Technical Reports on Language Technology
Report No 58, 2020
http://www.njas.helsinki.fi/salama

Ad hoc compounds in English to Swahili machine translation

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

Abstract

English compounds are normally composed of two or more consecutive words. It is also possible to form compounds using dash as a connector between words. Such constructions are difficult to handle, because they are not in the lexicon, and their interpretation must be carried out using heuristic guessing. However, it is possible to handle the members of the compound separately, and guessing can be avoided. In the Technical Report No. 57 I discussed the problem in English to Finnish machine translation. In this report I will discuss it from the viewpoint of English to Swahili machine translation. Although the basic approach in both is the same, the implementation is language-specific.

Key Words: *compounding, morphological analysis, machine translation.*

1 Introduction

A detailed introduction to the problem discussed here is in Technical Report No. 57, and there is no need to repeat it here. I will describe here the measures required for achieving correct translation of each type of *ad hoc* compounds, when the target language is Swahili.

We need a separate version of the morphological lexicon of English, where the Swahili glosses are added to those two sub-lexicons, where commonly occurring last parts of compounds are listed. The lexical entries are the same as was described in report No. 57, but the glosses are different. They are different not only as words. Also the structure of the compound must be formulated according to the rules of the target language. This involves the correct position of the compound parts, and also the class affiliation in the noun class system.

Below I will show in detail, how the translation proceeds.

2 Ad hoc compounds expressing measure

A fairly frequent is the compound type, where the last part of the compound expresses the unit, and the first part expresses the number of that unit. Examples are *three-day, seven-week, twelve-month, eight-foot, nine-metre,* etc.

In these examples, the first part is a numeral. It also can be a true number, such as *5-day, 7-week, 13-inch*. Especially if the count number is big, numbers are often used instead of numerals.

In this group of compounds, the last member is noun, expressing some measure. The list of this compound type, extracted from the morphological lexicon, is in (1).

(1)
```
-day # "= [ -a siku >10 ]";
-floor # "= [ -a ghorofa >10 ]";
-foot # "= [ -a futi >10 ]";
-goal # "= [ -a shabaha >10 ]";
-hour # "= [ -a saa >10 ]";
-inch # "= [ -a inchi >10 ]";
-long # "= [ -eneye urefu wa > ]";
-metre # "= [ -a mita >10 ]";
-million # "= [ -a milioni >10 ]";
-minute # "= [ -a dakika >10 ]";
-month # "= [ -a miezi >4 ]";
-page # "= [ -a kurasa >10 ]";
-pound # "= [ -a paundi >10 ]";
-phase # "= [ -a hatua >10 ]";
-stage # "= [ -a steji >10 ]";
-step # "= [ -a hatua >10 ]";
-storey # "= [ -a ghorofa >10 ]";
-story # "= [ -a ghorofa >10 ]";
-week # "= [ -a wiki >10 ]";
-year # "= [ -a miaka >4 ]";
-years # "year [ -a miaka >4 ]";
```

The first part on the line is the last section of the compound. The hash '#' indicates that the execution stops here, and no further suffixes can follow. Between the quotes is the interpretation of the string, in this case the gloss plus further instructions. The equal mark '=' means that the string in the beginning of the line will be copied as the stem. The angle bracket '>' means that the first member of the compound will be located after the gloss of the last part, that is, in the opposite order. The number after the angle bracket shows the noun class of the first member, in case it is an inflecting numeral or adjective.

In (2) are analysis results of compounds formed with some of the words in list (1).

(2)
```
"<three-day>"
      "three-day" A  [ -a siku >10 ]

"<four-month>"
      "four-month" A  [ -a miezi >4 ]

"<six-year>"
      "six-year" A  [ -a miaka >4 ]

"<five-month>"
      "five-month" A  [ -a miezi >4 ]

"<five-year>"
      "five-year" A  [ -a miaka >4 ]
```

```
"<six-day>"
      "six-day" A  [ -a siku >10 ]

"<six-month>"
      "six-month" A  [ -a miezi >4 ]

"<eight-foot>"
      "eight-foot" A  [ -a futi >10 ]

"<eight-metre>"
      "eight-metre" A  [ -a mita >10 ]
```

The compounds were interpreted as adjectives, because in context they behave as adjectives, although none of the units of the compound is an adjective. Swahili, as Bantu languages in general, have very few true adjectives. Therefore, adjectives must be formed using genitive or relative constructions. In all the cases in (2), adjectives are formed with the particle *a*, which will have the appropriate class prefix. After the gloss, the angle bracket '>' shows that the gloss of the first part, to be inserted later, will locate after the current gloss. The numbers 10 and 4 mean that the gloss, to be inserted later, must inflect according to the class, marked by the number.

After adding the rest of glosses, the result is as in (3).

