEN-CON 9/10-SG { of }
"<mapinduzi>"
        "mapinduzi" N 6-PLSG { the } DER:verb { revolution }
```

MWE isolated:

```
"<serikali>"
	"serikali" N 9/10-SG { the } MW-N>>
"<ya>"
	"ya" MW<>
"<mapinduzi>"
	"mapinduzi" <<MW { *revolutionary *government }
```

5 Performance of SALAMA in handling MWEs

The performance of SALAMA in handling MWEs was evaluated with a corpus of newspaper text containing 81.223 words (Corpus 1). In total there were 3.525 MWEs falling into various sub-categories. Each occurrence of a MWE was manually checked and the quality of translation was assessed. It is hard to find reliable criteria for evaluation, because often it is a question of taste, whether this or that translation is better. Therefore, attention was paid to clear mistakes. In the evaluation, only those word clusters were considered, which the system interpreted as MWEs. Because the system is rule-based, it was hard to find cases, where the interpretation totally failed. The importance of having a comprehensive system with a maximal number of cases covered is reflected in cases, where a rule applies to part of the structure, and the interpretation will be defective. An example of such a case is *chama cha upinzani*, which was translated as *opposing party*, although the more correct translation would be *opposition party*. The failure is due to the fact that there is a rule that rewrites *cha upinzani* as *opposing*, but the longer rule for the whole construction is not yet in the system. Because the order of rule application is long-first, in future this problem will be solved.

There was also one case, where the rule was too permissive. *Usambazaji kimaandishi wa maneno hayo* was translated as *the spreading by means of writing oral these* instead of *the spreading by means of writing of these words*. The translation is still clumsy, but it shows what went wrong. A rule isolated the sequence *wa maneno* and interpreted it as *oral*, which was wrong in this context. More constraining is needed for this rule.

The third type of mistakes is related to near-synonyms. We probably translate *wakati wa usiku* as *night time* rather than *nocturnal time*. Also *wananchi wa kawaida* are more naturally translated as *ordinary citizens* than *usual citizens*.

Apart from these few mistakes the translation was as it was intended to be. When also the section of noun compounding rules is complete in future and sufficient tests are made for judging the optimal constraints for each rule, the system can be expected to handle properly all types of MWEs.

6 Distribution of multiword expressions

The distribution of various types of MWEs was calculated from a corpus of fiction texts containing 1.181.450 words (Corpus2). Results of this study are displayed in Table 1 and Table 2.

Table 1. Frequency of multiword expressions in Corpus2.

	All	%	Unique	%

Words, total	1.181.450		96.519	
MWEs, total	41.684	3.5	2.441	2.5

The proportion of MWEs in Corpus 2 was 3.5 %, when all words were calculated. However, when duplicates were removed, the proportion was only 2.5 %, which means that in the average MWEs occurs in text more frequently than other word-forms.

Table 2. Types of multiword expressions in Corpus 2.

	All	%	Unique	%
MWEs, total	41.684	100	2.441	100
nouns	2.199	5.3	369	15.1
adjectives	10.914	26.2	840	34.4
adverbs	7.044	16.9	193	7.9
idioms	8.436	20.2	698	28.7
proverbs	91	0.2	72	2.9
pronouns	894	2.1	122	5.0
prepositions	10.819	26.0	142	5.8
conjunctions	1.287	3.1	5	0.2

Table 2 shows the distribution of various types of MWEs in Corpus 2. This includes the MWEs that were described in the tokenizer as well as the ones that were described with Constraint Grammar rules. There are significant differences in percentages depending on whether counting was made on the basis of all word-forms in the corpus or on the basis of unique occurrences. For example, prepositions occur frequently, although their total number is small. In fact, their number (142) is not as small as expected. The reason is that also the prepositions attached to personal pronouns were treated as frozen clusters. Pronouns include a large number of reduplicated forms.

The proportion of noun compounds is small. This is due to the fact that whereas most other types of MWEs have been described rather exhaustively in SALAMA, the work on noun compounding is still in progress. Therefore, the statistics above should not be considered final. Yet they give a rough picture of the work involved when we try to cope with MWEs in a rule-based MT system.

7 Conclusion

The appropriate handling of multiword expressions is necessary in constructing high-performing applications such as machine translation and automatic dictionary compilation. We have described how various types of MWEs can be described in the way that enhances accurate further processing. When MWEs have been described in the language analysis system, they can be automatically included as part of the corpus-based dictionary compilation thus creating an inventory of various uses of headwords. Even more important is the isolation of MWEs in machine translation, where all words and structures of the source text must be handled in the way that yields satisfactory translation. When MWEs are handled in the way described in this paper, automatic extraction and management of MWEs from raw text should not be a problem.

