Technical Reports on Language Technology
Report No 71, 2021
http://www.njas.helsinki.fi/salama

Translating unknown compounds from English to Finnish[1]

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

Abstract

The translation of unknown compounds is one of the most challenging tasks in machine translation (MT). If the members of the compound are written as separate words, there is only one method for handling them. In analysis, each member is analysed as a word of its own, and in a later phase they will be isolated as multiword expressions (MWE) and treated as single units. If the members are written together as a single word, either directly or using a dash '-' in between, we encounter problems already in morphological analysis. By default, a word boundary stops the analysis, and only continuous strings can be analysed. Part of such compounds can be analysed so that the lexicon structure allows such combinations of words. However, for this method to succeed, all the members of the compound must be as individual words in the lexicon.

In this report, I will discuss such cases, where all members of the compound are not in the lexicon, and they also must be handled in MT.

Key Words: *noun compound, language analysis, machine translation.*

1 Introduction

Earlier I wrote two reports on the so-called *ad hoc* compounds and their translation[2]. The assumption was that all the members of the compound were listed in the lexicon. The translation process was applied to English to Finnish and English to Swahili translation. The solutions were very different, because the structures of those languages are different. It was assumed that the compounds can be described as a union of individual words in the lexicon.

In this report we will discuss compounds, where at least one member is unknown, that is, it is not in the lexicon. Because the analysis system handles only continuous strings, it cannot handle strings with defective description in the lexicon.

[1] The report is issued under licence CC BY-NC

[2] http://www.njas.helsinki.fi/salama/ad-hoc-compounds-in-en-to-fi-mt.pdf
http://www.njas.helsinki.fi/salama/ad-hoc-compounds-in-en-to-swa-mt.pdf

The solution is that we split the compound into parts and inspect each part separately. In case there is a part that the lexicon does not recognise, we handle that part with the heuristic guesser, and the known part or parts will be analysed as individual words. As a result, we have a member with heuristic interpretation, while the other member or members are properly analysed.

Very often the unknown members are proper names, for example place names and people's names. These are categories, for which one cannot hope to maintain an up-to-date list.

In English, the unknown member of the compound is usually the first member, while to the latter part we can give an interpretation. The known member often gives a hint about the type of the unknown member. For example, it may suggest that the unknown member must be a person. Or it suggests that the unknown member must be a place, either a proper name or an ordinary word.

In the discussion below, I use the news corpus of about 10,000 sentences from the years 2017-2019. This data was the translation task in the WMT competition in those years. This data was chosen, because it was manually edited and considered nearly correct. Therefore, only those compounds are discussed, which occur in the corpus.

2 Pre-process the problematic compounds

In order to analyse such compounds, where one or more segments are unknown, we detach the components of the compound as separate words. We assume that the components are joined with a dash '-' in text. This is the normal method in all compounds in English, unless the components are written as separate words. It is not economical to treat all dash-joined (note! this might be a new compound) compounds in this way. Those compounds, which can be described in the lexicon, should be handled there.

How can we know, which compound types are problematic? We can only do this by considering each compound type separately. For example, the compound type 'five-day' is likely to have a number or numeral as its first part. This is a closed set, and it can be described in the lexicon. It is unlikely that a proper name would be the first part in this compound type. Therefore, we can process such compounds as single strings in the analyser.

3 Dividing the compounds into two groups

We can divide the compounds into two groups, (a) those that can be described safely in the morphological lexicon, and (b) those that can only partly be described in the morphological lexicon. With the latter group I mean such compounds, which may have an adjective, numeral, or noun as the left part of the compound, and alternatively a proper name as the left part.

The first group does not cause major problems and the translation of such compounds was described in earlier reports. We concentrate here on the latter group.

For the sake of clarity, I list the two compound types below (1 and 2). The lists were extracted from that version of the English morphological lexicon, which is applied to English to Finnish translation. Therefore, the glosses are in Finnish.

