Proper names and acronyms in English to Finnish Machine Translation

Hurskainen, Arvi

SALAMA - Swahili Language Manager
2018

Hurskainen , A 2018 , Proper names and acronyms in English to Finnish Machine Translation . in SALAMA - Swahili Language Manager : Technical reports on LT . Technical reports on language technology , no. 28 , SALAMA - Swahili Language Manager , Helsinki .

http://hdl.handle.net/10138/311052

unspecified
publishedVersion

Technical Reports in Language Technology
Report No 28, 20018
http://www.njas.helsinki.fi/salama

Proper names and acronyms in English to Finnish Machine Translation

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

Abstract

The translation of proper names is a complex process, and it is hardly possible to constract an error free translation system. Therefore, the practical aim of translating proper names is to make it as correct as possible. The problems include the fact that many words or clusters of words, which are used as proper names, are also used as ordinary words. It is often very hard to define which interpretation is correct in each case. If this problem is resolved, we have another problem. Should the proper name be translated or left as such? Acronyms form also a translation problem. If a string is identified as an acronym, it is not self-evident that it should be translated as corresponding acronym or as a real word or cluster of words. And another problem is whether it should be translated at all. These problems will be discussed below.

Key Words: *machine translation, proper names, acronyms.*

1 Introduction

Analysis systems of text are normally constructed so that the analysis result contains only lowercase letters. This is done for simplifying and speeding up the analysis process. However, when we translate the text, we need to know precisely whether a word has capital letters and where in the word they are. This is done by adding suitable tags to the analysis result. Therefore, the information on correct writing style is retained through the translation process.

Problems occur when conventions do not match between languages, that is, a captal letter in source language does not match with a capital letter in target language. Therefore, along the translation process, we must also manipulate the tags, so that correct information is maintained when the translation process

2 Translating acronyms

An acronym in one language normally matches an acronym in another language, that is, an acronym is translated as a corresponding acronym in target language. Some acronyms are not translated at all. In the third type, an acronym of source language matches with an ordinary word in target language. Consider example (1).

(1)

```
1     The   the   det:>2      @DN> %>N DET
2     country     country     subj:>5     @SUBJ %NH N NOM SG
3     called      call  mod:>2      @-FMAINV %VP EN
4     USA   usa   obj:>3      @OBJ %NH N NOM SG
5     has   have  main:>0     @+FMAINV %VA V PRES SG3
6     its   it    attr:>7     @A> %>N PRON GEN SG3
7     president   president   obj:>5      @OBJ %NH N NOM SG
8     in    in    mod:>7      @<NOM %N< PREP
9     Washington  washington  pcomp:>8    @<P %NH N NOM SG
10    .     .
```

In (1), *USA* is an acronym and it is written with all capital. *Washington* is a proper name, and it is written with capital initial. The stem of both words is written with lowercase all. The sentence-initial *The* also is written with capital initial, when it starts the sentence. Otherwise it is written with lowercase all. In order to retain this information, we need to add tags for each three types of capitalisation (2).

(2)

```
"<The>"
      "the" %DN> CAPINIT DET
"<country>"
      "country" %SUBJ N SG NOM DEF
"<called>"
      "call" %-FMAINV V EN
"<USA>"
      "usa" %OBJ CAPALL N SG
"<has>"
      "have" %+FMAINV V PRES SG3
"<its>"
      "it" %A> PRON GEN SG3
"<president>"
      "president" %OBJ N SG DEF
"<in>"
      "in" %<NOM PREP
"<Washington>"
      "washington" %<P CAP N SG NOM
"<.>"
      "."
```

The corresponding tags are CAPINIT, CAPALL, and CAP. These tags retain the information on the type of capitalisation of the word. Note that in the translation process we are only dealing with the analysis result, not with the surface text. When the translation proceeds, we must be able to handle the question of capitalisation also in target text. In (2), to the words *country* and *president* was added the tag DEF indicating that they are here in defined form (e.g. the country), and not in indefined form, whereby the tag would be INDEF (e.g. a country). This is done, because in later phase the articles are removed as redundant material. After adding the Finnish equivalents of the words and after semantic disambiguation, we get a result as in (3).