(3)

```
"<three-day>"
      "three-day" { -tatu , NOAFF tatu } INFL PL A  [ -a siku >10
]
"<four-month>"
      "four-month" { -nne , NOAFF nne } INFL PL A  [ -a miezi >4 ]
"<six-year>"
      "six-year" { sita } PL A  [ -a miaka >4 ]
"<five-month>"
      "five-month" { -tano } INFL PL A  [ -a miezi >4 ]
"<five-year>"
      "five-year" { -tano } INFL PL A  [ -a miaka >4 ]
"<six-day>"
      "six-day" { sita } PL A  [ -a siku >10 ]
"<six-month>"
      "six-month" { sita } PL A  [ -a miezi >4 ]
"<eight-foot>"
      "eight-foot" { -nane , NOAFF nane } INFL PL A  [ -a futi >10
]
"<eight-metre>"
      "eight-metre" { -nane , NOAFF nane } INFL PL A  [ -a mita
>10 ]
```

Now each compound has two sets of glosses, one for the first part and one for the last part. Some have more than one gloss alternative. We need only the first one and remove the rest (4).

(4)
```
"<three-day>"
      "three-day" { -tatu } INFL PL A  [ -a siku >10 ]
"<four-month>"
      "four-month" { -nne } INFL PL A  [ -a miezi >4 ]
"<six-year>"
      "six-year" { sita } PL A  [ -a miaka >4 ]
"<five-month>"
      "five-month" { -tano } INFL PL A  [ -a miezi >4 ]
"<five-year>"
      "five-year" { -tano } INFL PL A  [ -a miaka >4 ]
"<six-day>"
      "six-day" { sita } PL A  [ -a siku >10 ]
"<six-month>"
      "six-month" { sita } PL A  [ -a miezi >4 ]
"<eight-foot>"
      "eight-foot" { -nane } INFL PL A  [ -a futi >10 ]
"<eight-metre>"
      "eight-metre" { -nane } INFL PL A  [ -a mita >10 ]
```

The numerals prefixed with a dash '-' inflect, and those without do not. As the codes *>4* and *>10* show, the new gloss must come after the old gloss (5).

(5)
```
"<three-day>"
      "three-day" { -a siku >10 -tatu } INFL PL A
"<four-month>"
      "four-month" { -a miezi >4 -nne } INFL PL A
"<six-year>"
      "six-year" { -a miaka >4 sita } PL A
"<five-month>"
      "five-month" { -a miezi >4 -tano } INFL PL A
"<five-year>"
      "five-year" { -a miaka >4 -tano } INFL PL A
"<six-day>"
      "six-day" { -a siku >10 sita } PL A
"<six-month>"
      "six-month" { -a miezi >4 sita } PL A
"<eight-foot>"
      "eight-foot" { -a futi >10 -nane } INFL PL A
"<eight-metre>"
      "eight-metre" { -a mita >10 -nane } INFL PL A
```

Now the glosses are in correct order. The inflection tag is in front of the numeral. The inflection rules apply only to those numerals, which inflect (6).

(6)
```
"<three-day>"
      "three-day" { -a siku tatu } INFL PL A
```

```
"<four-month>"
      "four-month" { -a miezi minne } INFL PL A
"<six-year>"
      "six-year" { -a miaka sita } PL A
"<five-month>"
      "five-month" { -a miezi mitano } INFL PL A
"<five-year>"
      "five-year" { -a miaka mitano } INFL PL A
"<six-day>"
      "six-day" { -a siku sita } PL A
"<six-month>"
      "six-month" { -a miezi sita } PL A
"<eight-foot>"
      "eight-foot" { -a futi nane } INFL PL A
"<eight-metre>"
      "eight-metre" { -a mita nane } INFL PL A
```

If the inflecting numeral needed a prefix of class 4, it was inserted. In case the class was 10, nothing was added, because in this class the class prefix is added only in restricted cases. Therefore, we have *-a siku tatu*.

We put the compounds into context for seeing how they work in further processing (7).

(7)

```
"<three-day>"
      "three-day" { -a siku tatu } INFL PL A
      "three-day" { -tatu } INFL PL A
"<work>"
      "work { fanyA kazi } HUM-V" V vt vi INF
      "work { fanyA kazi } HUM-V" V vt vi IMP
      "work { fanyA kazi } HUM-V" V vt vi PRES SG1
      "work { fanyA kazi } HUM-V" V vt vi PRES SG2/PL2
      "work { fanyA kazi } HUM-V" V vt vi PRES PL1
      "work { fanyA kazi } HUM-V" V vt vi PRES PL3
      "work { 9SG 10PL kazi }" N SG
"<.>"
      "." **CLB
"<four-month>"
      "four-month" { -a miezi minne } INFL PL A
"<period>"
      "period { 7SG 8PL pindi }" N SG
"<.>"
      "." **CLB
"<six-year>"
      "six-year" { -a miaka sita } PL A
"<war>"
      "war { 9SG 10PL vita }" N SG
"<.>"
      "." **CLB
"<eight-foot>"
```

```
      "eight-foot" { -a futi nane } INFL PL A
"<rope>"
      "rope { shawishi } SVO" V vt INF
      "rope { shawishi } SVO" V vt IMP
      "rope { shawishi } SVO" V vt PRES SG1
      "rope { shawishi } SVO" V vt PRES SG2/PL2
      "rope { shawishi } SVO" V vt PRES PL1
      "rope { shawishi } SVO" V vt PRES PL3
      "rope { 9SG 10PL kamba }" N SG
"<.>"
      "." **CLB
"<eight-metre>"
      "eight-metre" { -a mita nane } INFL PL A
"<house>"
      "house { hifadhi } SVO" V vt INF
      "house { hifadhi } SVO" V vt IMP
      "house { hifadhi } SVO" V vt PRES SG1
      "house { hifadhi } SVO" V vt PRES SG2/PL2
      "house { hifadhi } SVO" V vt PRES PL1
      "house { hifadhi } SVO" V vt PRES PL3
      "house { 9SG 10PL nyumba }" N SG
"<.>"
      "." **CLB
```

We disambiguate the text (8).