(1)

```
LEXICON AdjExt
-baked # "= [ leivottu N1-C ] ";
-catching # "= <ILL [ pistävä N10 FRONT ]";
-causing # "= <PAR [ aiheuttava N10 ]";
-century # "= [ vuosisata N9-F ]";
-changing # "= <PAR [ muuttava N10 ]";
-clad # "= <NOM [ pukuinen N38 , päällysteinen N38 FRONT ]";
-class # "= <NOM [ luokkainen N38 ]";
-deep # "= <GEN [ syvyinen N38 FRONT ]";
-degree # "= <GEN [ asteen ]";
-elect # "= [ valittu N1-C ]";
-faced # "= <NOM [ kasvoinen N10 ]";
-filled # "= <GEN [ täyttämä N10 FRONT ]";
-free # "= <NOM [ vapaa N17 ]";
-game # "= <GEN [ pelinen N38 FRONT ]";
-handed # "= <NOM [ kätinen N38 FRONT ]";
-held # "= <GEN [ pitämä N10 FRONT ]";
-inspired # "= <GEN [ innoittama N10 ]";
-level # "= <NOM [ tasoinen N38 ]";
-like # "=  <GEN [ tapainen N38 ]";
-lived # "= <NOM [ ikäinen N38 ]";
-living # "= <NOM [ ikäinen N38 ]";
-looking # "= <GEN [ näköinen N38 FRONT ]";
-man # "= <NOM [ miehinen N38 FRONT ]";
-needed # "= [ tarvittu N1-C ]";
-year-old # "= [ vuotta vanha N9 ]";
-old # "= <PAR [ vanha N9 ]";
-olds # "= <PAR [ vanha N9 ] PL";
-page # "= <NOM [ sivuinen N38 ]";
-paid # "= [ maksettu N1-C ]";
-person # "= [ henkinen N38 FRONT ]";
-point # "= [ kohtainen N38 ]";
-ranked # "= <ALL [ arvioitu N1-F ]";
-ranking # "= <ALL [ arvioiva N10 ]";
-registered # "= [ rekisteröity N1-F FRONT ]";
-rich # "= <NOM [ rikas N41-A ]";
-round # "= [ kierroksinen N38 ]";
-scale # "= [ mittainen N38 ]";
-shaped # "= <GEN [ muotoinen N38 ]";
-sided # "= [ sivuinen N38 ]";
-sized # "= <GEN [ kokoinen N38 ]";
-sourced # "= <NOM [ lähtöinen N38 FRONT ]";
-stricken # "= <GEN [ lyömä N10 FRONT ]";
-strong # "= <GEN [ vahvuinen N38 ]";
-style # "= <NOM [ tyylinen N38 FRONT ]";
-tall # "= <GEN [ korkuinen N38 ]";
-term # "= [ ajan ]";
-thick # "= <GEN [ paksuinen N38 ]";
-time # "= [ kertainen N38 , aikainen N38 ]";
-traded # "= [ ostettu N1-C ]";
```

```
-winning # "= <PAR [ voittava N10 ]";
-won # "= [ voitettu N1-C ]";
-working # "= [ työskentelevä N10 FRONT ]";
```

(2)

```
LEXICON AdjExt2
--backed # "= <GEN [ tukema N10 ]";
--based # "= [ kotoisin oleva N10 <ELA , -pohjainen N38 <NOM ,
perustuva N10 <ILL ]";
--controlled # "= <GEN [ kontrolloima N10 ]";
--dominated # "= <GEN [ hallitsema N10 ]";
--fuelled # "= <NOM [ -käyttöinen N38 FRONT ]";
--funded # "= <GEN [ rahoittama N10 ]";
--led # "= <NOM [ -johtoinen N38 ]";
--made # "= [ -tekoinen N38 <NOM , tehty N1-F FRONT <INE ]";
--minded # "= <NOM [ -mielinen N38 ]";
--operated # "= <NOM [ -käyttöinen N38 FRONT ]";
--owned # "= <NOM [ -omisteinen N38 ]";
--related # "= <ILL [ liittyvä N10 FRONT ]";
--sponsored # "= <GEN [ sponsoroima N10 ]";
--wide # "= <GEN [ -laajuinen N38 ]";
```

List (1) contains all nonproblematic dash-connected compound types found in the corpus. List (2) contains only the problematic compound types.

4 Solution 1: Known and unknown left parts of List (2) compounds treated in different ways

Below I will test two methods of handling List (2) compounds. In the first method we split as separate words such compounds that likely have a proper name as left part.

Let us assume that in list (2), each word may have known members and unknown members as left part. We separate the compound members from each other in the pre-processing phase, if the left member is likely a proper name.

We do not pre-process, that is, separate from each other, such cases, where the left part is not a proper name. If it is not a proper name, it will likely be in the lexicon, and its interpretation is given in the analysis.