(3)
```
"<country>"
      "country" { maa N18 } %SUBJ MAA DEF N NOM SG
"<called>"
      "call" { nimeltä :2 } %-FMAINV V EN
"<USA>"
      "usa" { usa N2 } %OBJ MAA IN NOPROP CAPALL N SG
"<has>"
      "have" { olla V67b } %+FMAINV S-ADE O-PAR O-ACC-N V PRES SG3
"<its>"
      "it" { se Np11 FRONT } %A> OUT PRON GEN SG3
"<president>"
      "president" { presidentti N5-C FRONT } %OBJ HUM TITLE NOPROP
DEF N SG
"<in>"
      "in" { NOGLOSS M-INE } %<NOM PREP
"<Washington>"
      "washington" { washington N1b } %<P PLACE IN CAP N NOM SG
"<.>"
      "." { . }
```

We note in (3) that the definite article *The* was removed, and as a result the first word of the sentence is *country* with lowercase initial. Although now the first word of the sentence does not have the tag CAPINIT, it can be handled correctly on the basis of its position in the sentence. The Finnish eqivalent for *USA* is *usa* plus the tag CAPALL. It shows that all the letters of the word should be in capital. The equivalent for *Washington* is *washington*, the word with lowercase initial plus the tag CAP. This shows that only the first letter of the word should be capitalised. On the basis of all this information we get the translation as in (4).

(4)
Maalla nimeltä USA on sen presidentti Washingtonissa.

When we need to inflect an acronym, we should put the inflection part into lowercase, if the text otherwise is in lowercase. Therefore, the word will have uppercase letters and lowercase letters. The example in (5) illustrates this.

(5)
```
"<The>"
      "the" %DN> CAPINIT DET
"<president>"
      "president" %SUBJ N SG NOM DEF
"<of>"
      "of" %<NOM-OF PREP
"<USA>"
      "usa" %<P CAPALL N SG NOM
"<lives>"
      "live" %+FMAINV V PRES SG3
```

```
"<in>"
      "in" %ADVL PREP
"<Washington>"
      "washington" %<P CAP N SG NOM
"<.>"
      "."
```

When we add Finnish glosses and disambiguate we get the result as in (6).

(6)
```
"<president>"
      "president" { presidentti N5-C FRONT } %SUBJ HUM TITLE
NOPROP DEF N NOM SG
"<of>"
      "of" { NOGLOSS M-GEN } %<NOM-OF PREP
"<USA>"
      "usa" { usa N2 } %<P MAA IN NOPROP CAPALL N NOM SG
"<lives>"
      "live" { elää V53 FRONT } %+FMAINV O-ADE V PRES SG3
"<in>"
      "in" { NOGLOSS M-INE } %ADVL PREP
"<Washington>"
      "washington" { washington N1b } %<P PLACE IN CAP N NOM SG
"<.>"
      "." { . }
```

Now we add the inflection suffixes to demonstrate how the stem of the word and its suffixes can be kept apart (7).

(7)
```
"<president>"
      "president" { presidentt:i :N5-C FRONT } %SUBJ HUM TITLE
NOPROP DEF N  SG NOM
"<of>"
      "of" { NOGLOSS M-GEN } %<NOM-OF PREP
"<USA>"
      "usa" { usa:+n :N2 } %<P MAA IN NOPROP CAPALL N  SG GEN
"<lives>"
      "live" { el:ää+aa :V53 FRONT } %+FMAINV O-ADE V PRES  SG
"<in>"
      "in" { NOGLOSS M-INE } %ADVL PREP
"<Washington>"
      "washington" { washington:+issa :N1b } %<P ACE IN CAP N  SG
INE
"<.>"
      "." { . }
```

We see in (7) that the inflected Finnish equivalent of *USA* is *usa:+n*. The colon ':' marks the end of the stem, and '+' marks the beginning of the suffix. What is between these two codes is the last part of the base form, and it is dropped in inflection. In the word for *USA*,

there is nothing after the stem, and the inflection part will be attached directly to the stem. Now when we convert the letters to uppercase, the boundary mark ':' indicates that the letters after that should not be converted. Therefore, we get the translation as in (8).