(8)

```
"<three-day>"
      "three-day" { -a siku tatu } INFL PL A
"<work>"
      "work { 9SG 10PL kazi }" N SG
"<.>"
      "." **CLB
"<four-month>"
      "four-month" { -a miezi minne } INFL PL A
"<period>"
      "period { 7SG 8PL pindi }" N SG
"<.>"
      "." **CLB
"<six-year>"
      "six-year" { -a miaka sita } PL A
"<war>"
      "war { 9SG 10PL vita }" N SG
"<.>"
      "." **CLB
"<eight-foot>"
      "eight-foot" { -a futi nane } INFL PL A
"<rope>"
      "rope { 9SG 10PL kamba }" N SG
"<.>"
      "." **CLB
```

```
"<eight-metre>"
      "eight-metre" { -a mita nane } INFL PL A
"<house>"
      "house { 9SG 10PL nyumba }" N SG
"<.>"
      "." **CLB
```

Now when we have the compounds in context and we have disambiguated the readings, we can add the needed inflection codes (9).

(9)
```
"<three-day>"
      "three-day" { -a siku tatu } INFL PL A 9SG
"<work>"
      "work { 9SG 10PL kazi }" N SG
"<.>"
      "." **CLB
"<four-month>"
      "four-month" { -a miezi minne } INFL PL A 7SG
"<period>"
      "period { 7SG 8PL pindi }" N SG
"<.>"
      "." **CLB
"<six-year>"
      "six-year" { -a miaka sita } PL A 9SG
"<war>"
      "war { 9SG 10PL vita }" N SG
"<.>"
      "." **CLB
"<eight-foot>"
      "eight-foot" { -a futi nane } INFL PL A 9SG
"<rope>"
      "rope { 9SG 10PL kamba }" N SG
"<.>"
      "." **CLB
"<eight-metre>"
      "eight-metre" { -a mita nane } INFL PL A 9SG
"<house>"
      "house { 9SG 10PL nyumba }" N SG
"<.>"
      "." **CLB
```

Each compound has an inflection code at the end of the reading. Then we move the inflection codes in front of the word to be inflected (10).

(10)
```
"<three-day>"
      "three-day" { 9SG -a siku tatu } INFL PL A
"<work>"
      "work { 9SG 10PL kazi }" N SG
```

```
"<.>"
      "." **CLB
"<four-month>"
      "four-month" { 7SG -a miezi minne } INFL PL A
"<period>"
      "period { 7SG 8PL pindi }" N SG
"<.>"
      "." **CLB
"<six-year>"
      "six-year" { 9SG -a miaka sita } PL A
"<war>"
      "war { 9SG 10PL vita }" N SG
"<.>"
      "." **CLB
"<eight-foot>"
      "eight-foot" { 9SG -a futi nane } INFL PL A
"<rope>"
      "rope { 9SG 10PL kamba }" N SG
"<.>"
      "." **CLB
"<eight-metre>"
      "eight-metre" { 9SG -a mita nane } INFL PL A
"<house>"
      "house { 9SG 10PL nyumba }" N SG
"<.>"
      "." **CLB
```

The genitive connector -a will get its prefix according to the inflection tag (11).

(11)

```
"<three-day>"
      "three-day" { ya siku tatu } INFL PL A
"<work>"
      "work { 9SG 10PL kazi }" N SG
"<.>"
      "." **CLB
"<four-month>"
      "four-month" { cha miezi minne } INFL PL A
"<period>"
      "period { 7SG 8PL pindi }" N SG
"<.>"
      "." **CLB
"<six-year>"
      "six-year" { ya miaka sita } PL A
"<war>"
      "war { 9SG 10PL vita }" N SG
"<.>"
      "." **CLB
"<eight-foot>"
      "eight-foot" { ya futi nane } INFL PL A
"<rope>"
```

```
      "rope { 9SG 10PL kamba }" N SG
"<.>"
      "." **CLB
"<eight-metre>"
      "eight-metre" { ya mita nane } INFL PL A
"<house>"
      "house { 9SG 10PL nyumba }" N SG
"<.>"
      "." **CLB
```

The singular/plural ambiguity must be resolved according to the tag (12).

(12)

```
"<three-day>"
      "three-day" { ya siku tatu } INFL PL A
"<work>"
      "work { 9SG kazi }" N SG
"<.>"
      "." **CLB
"<four-month>"
      "four-month" { cha miezi minne } INFL PL A
"<period>"
      "period { 7SG pindi }" N SG
"<.>"
      "." **CLB
"<six-year>"
      "six-year" { ya miaka sita } PL A
"<war>"
      "war { 9SG vita }" N SG
"<.>"
      "." **CLB
"<eight-foot>"
      "eight-foot" { ya futi nane } INFL PL A
"<rope>"
      "rope { 9SG kamba }" N SG
"<.>"
      "." **CLB
"<eight-metre>"
      "eight-metre" { ya mita nane } INFL PL A
"<house>"
      "house { 9SG nyumba }" N SG
"<.>"
      "." **CLB
```

The tag 9SG is converted to surface form, which in this case is zero. The tag 7SG is converted to *ki* (13).