Our strategy is: if the left member is capital-initial, it is likely a proper name, and the pre-processing rules apply to such cases. If it is not capital-initial, it is likely listed in the dictionary, and the pre-processing rules do not apply to such cases.

The pre-processing can be implemented in many ways. In (3) is an example of a Perl script.

(3)

```
perl -pe 's/ ([A-ZÅÄÖ]\S+)\-(based)/ $1 --$2/gm'
```

The rule says: If there is a dash in the environment, where immediately on the left there is anything except the word boundary, and before the word boundary there is a capital letter, and immediately on the right there is the string 'based', then copy the left part, put an empty space, and then put two dashes and copy the right part.

In (4) we have examples of how the rule works with compounds that have -based as their last part. The compounds are in original form in (4) and after pre-processing in (5).

(4)
Mombasa-based
Machakos-based
foreign-based
Foreign-based
meat-based
Notting Hill-based

(5)
Mombasa --based
Machakos --based
foreign-based
Foreign --based
meat-based
Notting Hill --based

We see that if the first member of the compound is capital-initial, the rule applies to it. The compounds that are normally lower-case initial are problematic, because in sentence-initial position thy are capital-initial. Therefore, the rule over-generates in certain contexts. The damage is minimal, however, because the structure can still be handled as separate units.

In (6) we see how the compounds are analysed after pre-processing.

(6)
```
(a) "<Mombasa>"
      "mombasa" PROPN

"<--based>"
      "--based" A [ kotoisin oleva N10 <ELA , -pohjainen N38 <NOM
, perustuva N10 <ILL ]

(b) "<Machakos>"
      "machakos" PROPN CAP Heur
"<--based>"
      "--based" A [ kotoisin oleva N10 <ELA , -pohjainen N38 <NOM
, perustuva N10 <ILL ]

(c) "<foreign-based>"
      "foreign-based" A [ kotoisin oleva N10 <ELA , -pohjainen N38
<NOM , perustuva N10 <ILL ]

(d) "<Foreign>"
      "Foreign" CAP A
```

```
"<--based>"
      "--based" A [ kotoisin oleva N10 <ELA , -pohjainen N38 <NOM
, perustuva N10 <ILL ]

(e) "<meat-based>"
      "meat-based" A  [ kotoisin oleva N10 <ELA , -pohjainen N38
<NOM , perustuva N10 <ILL ]

(f) "<Notting>"
      "notting" PROPN CAP Heur
"<Hill>"
      "Hill" CAP N SG
"<--based>"
      "--based" A [ kotoisin oleva N10 <ELA , -pohjainen N38 <NOM
, perustuva N10 <ILL ]
```

Cases (a) and (b) above have a word meaning place as the left member, but the analysis is different. *Mombasa* is analysed as proper name, but *Machakos* is interpreted as proper name by means of guessing. It is not in the lexicon. The cases (c) and (d) are the same compound, but the latter one is sentence-initial. Therefore, they were analysed in different ways. The analysis of case (e) comes directly from the lexicon.

Case (f) has two members on the left side, but they are written as separate words in original text. In English, such strange writing seems possible. In Finnish, if the left part contains more than one separate word, the right part is written as a separate word (for example, *Notting Hill -based*). If this would be the practice also in English, it would be easy to handle such cases in translation. Now it is impossible to know whether *Notting* belongs to the compound or not.

We see above that only the right part '-based' has a Finnish gloss, in fact three alternative glosses. The left part is without gloss. This derives from the structure of the lexicon. In it, the entries are normally without glosses. Only some closed sets, such as the ones we discuss here, have glosses in the lexicon. The glosses for other members will be added later in the translation process.

Next we add the glosses of the left part to the compounds, which were not detached in the pre-processing phase (7).