(8)
USA:n presidentti elää Washingtonissa.

This is the normal convention when acronyms are inflected. Optionally the colon ':' can be omitted.

There are also cases when acronyms do not match. That is, an acronym in English is translated with an ordinary word. By default, such words would be translated with capital all, and this is not sometimes desirable. Consider the example in (9).

(9)
```
"<This>"
      "this" %DN> CAPINIT DET DEM SG
"<year>"
      "year" %ADVL N SG NOM
"<,>"
      ","
"<the>"
      "the" %DN> DET
"<CPI>"
      "cpi" %SUBJ CAPALL ABBR NOM SG
"<has>"
      "have" %+FAUXV V PRES SG3
"<risen>"
      "rise" %-FMAINV V EN
"<.>"
      "."
```

Note that the acronym CPI has a tag CAPALL, because the rule defines so. After adding Finnish glosses and after disambiguation the result is as in (10).

(10)
```
"<This>"
      "this" { tämä Np1 FRONT } %DN> CAPINIT DET DEM SG
"<year>"
      "year" { vuosi N27 } %ADVL TIME N NOM SG
"<,>"
      "," { NOGLOSS }
"<CPI>"
      "cpi" { kuluttajahintaindeksi N5 FRONT } %SUBJ NOCAP NOPROP
ABBR NOM SG
"<has>"
      "have" { olla V67b } %+FAUXV HAVE-PERF V PRES SG3
"<risen>"
      "rise" { nousta V66 } %-FMAINV V EN
"<.>"
```

```
      "." { . }
```

The acronym *CPI* (consumer price index) is translated as *kuluttajahintaindeksi*, because Finnish does not use an acronym in this context. We also see that the tag CAPALL has disappeared, and it has been replaced with the tag NOCAP. When the trigger CAPALL for capitalising all words has disappeared, the words will be left to lowercase format (11).

(11)
Tänä vuotena kuluttajahintaindeksi on noussut.

There are also cases where both types of translation, the acronym and the real word, would be justified. Consider the case in (12).

(12)
```
"<This>"
      "this" %DN> CAPINIT DET DEM SG
"<year>"
      "year" %ADVL N SG NOM
"<,>"
      ","
"<the>"
      "the" %DN> DET
"<GNP>"
      "gnp" %SUBJ CAPALL ABBR NOM SG
"<has>"
      "have" %+FAUXV V PRES SG3
"<risen>"
      "rise" %-FMAINV V EN
"<.>"
      "."
```

When we add Finnish flosses, the result is as in (13).

(13)
```
"<This>"
      "this" { tämä Np1 FRONT } %DN> CAPINIT DET DEM SG
      "this" { nämä Np2 FRONT } %DN> CAPINIT DET DEM SG
      "this" { PROP-CAND } %DN> CAPINIT DET DEM SG
"<year>"
      "year" { vuosi N27 } TIME %ADVL N NOM SG
"<,>"
      "," { , }
      "," { NOGLOSS }
"<GNP>"
      "gnp" { bruttokansantulo N1 } NOCAP NOPROP %SUBJ ABBR NOM SG
      "gnp" { bruttokansantuote N48-C } NOCAP NOPROP %SUBJ ABBR
      NOM SG
      "gnp" { bkt N2 } NOCAP NOPROP %SUBJ ABBR NOM SG
      "gnp" { PROP-CAND } NOCAP NOPROP %SUBJ ABBR NOM SG
```

```
"<has>"
      "have" { olla V67b S-ADE O-PAR O-ACC-N } %+FAUXV V PRES SG3
      "have" { olla V67b HAVE-PERF } %+FAUXV V PRES SG3
      "have" { omistaa V67 } %+FAUXV V PRES SG3
      "have" { NOGLOSS } %+FAUXV V PRES SG3
      "have" { O-ACC-N } %+FAUXV V PRES SG3
      "have" { en ole } %+FAUXV V PRES SG3
      "have" { en ollut } %+FAUXV V PRES SG3
      "have" { et ole } %+FAUXV V PRES SG3
      "have" { ei ole O-PAR } %+FAUXV V PRES SG3
      "have" { ei ollut O-PAR } %+FAUXV V PRES SG3
      "have" { emme ole } %+FAUXV V PRES SG3
      "have" { ette ole } %+FAUXV V PRES SG3
      "have" { eivät ole :2 } %+FAUXV V PRES SG3
      "have" { eivät olleet :3 } %+FAUXV V PRES SG3
      "have" { saada V63 O-ACC } %+FAUXV V PRES SG3
"<risen>"
      "rise" { nousta V66 } %-FMAINV V EN
"<.>"
      "." { . }
```