(13)

```
"<three-day>"
      "three-day" { ya siku tatu } INFL PL A
```

```
"<work>"
      "work { kazi }" N SG
"<.>"
      "." **CLB
"<four-month>"
      "four-month" { cha miezi minne } INFL PL A
"<period>"
      "period { kipindi }" N SG
"<.>"
      "." **CLB
"<six-year>"
      "six-year" { ya miaka sita } PL A
"<war>"
      "war { vita }" N SG
"<.>"
      "." **CLB
"<eight-foot>"
      "eight-foot" { ya futi nane } INFL PL A
"<rope>"
      "rope { kamba }" N SG
"<.>"
      "." **CLB
"<eight-metre>"
      "eight-metre" { ya mita nane } INFL PL A
"<house>"
      "house { nyumba }" N SG
"<.>"
      "." **CLB
```

The final translation is in (14).

(14)
kazi ya siku tatu.
kipindi cha miezi minne.
vita ya miaka sita.
kamba ya futi nane.
nyumba ya mita nane.

3 Ad hoc compounds with adjective as last member

Another group of *ad hoc* compounds has an adjective as the last member. It is not as homogenous as the group above, and there are several ways of implementing the translation. Also here, the special problem is the scarcity of true adjectives in Swahili. In (15) is this mixed list, extracted from the morphological lexicon.

(15)

```
-backed # "= [ -li-egemewa na > ]";
-baked # "= [ -li-okwa > ]";
-based # "= [ -li- na msingi katika > ]";
```

```
-catching # "= [ -na-daka < ]";
-causing # "= [ -na-sababisha > ]";
-century # "= [ -a karne > ]";
-changing # "= [ -na-badilisha > , -na-badilika < ]";
-class # "= [ -a darasa la > , -a darasa >5 ]";
-controlled # "= [ -li-dhibitiwa na > ]";
-deep # "= [ -a kina cha > , -a kina >7 ]";
-degree # "= [ -a daraja la > , -a daraja >5 ]";
-dominated # "= [ -li-tawaliwa na > ]";
-elect # "= [ -li-chaguliwa < ]";
-faced # "= [ -enye uso wa > , -enye uso >11 ]";
-filled # "= [ -li-jaa > ]";
-free # "= [ -si- na > ]";
-fuelled # "= [ -li-endeshwa na > ]";
-funded # "= [ -li-lipwa na > ]";
-game # "= [ -enye michezo ya > , -enye michezo >4 ]";
-handed # "= [ -enye mkono wa > , -enye mkono >3 ]";
-held # "= [ -li-shikwa na > ]";
-inspired # "= [ -li-sajiishwa na > ]";
-led # "= [ -li-ongozwa na > ]";
-level # "= [ -a ngazi ya > , -a ngazi >9 ]";
-like # "= [ -na-fananisha > ]";
-lived # "= [ -enye maisha > ]";
-living # "= [ -na-ishi > ]";
-looking # "= [ -enye sura ya > , -enye sura >9 ]";
-made # "= [ -li-undwa > ]";
-man # "= [ -a watu wa > , -a watu >2 ]";
-minded # "= [ -enye nia ya > , -enye nia >9 ]";
-needed # "= [ -li-takiwa > ]";
-year-old # "= [ -enye umri wa miaka >4 ]";
-old # "= [ -enye umri wa > , -enye umri >11 ]";
-olds # "= [ -enye umri wa > , -enye umri >11 ]";
-operated # "= [ -li-endeshwa na > ]";
-owned # "= [ -li-milikiwa na > ]";
-page # "= [ -a kurasa > ]";
-paid # "= [ -li-lipwa > ]";
-person # "= [ -a watu wa > , -a watu >2 ]";
-point # "= [ -enye ponti ya > , -enye pointi >9 ]";
-ranked # "= [ -enye wadhifa wa > , -enye wadifa >11]";
-ranking # "= [ -enye wadhifa wa > , -enye wadifa >11]";
-registered # "= [ -li-sajiliwa > ]";
-related # "= [ -li- na uhusiano na > ]";
-rich # "= [ -enye -- -ingi ]";
-round # "= [ -a mazunguko < ]";
-scale # "= [ -a skeli ya > , -a skeli >9 ]";
-shaped # "= [ -enye muundo wa > ]";
-sided # "= [ -enye pande > ]";
-sized # "= [ -enye saizi ya > , -enye saizi >9 ]";
-sourced # "= [ -enye asili ya > , -enye asili >9 ]";
-sponsored # "= [ -li-fadhilishwa na > ]";
-stricken # "= [ -li-pigwa na > ]";
```

```
-style # "= [ -enye staili ya > , -enye staili >9 ]";
-tall # "= [ -enye urefu wa > - -enye urefu >11 ]";
-thick # "= [ -enye unene wa > , -enye unene >11 ]";
-time # "= [ -a muda wa > , -a mara > ]";
-traded # "= [ -li-uzwa > ]";
-wide # "= [ -enye upana wa > -enye upana >11 ]";
-winning # "= [ -na-shinda > ]";
-won # "= [ -li-shinda > ]";
-working # "= [ -na-fanya kazi > ]";
```

There are three ways of constructing adjectival expressions. One method is to use the genitive particle *a*, which gets its prefix according to the noun class of its head noun. Another method is to use the relative particle *enye*, which also requires a class prefix. These two methods are very similar and often either of them can be used.