(7)

```
(a) "<Mombasa>"
      "mombasa" PROPN
"<--based>"
      "--based" A [ kotoisin oleva N10 <ELA , -pohjainen N38 <NOM
, perustuva N10 <ILL ]

(b) "<Machakos>"
      "machakos" PROPN CAP Heur
"<--based>"
      "--based" A [ kotoisin oleva N10 <ELA , -pohjainen N38 <NOM
, perustuva N10 <ILL ]

(c) "<foreign-based>"
```

```
      "foreign { ulkomainen NEN N38 , ulko-- COMP , vieras N41 ,
ulkomaalainen NEN N38 }-based" A [ kotoisin oleva N10 <ELA , -
pohjainen N38 <NOM , perustuva N10 <ILL ]

(d) "<Foreign>"
      "Foreign" CAP A
"<--based>"
      "--based" A [ kotoisin oleva N10 <ELA , -pohjainen N38 <NOM
, perustuva N10 <ILL ]

(e) "<meat-based>"
      "meat { liha N9 , liha-- COMP }-based" A [ kotoisin oleva
N10 <ELA , -pohjainen N38 <NOM , perustuva N10 <ILL ]
      "meat { liha N9 , liha-- COMP }-based" N

(f) "<Notting>"
      "notting" PROPN CAP Heur
"<Hill>"
      "Hill" CAP N SG
"<--based>"
      "--based" A [ kotoisin oleva N10 <ELA , -pohjainen N38 <NOM
, perustuva N10 <ILL ]
```

In this phase we added the glosses of those words, which were in the lexicon. Such glosses are surrounded with curly brackets *{* and *}*, whereas the right part glosses, surrounded with *[* and *]*, are listed already in the lexicon. These two types of bracketing help in constructing the final forms.

The separated left parts are still without glosses. We do this in the next phase (8).

(8)

```
(a) "<Mombasa>"
      "mombasa" { Mombasa } PROPN
"<--based>"
      "--based" A [ kotoisin oleva N10 <ELA , -pohjainen N38 <NOM
, perustuva N10 <ILL ]

(b) "<Machakos>"
      "machakos" { Machakos } PROPN CAP Heur
"<--based>"
      "--based" A [ kotoisin oleva N10 <ELA , -pohjainen N38 <NOM
, perustuva N10 <ILL ]

(c) "<foreign-based>"
      "foreign" { ulkomainen NEN N38 , vieras-- COMP , vieras N41
, ulkomaalainen NEN N38 }-based A [ kotoisin oleva N10 <ELA , -
pohjainen N38 <NOM , perustuva N10 <ILL ]

(d) "<Foreign>"
      "Foreign" { ulkomainen NEN N38 , vieras-- COMP , vieras N41
, ulkomaalainen NEN N38 } CAP A
```

```
"<--based>"
      "--based" A [ kotoisin oleva N10 <ELA , -pohjainen N38 <NOM
, perustuva N10 <ILL ]

(e) "<meat-based>"
      "meat" { liha N9 , liha-- COMP }-based A [ kotoisin oleva
N10 <ELA , -pohjainen N38 <NOM , perustuva N10 <ILL ]
      "meat" { liha N9 , liha-- COMP }-based N

(f) "<Notting>"
      "notting" { Notting } PROPN CAP Heur
"<Hill>"
      "Hill" { Hill } CAP N SG
"<--based>"
      "--based" A [ kotoisin oleva N10 <ELA , -pohjainen N38 <NOM
, perustuva N10 <ILL ]
```

The original reading was copied as a gloss, if it had a tag *Heur* (that is, the interpretation was done by heuristic means), such as *Machakos*, or if it was listed in lexicon, as *Mombasa*.

Now there are glosses surrounded with *[* and *]*, and glosses surrounded with *{* and *}*. Readings of both gloss types must be cascaded, so that we can perform disambiguation (9).