We see that the acronym *GNP* has three grosses in Finnish, one of them an acronym *bkt*. If this would be a normal acronym with the code CAPALL, it would be converted into *BKT*, as it should be. Now, because of the non-acronym interpretation, the tag CAPALL is replaced with NOCAP, the conversion would not happen. Therefore, in cases where the acronym has an acronym interpretation and a non-acronym interpretation, the acronym interpretation should be glossed directly as acronym in uppercase, as in (14).

(14)

```
"<GNP>"
      "gnp" { bruttokansantulo N1 } NOCAP NOPROP %SUBJ ABBR NOM SG
      "gnp" { bruttokansantuote N48-C } NOCAP NOPROP %SUBJ ABBR
NOM SG
      "gnp" { BKT N2 } NOCAP NOPROP %SUBJ ABBR NOM SG
      "gnp" { PROP-CAND } NOCAP NOPROP %SUBJ ABBR NOM SG
```

Now the English acronym can ba translated as acronym in target language (15).

(15)
This year, the GNP has risen.
Tänä vuotena BKT on noussut.

There has been an increase in GNP.
On ollut lisäys BKT:ssa.

3 Proper names

From the viewpoint of machine translation, proper names can be divided into those which are listed in the dictionary, and those which are unknown to the system. Both types are discussed below.

3.1 Proper names known to the system

Proper names can consist of a single word or a cluster of words. If possible, clusters of words should be treated as multiword expressions (MWE). Single-word proper names can be divided into those which should be translated and those that should be transferred to target language as such. Especially the translated ones should be listed in the dictionary for getting correct translated form. Fortunately, many names of this group are common in language and their number is limited. Most proper names, however, need no translation, and they should be treated as a separate group.

Typical translated names are place names and most country names. Also here applies the principle of proximity. The closer the place is and the more frequent the contacts with the place are, the more likely the name is translated. Because there are no global means for translating such names, they should be listed in the dictionary. When listing them in the dictionary, it is also easier to control that each name inflects properly. Consider the example in (16).

(16)

```
"<France>"
      "france" %SUBJ CAPINIT N SG NOM
"<and>"
      "and" %CC CC
"<Germany>"
      "germany" %SUBJ CAP N SG NOM
"<are>"
      "be" %+FMAINV V PRES PL
"<countries>"
      "country" %PCOMPL-S N PL NOM
"<in>"
      "in" %<NOM PREP
"<Europe>"
      "europe" %<P CAP N SG NOM
"<.>"
      "."
```

In (16), *France* is the first word of the sentence and therefore capitalised with CAPINIT. It could also have the tag CAP, because it is a proper name, but it would be redundant. The name *Germany* is a proper name with a tag CAP, and so is also *Europe*. All three words are proper nouns also in Finnish and are translated using capital initials. The translation is in (17)'.

(17)
Ranska ja Saksa ovat maita Euroopassa.

More problematic are such cases, where the word is a proper name in English but not in target language. Consider the examples in (18).

(18)
```
"<His>"
      "he" %A> CAPINIT PRON PERS GEN SG3
"<mother>"
      "mother" %A> N SG NOM DEF
"<tongue>"
      "tongue" %SUBJ N SG NOM
"<is>"
      "be" %+FMAINV V PRES SG3
"<Finnish>"
      "finnish" %PCOMPL-S CAP N SG NOM
"<.>"
      "."
"<<s>>"
      "<s>"
"<This>"
      "this" %SUBJ CAPINIT PRON DEM SG
"<is>"
      "be" %+FMAINV V PRES SG3
"<Finnish>"
      "finnish" %A> CAP A ABS
"<government>"
      "government" %PCOMPL-S N NOM
"<.>"
      "."
```

Both sentences include the word *Finnish*, but one is analysed as a noun and another as an adjective. We cannot, however, translate *Finnish government* as *suomalainen hallitus*. This is one problem; an adjective cannot be translated as adjective. Another problem is capitalisation. If the word is interpreted as a country name, it is capitalised. If it is interpreted as a language name, it is written in lowercase. Therefore, the sentences should be translated as in (19).