The third method is more complex. In it, the adjectival expression is formed of a verb with relative prefix. In (16) is a schematic structure of forming adjectival expressions using this method.

(16)

SP	TAM	REL	STEM
-	*na*	-	*uzwa* > *-na-uzwa*
-	*li*	-	*uzwa* > *-li-uzwa*
-	*si*	-	*uzwa* > *-si-uzwa*
-	*li*	-	*na* > *-li- na*
-	*si*	-	*na* > *-si- na*

The three first examples are structures of the verb *uza* (to sell). The two last structures express possession, and no true verb is involved.

The slot for the class concord is marked with a dash '-', and it varies according to the noun class of its head.

In (15), the angle bracket shows the position of the first member of the compound in target language.

Various types of examples of compounds are in (17), shown after morphological analysis.

(17)

```
"<hill-based>"
      "hill-based" A [ -li- na msingi katika > ]

"<smoke-causing>"
      "smoke-causing" A [ -na-sababisha > ]

"<life-changing>"
      "life-changing" A [ -na-badilisha > , -na-badilika < ]

"<state-controlled>"
      "state-controlled" PREFR A [ -li-dhibitiwa na > ]
```

```
"<male-dominated>"
      "male-dominated" A [ -li-tawaliwa na > ]

"<president-elect>"
      "president-elect" N SG
      "president-elect" A [ -li-chaguliwa < ]

"<sex-fuelled>"
      "sex-fuelled" A [ -li-endeshwa na > ]

"<video-inspired>"
      "video-inspired" A [ -li-sajiishwa na > ]

"<opposition-led>"
      "opposition-led" A [ -li-ongozwa na > ]

"<investment-related>"
      "investment-related" A [ -li- na uhusiano na > ]

"<funnel-shaped>"
      "funnel-shaped" A [ -enye muundo wa > ]

"<state-sponsored>"
      "state-sponsored" PREFR A [ -li-fadhilishwa na > ]

"<nation-wide>"
      "nation-wide" A [ -enye upana wa > -enye upana >11 ]

"<knee-deep>"
      "knee-deep" A [ -a kina cha > , -a kina >7 ]
```

Some examples have two alternative glosses. The first one is for compounds, where the first member is a noun. The second gloss is for cases, where the first member is an adjective or numeral. The selection must be made when the gloss of the first member is known (18).

(18)

```
"<hill-based>"
      "hill-based" { 7SG 8PL lima , 7SG 8PL duta , 7SG 8PL limbo }
A [ -li- na msingi katika > ]

"<smoke-causing>"
      "smoke-causing" { 3SG oshi } A [ -na-sababisha > ]

"<life-changing>"
      "life-changing" { 6PLSG isha , 9SG 10PL hayati , 11SG hai }
A [ -na-badilisha > , -na-badilika < ]
```

```
"<state-controlled>"
      "state-controlled" { 9SG 10PL serikali , 9SG 10PL hali , 9SG
10PL dola } PREFR A [ -li-dhibitiwa na > ]

"<male-dominated>"
      "male-dominated" { 1SG 2PL anaume , 9SG 6PL dume , 1SG 2PL
anamme } HUM A [ -li-tawaliwa na > ]

"<president-elect>"
      "president-elect { 9SG rais aliyechaguliwa } HUM" N SG
      "president-elect { 9SG rais aliyechaguliwa } HUM" A [ -li-
chaguliwa < ]

"<sex-fuelled>"
      "sex-fuelled" { 9SG 10PL jinsia , 11SG jinsia } A [ -li-
endeshwa na > ]

"<video-inspired>"
      "video-inspired" { 9SG 10PL video } A [ -li-sajiishwa na > ]

"<opposition-led>"
      "opposition-led" { 11SG pinzani , 5SG 6PL pingano , 5SG 6PL
kinzano } A [ -li-ongozwa na > ]

"<investment-related>"
      "investment-related" { 11SG wekezaji , 7SG 8PL tegauchumi }
A [ -li- na uhusiano na > ]

"<funnel-shaped>"
      "funnel-shaped" { 9SG 10PL faneli , 3SG 4PL lizamu } A [ -
enye muundo wa > ]
      "funnel-shaped" A
"<state-sponsored>"
      "state-sponsored" { 9SG 10PL serikali , 9SG 10PL hali , 9SG
10PL dola } PREFR A [ -li-fadhilishwa na > ]

"<nation-wide>"
      "nation-wide" { 5SG 6PL taifa , 5SG 6PL dola } A [ -enye
upana wa > , -enye upana >11 ]

"<knee-deep>"
      "knee-deep" { 5SG 6PL goti , 5SG 6PL futi , 5SG 6PL ondo } A
[ -a kina cha > , -a kina >7 ]
      "knee-deep" A
```

We leave only the first gloss of the first member of the compound, because the other ones are redundant (19).