(9)

```
(a) "<Mombasa>"
      "mombasa" { Mombasa } PROPN
"<--based>"
      "--based" A [ kotoisin oleva N10 <ELA ]
      "--based" A [ -pohjainen N38 <NOM ]
      "--based" A [ perustuva N10 <ILL ]

(b) "<Machakos>"
      "machakos" { Machakos } PROPN CAP Heur
"<--based>"
      "--based" A [ kotoisin oleva N10 <ELA ]
      "--based" A [ -pohjainen N38 <NOM ]
      "--based" A [ perustuva N10 <ILL ]

(c) "<foreign-based>"
      "foreign" { ulkomainen NEN N38 } -based A [ kotoisin oleva
N10 <ELA ]
      "-based A [ -pohjainen N38 <NOM ]
      "-based A [ perustuva N10 <ILL ]
      "foreign" { vieras-- COMP , vieras N41 , ulkomaalainen NEN
N38 } -based A [ kotoisin oleva N10 <ELA ]
      "-based A [ -pohjainen N38 <NOM ]
      "-based A [ perustuva N10 <ILL ]
```

```
(d) "<Foreign>"
      "Foreign" { ulkomainen NEN N38 } CAP A
      "Foreign" { vieras-- COMP } CAP A
      "Foreign" { vieras N41 } CAP A
      "Foreign" { ulkomaalainen NEN N38 } CAP A
"<--based>"
      "--based" A [ kotoisin oleva N10 <ELA ]
      "--based" A [ -pohjainen N38 <NOM ]
      "--based" A [ perustuva N10 <ILL ]

(e) "<meat-based>"
      "meat" { liha N9 } -based A [ kotoisin oleva N10 <ELA ]
      "-based A [ -pohjainen N38 <NOM ]
      "-based A [ perustuva N10 <ILL ]
      "meat" { liha-- COMP } -based A [ kotoisin oleva N10 <ELA ]
      "-based A [ -pohjainen N38 <NOM ]
      "-based A [ perustuva N10 <ILL ]
      "meat" { liha N9 } -based N
      "meat" { liha-- COMP } -based N

(f) "<Notting>"
      "notting" { Notting } PROPN CAP Heur
"<Hill>"
      "Hill" { Hill } CAP N SG
"<--based>"
      "--based" A [ kotoisin oleva N10 <ELA ]
      "--based" A [ -pohjainen N38 <NOM ]
      "--based" A [ perustuva N10 <ILL ]
```

When we look at the result above, we see that some of the readings are not in correct format for disambiguation. In fact, only those readings, which were detached before analysis, are in correct format.

We will see, what happens, if we detach all those compounds, which have the string *based* as right member (10).

(10)

```
(a) "<Mombasa>"
      "mombasa" { Mombasa } PROPN
"<--based>"
      "--based" A [ kotoisin oleva N10 <ELA ]
      "--based" A [ -pohjainen N38 <NOM ]
      "--based" A [ perustuva N10 <ILL ]

(b) "<Machakos>"
      "machakos" { Machakos } PROPN CAP Heur
"<--based>"
      "--based" A [ kotoisin oleva N10 <ELA ]
      "--based" A [ -pohjainen N38 <NOM ]
      "--based" A [ perustuva N10 <ILL ]
```

```
(c) "<foreign>"
      "foreign" { ulkomainen NEN N38 } A
      "foreign" { vieras-- COMP } A
      "foreign" { vieras N41 } A
      "foreign" { ulkomaalainen NEN N38 } A
 "<--based>"
      "--based" A [ kotoisin oleva N10 <ELA ]
      "--based" A [ -pohjainen N38 <NOM ]
      "--based" A [ perustuva N10 <ILL ]

(d) "<Foreign>"
      "Foreign" { ulkomainen NEN N38 } CAP A
      "Foreign" { vieras-- COMP } CAP A
      "Foreign" { vieras N41 } CAP A
      "Foreign" { ulkomaalainen NEN N38 } CAP A
"<--based>"
      "--based" A [ kotoisin oleva N10 <ELA ]
      "--based" A [ -pohjainen N38 <NOM ]
      "--based" A [ perustuva N10 <ILL ]

(e) "<meat>"
      "meat" { liha N9 } N SG
      "meat" { liha-- COMP } N SG
"<--based>"
      "--based" A [ kotoisin oleva N10 <ELA ]
      "--based" A [ -pohjainen N38 <NOM ]
      "--based" A [ perustuva N10 <ILL ]

(f) "<Notting>"
      "notting" { Notting } PROPN CAP Heur
"<Hill>"
      "Hill" { Hill } CAP N SG
"<--based>"
      "--based" A [ kotoisin oleva N10 <ELA ]
      "--based" A [ -pohjainen N38 <NOM ]
      "--based" A [ perustuva N10 <ILL ]
```

5 Solution2: Treat all cases in List (2) in the same way

Now all readings in (10) are in correct format for disambiguation. We can conclude that it might be better to detach all compounds that have as right part any of the strings in (2). The solution makes the semantic disambiguation easier.

What are the criteria for disambiguating the above readings? We note that there are obvious combination pairs on both sides. The tag COMB on the left side indicates that it should be combined with some string on the right side. The dash '-' in the beginning of the gloss on the right side shows that it is the correct string to be combined.