(19)
Hänen äidinkielensä on suomi.
Tämä on Suomen hallitus.

More examples on how the word *English* should be translated in Finnish are in (20).

(20)
(a)
```
"<His>"
      "he" %A> CAPINIT PRON PERS GEN SG3
"<mother>"
      "mother" %A> N SG NOM DEF
```

```
"<tongue>"
      "tongue" %SUBJ N SG NOM
"<is>"
      "be" %+FMAINV V PRES SG3
"<Finnish>"
      "finnish" %PCOMPL-S CAP N SG NOM
"<.>"
      "."
```

(b)

```
"<He>"
      "he" %SUBJ CAPINIT PRON PERS NOM SG3
"<speaks>"
      "speak" %+FMAINV V PRES SG3
"<Finnish>"
      "finnish" %OBJ CAP N SG
"<.>"
      "."
```

(c)

```
"<He>"
      "he" %SUBJ CAPINIT PRON PERS NOM SG3
"<speaks>"
      "speak" %+FMAINV V PRES SG3
"<Finnish>"
      "finnish" %A> CAP A ABS
"<language>"
      "language" %OBJ N SG
"<.>"
      "."
```

(d)

```
"<This>"
      "this" %SUBJ CAPINIT PRON DEM SG
"<is>"
      "be" %+FMAINV V PRES SG3
"<Finnish>"
      "finnish" %A> CAP A ABS
"<capital>"
      "capital" %PCOMPL-S N SG NOM
"<.>"
      "."
```

(e)

```
"<This>"
      "this" %SUBJ CAPINIT PRON DEM SG
"<is>"
      "be" %+FMAINV V PRES SG3
"<Finnish>"
      "finnish" %A> CAP A ABS
"<government>"
      "government" %PCOMPL-S N NOM
"<.>"
      "."
```

(f)

```
"<This>"
      "this" %SUBJ CAPINIT PRON DEM SG
"<is>"
      "be" %+FMAINV V PRES SG3
"<the>"
      "the" %DN> DET
"<Finnish>"
      "finnish" %A> CAP A ABS DEF
"<Ministry>"
      "ministry" %PCOMPL-S CAP N SG NOM DEF DEF
"<of>"
      "of" %<NOM-OF PREP
"<Foreign>"
      "foreign" %A> CAP A ABS
"<Affairs>"
      "affair" %<P CAP N PL NOM
"<.>"
      "."
```

(g)

```
"<He>"
      "he" %SUBJ CAPINIT PRON PERS NOM SG3
"<is>"
      "be" %+FMAINV V PRES SG3
"<Finnish>"
      "finnish" %PCOMPL-S CAP N SG NOM
"<.>"
      "."
```

In (a), (b) and (c), the word *Finnish* is in the sense of *language*. In (a) and (b), it is analysed as noun, and in (c) it is analysed as adjective. In all three cases, however, the word must be translated as a noun and written in lowercase. In (d), the word *Finnish* is analysed as adjective, but it must be translated as noun with capital initial. The words *Finnish capital* could also be rephrased as *capital of Finland*, whereby no ambiguity would be involved. The same problematic concerns also (e), where *Finnish government* could be rephrased as *government of Finland*. In (f), such rephrasing is not possible, and the word *Finnish*, analysed as adjective, must be translated as a noun and counry name with cpital initial. Finally, in (g) we have a case, where *Finnish* means nationality and is therefore written in lowercase. The translation of the sentences is in (21).