(19)

```
"<hill-based>"
      "hill-based" { 7SG 8PL lima } A [ -li- na msingi katika > ]
```

```
"<smoke-causing>"
      "smoke-causing" { 3SG oshi } A [ -na-sababisha > , -na-
sababika < ]
"<life-changing>"
      "life-changing" { 6PLSG isha } A [ -na-badilisha > , -na-
badilika < ]
"<state-controlled>"
      "state-controlled" { 9SG 10PL serikali } PREFR A [ -li-
dhibitiwa na > ]
"<male-dominated>"
      "male-dominated" { 1SG 2PL anaume } HUM A [ -li-tawaliwa na
> ]
"<president-elect>"
      "president-elect" SG
      "president-elect { 9SG rais aliyechaguliwa } HUM" N SG
      "president-elect { 9SG rais } HUM" A [ -li-chaguliwa < ]
"<sex-fuelled>"
      "sex-fuelled" { 9SG 10PL jinsia } A [ -li-endeshwa na > ]
"<video-inspired>"
      "video-inspired" { 9SG 10PL video } A [ -li-sajiishwa na > ]
"<opposition-led>"
      "opposition-led" { 11SG pinzani } A [ -li-ongozwa na > ]
"<investment-related>"
      "investment-related" { 11SG wekezaji } A [ -li- na uhusiano
na > ]
"<funnel-shaped>"
      "funnel-shaped" { 9SG 10PL faneli } A [ -enye muundo wa > ]
"<state-sponsored>"
      "state-sponsored" { 9SG 10PL serikali } PREFR A [ -li-
fadhilishwa na > ]
"<nation-wide>"
      "nation-wide" { 5SG 6PL taifa } A [ -enye upana wa > , -enye
upana >11 ]
"<knee-deep>"
      "knee-deep" { 5SG 6PL goti } A [ -a kina cha > , -a kina >7
]
```

Now we disambiguate the glosses of the second member of the compound. In case the first member is a noun, we select the first gloss. We see that in all examples this is the case.

The compounds *smoke-causing* and *life-changing* in themselves are ambiguous, because they can be either nouns or adjectives. Only the context helps in disambiguating these compounds.

The compound *president-elect* is a noun, and it does not fit into this group. Either its adjective definition should be changed into noun, or it should be handled directly in the morphological lexicon.

The fully disambiguated result is in (20).

(20)

```
"<hill-based>"
```

```
      "hill-based" { 7SG 8PL lima } A [ -li- na msingi katika > ]
"<smoke-causing>"
      "smoke-causing" { 3SG oshi } A [ -na-sababisha > ]
"<life-changing>"
      "life-changing" { 6PLSG isha } A [ -na-badilisha > ]
"<state-controlled>"
      "state-controlled" { 9SG 10PL serikali } PREFR A [ -li-
dhibitiwa na > ]
"<male-dominated>"
      "male-dominated" { 1SG 2PL anaume } HUM A [ -li-tawaliwa na
> ]
"<president-elect>"
      "president-elect { 9SG rais aliyechaguliwa } HUM" N SG
      "president-elect { 9SG rais aliyechaguliwa } HUM" A [ -li-
chaguliwa < ]
"<sex-fuelled>"
      "sex-fuelled" { 9SG 10PL jinsia } A [ -li-endeshwa na > ]
"<video-inspired>"
      "video-inspired" { 9SG 10PL video } A [ -li-sajiishwa na > ]
"<opposition-led>"
      "opposition-led" { 11SG pinzani } A [ -li-ongozwa na > ]
"<investment-related>"
      "investment-related" { 11SG wekezaji } A [ -li- na uhusiano
na > ]
"<funnel-shaped>"
      "funnel-shaped" { 9SG 10PL faneli } A [ -enye muundo wa > ]
"<state-sponsored>"
      "state-sponsored" { 9SG 10PL serikali } PREFR A [ -li-
fadhilishwa na > ]
"<nation-wide>"
      "nation-wide" { 5SG 6PL taifa } A [ -enye upana wa > ]
"<knee-deep>"
      "knee-deep" { 5SG 6PL goti } A [ -a kina cha > ]
```

The plural tags are removed as redundant (21).

(21)

```
"<hill-based>"
      "hill-based" { 7SG lima } A [ -li- na msingi katika > ]
"<smoke-causing>"
      "smoke-causing" { 3SG oshi } A [ -na-sababisha > ]
"<life-changing>"
      "life-changing" { 6PLSG isha } A [ -na-badilisha > ]
"<state-controlled>"
      "state-controlled" { 9SG serikali } PREFR A [ -li-dhibitiwa
na > ]
"<male-dominated>"
      "male-dominated" { 1SG anaume } HUM A [ -li-tawaliwa na > ]
"<president-elect>"
      "president-elect { 9SG rais aliyechaguliwa } HUM" N SG
```

```
      "president-elect { 9SG rais aliyechaguliwa } HUM" A [ -li-
chaguliwa < ]
"<sex-fuelled>"
      "sex-fuelled" { 9SG jinsia } A [ -li-endeshwa na > ]
"<video-inspired>"
      "video-inspired" { 9SG video } A [ -li-sajiishwa na > ]
"<opposition-led>"
      "opposition-led" { 11SG pinzani } A [ -li-ongozwa na > ]
"<investment-related>"
      "investment-related" { 11SG wekezaji } A [ -li- na uhusiano
na > ]
"<funnel-shaped>"
      "funnel-shaped" { 9SG faneli } A [ -enye muundo wa > ]
"<state-sponsored>"
      "state-sponsored" { 9SG serikali } PREFR A [ -li-fadhilishwa
na > ]
"<nation-wide>"
      "nation-wide" { 5SG taifa } A [ -enye upana wa > ]
"<knee-deep>"
      "knee-deep" { 5SG goti } A [ -a kina cha > ]
```

The gloss of the first member of the compound is moved to the appropriate place (22).