References

Alonge, A. 2006. The Italian Metaphor Database. *Proceedings of the 5th International Conference on Language Resources and Evaluation*, 24-26.5. 2006, Genova, pp. 455-460

Banko, M. and Moore, R. 2004. Part-of-Speech Tagging in Context. *Proceedings of the 20th International Conference on Computational Linguistics*, Geneva 23-27.8. 2004, pp. 556-561

Chuwa, A. 1995. *Phraseological Units and Dictionary: The Case of Swahili Language.* Ph.D. diss. University of Warsaw.

Hurskainen A. 1992. A Two-Level Computer Formalism for the Analysis of Bantu Morphology. An Application to Swahili. *Nordic Journal of African Studies* 1(1): 87-122. Retrieved November 10, 2007, from http://www.njas.helsinki.fi

Hurskainen A. 1996. Disambiguation of morphological analysis in Bantu languages. *Proceedings of COLING-96*, pp. 568-573.

Hurskainen A. 2003. New Approaches in Corpus-Based Computational Lexicography. *Lexikos* 13 (AFRILEX-reeks/series 13: 2003): 111-132.

Hurskainen A. 2004a. Optimizing Disambiguation in Swahili. In *Proceedings of COLING-04, The 20th International Conference on Computational Linguistics*, Geneva 23-27.8. 2004. Pp. 254-260. Retrieved November 10, 2007, from http://www.njas.helsinki.fi

Hurskainen, A. 2004b. Swahili Language Manager: A Storehouse for Developing Multiple Computational Applications. *Nordic Journal of African Studies* 13(3): 363–397. Retrieved November 10, 2007, from http://www.njas.helsinki.fi

Karlsson, F. 1995. Designing a parser for unrestricted text. Karlsson, F. et al (Eds.), *Constraint Grammar: A Language-Independent System for Parsing Unrestricted Text*: 1-40. Berlin: Mouton de Gryuter.

Mendes, A., Antunes, S., Nascimento, M., Casteleiro, J., Pereira, L. and Sá, T. 2006. COMBINA-PT: A Large Corpus-extracted and Hand-checked Lexical Database of Portuguese Multiword Expressions. *Proceedings of the 5th International Conference on Language Resources and Evaluation*, 24-26.5. 2006, Genova, pp. 1900-05

Mota, C., Carvalho, P., and Ranchhod, E. 2004.Multiword Lexical Acquisition and Dictionary Formalization. *Proceeding of the Workshop on Enhancing and Using Electronic Dictionaries*, in conjunction with COLING 2004, 29.8. 2004, Geneve, pp. 73-76

Neumann, G., Fellbaum, C., Geyken, A., Herold, A., Hümmer, C., Körner, F., Kramer, U., Krell, K., Sokirko, A., Stantcheva D. and Stathi, E. 2004. A Corpus-based Lexical Resource of German Idioms. *Proceeding of the Workshop on Enhancing and Using Electronic Dictionaries*, in conjunction with COLING 2004, 29.8. 2004, Geneve, pp. 48-52

Ney, H. and Popovic, M. 2004. Improving Word Alignment Quality using Morpho-syntactic Information. *Proceedings of the 20th International Conference on Computational Linguistics*, Geneva 23-27.8. 2004, pp. 310-314

Schrader. B. 2006. Non-probabilistic alignment of rare German and English nominal expressions. *Proceedings of the 5th International Conference on Language Resources and Evaluation*, 24-26.5. 2006, Genova, pp. 1274-77

Sharoff, S., Babych, B. and Hartley A. 2006. Using collocations from comparable corpora to find translation equivalents. *Proceedings of the 5th International Conference on Language Resources and Evaluation*, 24-26.5. 2006, Genova, pp. 465-470

Tapanainen, P. 1996. *The Constraint Grammar Parser CG-2.* Publications No. 27. Department of General Linguistics, University of Helsinki.

Tapanainen, P. 1999. *Parsing in two frameworks: finite-state and functional dependency grammar*. Ph.D. thesis, Department of General Linguistics, University of Helsinki.

Tiedemann, J. 2004. Word to word alignment strategies. *Proceedings of the 20th International Conference on Computational Linguistics*, Geneva 23-27.8. 2004, pp. 212-218

Wamitila, K.W. 1999. *Kamusi ya Misemo na Nahau*. Nairobi: Longhorn Publishers.

Wamitila, K.W. 2001. *Kamusi ya Methali*. Nairobi: Longhorn Publishers.