(21)
(a) *Hänen äidinkielensä on suomi.*
(b) *Hän puhuu suomea.*
(c) *Hän puhuu suomen kieltä.*
(d) *Tämä on Suomen pääkaupunki.*
(e) *Tämä on Suomen hallitus.*
(f) *Tämä on Suomen Ulkoasiain Ministeriö.*
(g) *Hän on suomalainen.*

3.2 Proper names not known to the system

It is not possible to maintain a covering list of proper names, because new names are constantly found in text. Therefore, we must have heuristic means for handling such cases. Consider the case in (22).

(22)

```
"<Barack>"
      "barack" { parakki N5-A } %A> CAPINIT Heur N NOM SG
      "barack" { PROP-CAND } %A> CAPINIT Heur N NOM SG
"<Obama>"
      "obama" { Obama N10 } HUM %SUBJ CAP N NOM SG
      "obama" { PROP-CAND } HUM %SUBJ CAP N NOM SG
"<is>"
      "be" { olla V67b BE V-4INF-TRA } O-LOC1 %+FMAINV V PRES SG3
      "be" { olla V67b V-3INF-ILL } O-LOC1 %+FMAINV V PRES SG3
      "be" { olla V67b V-3INF-INE } O-LOC1 %+FMAINV V PRES SG3
      "be" { olla V67b BE O-PAR } O-LOC1 %+FMAINV V PRES SG3
      "be" { eivät ole :2 V-4INF-TRA } O-LOC1 %+FMAINV V PRES SG3
      "be" { eivät olleet :3 V-4INF-TRA } O-LOC1 %+FMAINV V PRES
SG3
      "be" { eivät ole :2 O-PAR V-4INF-TRA } O-LOC1 %+FMAINV V
PRES SG3
      "be" { eivät olleet :3 O-PAR V-4INF-TRA } O-LOC1 %+FMAINV V
PRES SG3
      "be" { emme :6 } O-LOC1 %+FMAINV V PRES SG3
      "be" { emme ole V-3INF-INE } O-LOC1 %+FMAINV V PRES SG3
      "be" { emme olleet V-3INF-INE } O-LOC1 %+FMAINV V PRES SG3
      "be" { ei ole V-4INF-TRA } O-LOC1 %+FMAINV V PRES SG3
      "be" { ei ollut V-4INF-TRA } O-LOC1 %+FMAINV V PRES SG3
      "be" { ei ole O-PAR V-4INF-TRA } O-LOC1 %+FMAINV V PRES SG3
      "be" { NOGLOSS } O-LOC1 %+FMAINV V PRES SG3
      "be" { joka Np13 } O-LOC1 %+FMAINV V PRES SG3
      "be" { jotka Np14 } O-LOC1 %+FMAINV V PRES SG3
      "be" { tulla V67 V-3INF-ILL } O-LOC1 %+FMAINV V PRES SG3
"<former>"
      "former" { aiempi N16-H } %A> DEF A ABS
      "former" { entinen NEN N38 FRONT } %A> DEF A ABS
"<president>"
      "president" { presidentti N5-C FRONT } HUM TITLE NOPROP
%PCOMPL-S DEF N NOM SG
      "president" { puhemies N42 FRONT } HUM TITLE NOPROP %PCOMPL-
S DEF N NOM SG
      "president" { puheenjohtaja N10 } HUM TITLE NOPROP %PCOMPL-S
DEF N NOM SG
"<of>"
      "of" { M-LOC2 } %<NOM-OF PREP
      "of" { NOGLOSS M-GEN } %<NOM-OF PREP
      "of" { NOGLOSS M-ELA } %<NOM-OF PREP
      "of" { NOGLOSS M-ABL } %<NOM-OF PREP
```

```
      "of" { NOGLOSS M-ACC-N } %<NOM-OF PREP
      "of" { NOGLOSS M-PAR } %<NOM-OF PREP
      "of" { NOGLOSS :2 } %<NOM-OF PREP
"<USA>"
      "usa" { usa N2 } MAA IN NOPROP %<P CAPALL N NOM SG
      "usa" { PROP-CAND } MAA IN NOPROP %<P CAPALL N NOM SG
"<.>"
      "." { . }
```

For finding unknown proper names in text, the first thing to do is to find and mark such words, which are written with capitial initial in the sentence. Those words, which are not sentence-initial, are good candidates for this role. The marking is done by giving an extra interpretation with the tag PROP-CAND meaning that this is a candidate for being analysed as a proper name. Now all such words, which might have this interpretation, are marked. Other words, if they have not already been marked with PROPNAME, can in no case be interpreted as proper names.