(22)

```
"<hill-based>"
      "hill-based" { -li- na msingi katika 7SG lima } A
"<smoke-causing>"
      "smoke-causing" { -na-sababisha 3SG oshi } A
"<life-changing>"
      "life-changing" { -na-badilisha 6PLSG isha } A
"<state-controlled>"
      "state-controlled" { -li-dhibitiwa na 9SG serikali } PREFR A
"<male-dominated>"
      "male-dominated" { -li-tawaliwa na 1SG anaume } HUM A
"<president-elect>"
      "president-elect { 9SG rais aliyechaguliwa } HUM" N SG
      "president-elect { -li-chaguliwa < 9SG rais aliyechaguliwa }
HUM" A
"<sex-fuelled>"
      "sex-fuelled" { -li-endeshwa na 9SG jinsia } A
"<video-inspired>"
      "video-inspired" { -li-sajiishwa na 9SG video } A
"<opposition-led>"
      "opposition-led" { -li-ongozwa na 11SG pinzani } A
"<investment-related>"
      "investment-related" { -li- na uhusiano na 11SG wekezaji } A
"<funnel-shaped>"
      "funnel-shaped" { -enye muundo wa 9SG faneli } A
"<state-sponsored>"
      "state-sponsored" { -li-fadhilishwa na 9SG serikali } PREFR
A
```

```
"<nation-wide>"
      "nation-wide" { -enye upana wa 5SG taifa } A
"<knee-deep>"
      "knee-deep" { -a kina cha 5SG goti } A
```

The noun class code is converted to surface form and joined to the noun stem (23).

(23)
```
"<hill-based>"
      "hill-based" { -li- na msingi katika kilima } A
"<smoke-causing>"
      "smoke-causing" { -na-sababisha moshi } A
"<life-changing>"
      "life-changing" { -na-badilisha maisha } A
"<state-controlled>"
      "state-controlled" { -li-dhibitiwa na serikali } PREFR A
"<male-dominated>"
      "male-dominated" { -li-tawaliwa na mwanaume } HUM A
"<president-elect>"
      "president-elect { rais aliyechaguliwa } HUM" N SG
      "president-elect { -li-chaguliwa < rais aliyechaguliwa }
HUM" A
"<sex-fuelled>"
      "sex-fuelled" { -li-endeshwa na jinsia } A
"<video-inspired>"
      "video-inspired" { -li-sajiishwa na video } A
"<opposition-led>"
      "opposition-led" { -li-ongozwa na upinzani } A
"<investment-related>"
      "investment-related" { -li- na uhusiano na uwekezaji } A
"<funnel-shaped>"
      "funnel-shaped" { -enye muundo wa faneli } A
"<state-sponsored>"
      "state-sponsored" { -li-fadhilishwa na serikali } PREFR A
"<nation-wide>"
      "nation-wide" { -enye upana wa taifa } A
"<knee-deep>"
      "knee-deep" { -a kina cha goti } A
```

Now each compound has only one gloss with adjectival function, and they can now be put into context and translated. We take a few examples (24).

(24)
```
"<investment-related>"
      "investment { 11SG wekezaji , 7SG 8PL tegauchumi }-related"
A [ -li- na uhusiano na > ]
"<society>"
      "society { 9SG 10PL jamii }" N SG
"<.>"
      "." **CLB
```

```
"<video-inspired>"
      "video { 9SG 10PL video }-inspired" A [ -li-sajiishwa na > ]
A-REL
      "video { 9SG 10PL video }-inspired" A  [ -li-sajiishwa na >
] A-REL
      "video { 9SG 10PL video }-inspired" N
"<program>"
      "program { tengenezA programu } HUM-V SVO" V vt INF
      "program { tengenezA programu } HUM-V SVO" V vt IMP
      "program { tengenezA programu } HUM-V SVO" V vt PRES SG1
      "program { tengenezA programu } HUM-V SVO" V vt PRES SG2/PL2
      "program { tengenezA programu } HUM-V SVO" V vt PRES PL1
      "program { tengenezA programu } HUM-V SVO" V vt PRES PL3
      "program { 9SG 10PL programu }" N SG
"<.>"
      "." **CLB
"<sex-fuelled>"
      "sex { 9SG 10PL jinsia , 11SG jinsia }-fuelled" A [ -li-
endeshwa na > ] A-REL
      "sex-fuelled" A
"<program>"
      "program { tengenezA programu } HUM-V SVO" V vt INF
      "program { tengenezA programu } HUM-V SVO" V vt IMP
      "program { tengenezA programu } HUM-V SVO" V vt PRES SG1
      "program { tengenezA programu } HUM-V SVO" V vt PRES SG2/PL2
      "program { tengenezA programu } HUM-V SVO" V vt PRES PL1
      "program { tengenezA programu } HUM-V SVO" V vt PRES PL3
      "program { 9SG 10PL programu }" N SG
"<.>"
      "." **CLB
"<state-controlled>"
      "state { 9SG 10PL serikali , 9SG 10PL hali , 9SG 10PL dola
}-controlled" PREFR A [ -li-dhibitiwa na > ] A-REL
"<program>"
      "program { tengenezA programu } HUM-V SVO" V vt INF
      "program { tengenezA programu } HUM-V SVO" V vt IMP
      "program { tengenezA programu } HUM-V SVO" V vt PRES SG1
      "program { tengenezA programu } HUM-V SVO" V vt PRES SG2/PL2
      "program { tengenezA programu } HUM-V SVO" V vt PRES PL1
      "program { tengenezA programu } HUM-V SVO" V vt PRES PL3
      "program { 9SG 10PL programu }" N SG
"<.>"
      "." **CLB
"<horse-faced>"
      "horse-faced" { 9SG 10PL farasi } AN A [ -enye uso wa > , -
enye uso >11 ]
"<statue>"
      "statue { 9SG 10PL sanamu , 5SG 6PL sanamu }" N SG
"<.>"
      "." **CLB
```