If there is no COMB tag on the left side, we must look for criteria for disambiguation. There are two alternatives left, *kotoisin oleva* and *perustuva*. If the proper name is a human being, the latter choice is correct. If the proper name is a place, the first alternative is correct. But how can we know whether it is a human being or place? The answer is: we cannot know for sure. However, we can resort to likelihood. If there is the word *-based* as the right member, it is almost sure that the left side proper name is a place. Therefore, this interpretation is the default, and it will be chosen if no rule applies. The result after disambiguation is in (11).

(11)

```
"<Mombasa>"
	"mombasa" { Mombasa } PROPN
"<--based>"
	"--based" A [ kotoisin oleva N10 <ELA ]
"<Machakos>"
	"machakos" { Machakos } PROPN CAP Heur
"<--based>"
	"--based" A [ kotoisin oleva N10 <ELA ]
"<foreign>"
	"foreign" { vieras-- COMP } A
"<--based>"
	"--based" A [ -pohjainen N38 <NOM ]
"<Foreign>"
	"Foreign" { Vieras-- COMP } CAP A
"<--based>"
	"--based" A [ -pohjainen N38 <NOM ]
"<meat>"
	"meat" { liha-- COMP } N SG
"<--based>"
	"--based" A [ -pohjainen N38 <NOM ]
"<Notting>"
	"notting" { Notting } PROPN CAP Heur
"<Hill>"
	"Hill" CAP N SG
"<--based>"
	"--based" A [ kotoisin oleva N10 <ELA ]
```

Now all readings have been disambiguated and we can proceed with processing. We reformat the readings, so that each compound is on a single line (12).

(12)

```
"<Mombasa>"
	"mombasa" { Mombasa } PROPN "<--based>" "--based" A [
kotoisin oleva N10 <ELA ]
 "<Machakos>"
	"machakos" { Machakos } PROPN CAP Heur "<--based>" "--based"
A [ kotoisin oleva N10 <ELA ]
 "<foreign>"
	"foreign" { vieras-- COMP } A "<--based>" "--based" A [ -
pohjainen N38 <NOM ]
 "<Foreign>"
	"Foreign" { Vieras-- COMP } CAP A "<--based>" "--based" A [
-pohjainen N38 <NOM ]
 "<meat>"
```

```
      "meat" { liha-- COMP } N SG "<--based>" "--based" A [ -
pohjainen N38 <NOM ]
 "<Notting Hill>"
      <MW "Notting Hill" { Notting Hill kotoisin oleva } CAP ELA N
"<--based>" "--based" A [ N10 ] SG NOM
```

Note that the words *Notting* and *Hill* were isolated as a multiword expression. We join the cases, where the members can be directly joined (13).

```
(13)
"<Mombasa>"
      "mombasa" { Mombasa } PROPN "<--based>" "--based" A [
kotoisin oleva N10 <ELA ]
 "<Machakos>"
      "machakos" { Machakos } PROPN CAP Heur "<--based>" "--based"
A [ kotoisin oleva N10 <ELA ]
 "<foreign>"
      "foreign" { vieraspohjainen N38 <NOM ]
 "<Foreign>"
      "Foreign" { Vieraspohjainen N38 <NOM ]
 "<meat>"
      "meat" { lihapohjainen N38 <NOM ]
 "<Notting Hill>"
      <MW "Notting Hill" { Notting Hill kotoisin oleva } CAP ELA N
"<--based>" "--based" A [ N10 ] SG NOM
```

We still need to process the cases, where the production of the form is more complex, also involving inflection. This is the case, where the left member is a proper name. We first add the inflection class to the proper names (14). Note that the name *Hill* requires front inflection in Finnish.

```
(14)
"<Mombasa>"
      "mombasa" { Mombasa N9 } PROPN "<--based>" "--based" A [
kotoisin oleva N10 <ELA ]
 "<Machakos>"
      "machakos" { Machakos N39 } PROPN CAP Heur "<--based>" "--
based" A [ kotoisin oleva N10 <ELA ]
 "<foreign>"
      "foreign" { vieraspohjainen N38 <NOM ]
 "<Foreign>"
      "Foreign" { Vieraspohjainen N38 <NOM ]
 "<meat>"
      "meat" { lihapohjainen N38 <NOM ]
 "<Notting Hill>"
      <MW "Notting Hill" { Notting Hill N1b FRONT kotoisin oleva }
CAP ELA N  "<--based>" "--based" A [ N10 ] SG NOM
```

Then we add the inflection codes, which help in giving the words in gloss the correct form (15).