We see that after analysis and insertion of glosses there is a lot of ambiguity. Part of ambiguity is on the same line, and in other words the ambiguity is cascaded. This is due to different processes. Handling compounds goes through a different route compared with ordinary words. We resolve both types of ambiguity (25).

(25)

```
"<investment-related>"
      "investment-related" { -li- na uhusiano na u+wekezaji } A
"<society>"
      "society { 9SG jamii }" N SG
"<.>"
      "." **CLB
"<video-inspired>"
      "video-inspired" { -li-sajiishwa na video } A  A-REL
"<program>"
      "program { 9SG programu }" N SG
"<.>"
      "." **CLB
"<sex-fuelled>"
      "sex-fuelled" { -li-endeshwa na jinsia } A  A-REL
"<program>"
      "program { 9SG programu }" N SG
"<.>"
      "." **CLB
"<state-controlled>"
      "state-controlled" { -li-dhibitiwa na serikali } PREFR A  A-
REL
"<program>"
      "program { 9SG programu }" N SG
"<horse-faced>"
      "horse-faced" { -enye uso wa > 9SG farasi } AN A
"<statue>"
      "statue { 9SG sanamu }" N SG
"<.>"
      "." **CLB
```

We add inflection codes to compounds, using the following noun as a key (26).

(26)

```
"<investment-related>"
      "investment-related" { -li- na uhusiano na u+wekezaji } A
9SG
"<society>"
      "society { 9SG jamii }" N SG
"<.>"
      "." **CLB
"<video-inspired>"
      "video-inspired" { -li-sajiishwa na video } A  A-REL 9SG
"<program>"
      "program { 9SG programu }" N SG
```

```
"<.>"
      "." **CLB
"<sex-fuelled>"
      "sex-fuelled" { -li-endeshwa na jinsia } A  A-REL 9SG
"<program>"
      "program { 9SG programu }" N SG
"<.>"
      "." **CLB
"<state-controlled>"
      "state-controlled" { -li-dhibitiwa na serikali } PREFR A  A-
REL 9SG
"<program>"
      "program { 9SG programu }" N SG
"<.>"
      "." **CLB
"<horse-faced>"
      "horse-faced" { -enye uso wa farasi } AN A 9SG
"<statue>"
      "statue { 9SG sanamu }" N SG
"<.>"
      "." **CLB
```

With the help of the inflection code we convert the dashes into surface forms (27).

(27)

```
"<investment-related>"
      "investment-related" { iliyo na uhusiano na u+wekezaji } A
9SG
"<society>"
      "society { 9SG jamii }" N SG
"<.>"
      "." **CLB
"<video-inspired>"
      "video-inspired" { iliyosajiishwa na video } A  A-REL 9SG
"<program>"
      "program { 9SG programu }" N SG
"<.>"
      "." **CLB
"<sex-fuelled>"
      "sex-fuelled" { iliyoendeshwa na jinsia } A  A-REL 9SG
"<program>"
      "program { 9SG programu }" N SG
"<.>"
      "." **CLB
"<state-controlled>"
      "state-controlled" { iliyodhibitiwa na serikali } PREFR A
A-REL 9SG
"<program>"
      "program { 9SG programu }" N SG
"<.>"
      "." **CLB
```

```
"<horse-faced>"
      "horse-faced" { yenye uso wa farasi } AN A 9SG
"<statue>"
      "statue { 9SG sanamu }" N SG
"<.>"
      "." **CLB
```

Now we can get the final translation (28).

(28)
jamii iliyo na uhusiano na uwekezaji.
programu iliyosajiishwa na video.
programu iliyoendeshwa na jinsia.
programu iliyodhibitiwa na serikali.
sanamu yenye uso wa farasi.

4 Conclusion

The above discussion and demonstration show that it is possible to translate also such *ad hoc* compounds, which are not included into the morphological lexicon. The conventional method of using heuristic guessing as additional aid is too unreliable and defective for handling these compounds. I have shown that by splitting the compound into units it is possible to handle each component separately, and the final result of the processing the compound is a single lexical entry coupled with linguistic tags. This result can then be further directed to the normal translation routines. Although the approach here is the same as in translating from English to Finnish (Report No. 57), the production of the translation of the compounds must follow language-specific rules.