(15)

```
"<Mombasa>"
      "mombasa" { Mombasa N9 kotoisin oleva } PROPN "<--based>" "-
-based" A [ N10 <ELA ] SG NOM
"<Machakos>"
      "machakos" { Machakos N39 kotoisin oleva } PROPN CAP Heur
"<--based>" "--based" A [ N10 <ELA ] SG NOM
"<foreign>"
      "foreign" { vieraspohjainen N38 <NOM ]
"<Foreign>"
      "Foreign" { Vieraspohjainen N38 <NOM ]
"<meat>"
      "meat" { lihapohjainen N38 <NOM ]
 "<Notting Hill>"
      <MW "Notting Hill" { Notting Hill N1b FRONT kotoisin oleva }
CAP <ELA N  "<--based>" "--based" A [ N10 ] SG NOM
```

The code NOM means that the gloss must have the nominative form in this context. The proper names have the inflection class code (N9, N39 and N1b), but their inflection code <ELA is in the wrong place. In the next phase the code is moved to the left (16).

(16)

```
"<Mombasa>"
      "mombasa" { Mombasa N9 kotoisin oleva } PROPN <ELA "<--
based>" "--based" A [ N10 ] SG NOM
"<Machakos>"
      "machakos" { Machakos N39 kotoisin oleva } PROPN <ELA CAP
Heur "<--based>" "--based" A [ N10 ] SG NOM
"<foreign>"
      "foreign" { vieraspohjainen N38 <NOM ]
"<Foreign>"
      "Foreign" { Vieraspohjainen N38 <NOM ]
"<meat>"
      "meat" { lihapohjainen N38 <NOM ]
 "<Notting Hill>"
      <MW "Notting Hill" { Notting Hill N1b FRONT kotoisin oleva }
CAP <ELA N  "<--based>" "--based" A [ N10 ] SG NOM
```

The gloss in compounds with a proper noun on the left has two inflecting elements. In (16) above, the inflection code is immediately after the proper noun, but the code of the last part is far on the right. We move it to the correct place (17).

(17)

```
"<Mombasa>"
      "mombasa" { Mombasa N9 kotoisin oleva N10 } PROPN <ELA "<--
based>" "--based" A SG NOM
"<Machakos>"
      "machakos" { Machakos N39 kotoisin oleva N10 } PROPN <ELA
CAP Heur "<--based>" "--based" A SG NOM
```

```
"<foreign>"
      "foreign" { vieraspohjainen N38 <NOM ]
"<Foreign>"
      "Foreign" { Vieraspohjainen N38 <NOM ]
"<meat>"
      "meat" { lihapohjainen N38 <NOM ]
"<Notting Hill>"
      <MW "Notting Hill" { Notting Hill N1b FRONT kotoisin oleva
N10 } CAP <ELA N  "<--based>" "--based" A SG NOM
```

The glosses are converted to surface form (18).

(18)

```
"<Mombasa>"
      "mombasa" { Mombas:a+asta :N9 kotoisin olev:a :N10 } PROPN
<ELA "<--based>" "--based" A SG NOM
"<Machakos>"
      "machakos" { Machako:s+ksesta :N39 kotoisin olev:a :N10 }
PROPN <ELA CAP Heur "<--based>" "--based" A SG NOM
"<foreign>"
      "foreign" { vieraspohjai:nen :N38 <NOM ]
"<Foreign>"
      "Foreign" { Vieraspohjai:nen :N38 <NOM ]
"<meat>"
      "meat" { lihapohjai:nen :N38 <NOM ]
"<Notting>"
      <MW "Notting Hill" { Notting Hill:+istä :N1b FRONT kotoisin
olev:a :N10 } CAP <ELA N "<--based>" "--based" A SG NOM
```

Final pruning leaves only the translated text (19).

(19)
Mombasasta kotoisin oleva
Machakoksesta kotoisin oleva
vieraspohjainen
Vieraspohjainen
lihapohjainen
Notting Hillistä kotoisin oleva

6 Conclusion

We have seen that the translation of unknown compounds is a complex process. It is not problematic only because of defective information on the compound structure. The implementation of such structures must be integrated with the normal translation process, so that it does not cause harm to the whole process. In this report I have discussed the translation problem using a single right side member -*based*. I have investigated all its possible combinations and translated all of them simultaneously. All other cases in list (2) could possibly be handled in the same way, but this remains to be tested.