In (22) above, the words *Barck*, *Obama*, and *USA* have such a tag. We see that *Obama* is listed in the dictionary, but not as proper name. The word *Barack* is not listed as a proper name, but it is listed as an ordinary noun meaning *parakki*.

Whether the noun is a proper name or not is decided on the basis of context constraints. The example is idsambiguated in (23).

(23)

```
"<Barack>"
      "barack" { Barack } %A> CAP PROPNAME CAPINIT Heur N NOM SG
"<Obama>"
      "obama" { Obama } HUM %SUBJ CAP PROPNAME N NOM SG
"<is>"
      "be" { olla V67b BE V-4INF-TRA } O-LOC1 %+FMAINV V PRES SG3
"<former>"
      "former" { aiempi N16-H } %A> DEF A ABS
"<president>"
      "president" { presidentti N5-C FRONT } HUM TITLE NOPROP
%PCOMPL-S DEF N NOM SG
"<of>"
      "of" { NOGLOSS M-GEN } %<NOM-OF PREP
"<USA>"
      "usa" { usa N2 } MAA IN NOPROP %<P CAPALL N NOM SG
"<.>"
      "." { . }
```

The interpretation of proper name for *Obama* was chosen on the basis of two criteria. The word must be in the list of candidates for proper names. This list includes words, which may occur in text as proper names and ordinary words. The second criterion is that it must meet the required context conditions, in this case that the next word has a tag CAP.

Also, for the word *Obama*, the system chose the PROP-CAND interpretation. However, this was not done on the basis of the PROP-CAND list, because the word is not

found there. The selection was done on the basis of the previous word, whic has the PROP-CAND tag and therefore *Obama* is a good candidate for proper name.
For *USA*, the interpretetion of PROP-CAND was not chosen, because it has the tag NOPROP. Instead, the interpretation listed in the dictionary was chosen.

Note that although *Obama* was listed in the dictionary, the system missed it and chose the heuristic PROP-CAND interpretation. Now both *Barack* and *Obama* are without inflection tags, because the gloss is simply the copy of the English word. Below we shall see, how we can add inflection tags to unknown proper names (24).

(24)
```
"<Barack>"
      "barack" { Barack N1b } %A> CAP PROPNAMEINIT Heur N NOM SG
"<Obama>"
      "obama" { Obama N9 } HUM %SUBJ CAP PROPNAME N NOM SG
"<is>"
      "be" { olla V67b BE V-4INF-TRA } O-LOC1 %+FMAINV V PRES SG3
"<former>"
      "former" { aiempi N16-H } %A> DEF A ABS
"<president>"
      "president" { presidentti N5-C FRONT } HUM TITLE NOPROP
%PCOMPL-S DEF N NOM SG
"<of>"
      "of" { NOGLOSS M-GEN } %<NOM-OF PREP
"<USA>"
      "usa" { usa N2 } MAA IN NOPROP %<P CAPALL N NOM SG
"<.>"
      "." { . }
```

Inflection tags for *Barack* and *Obama* were added on the basis of the letter combination at the end of the word. This method is prone to errors, but better than nothing. It is also difficult to decide between front/back inflection. The translation is in (25).

(25)
Barack Obama on aiempi USA:n presidentti.

The proper names can also be inflected, as is demonstrated in (26).

(26)
```
"<Barack>"
      "barack" { Barack N1b } %A> CAP PROPNAMEINIT Heur N SG NOM
"<Obama>"
      "obama" { Obama N10 } %SUBJ HUM CAP N SG ADE
"<has>"
      "have" { olla V67b } %+FMAINV S-ADE O-PAR O-ACC-N V PRES SG
"<a lot of>"
      "a lot of" { paljon } %DN> M-PAR DET PAR
"<potential>"
      "potential" { potentiaali N5 } %OBJ N SG PAR
"<.>"
```

```
      "." { . }
```

Note that *Barack* is in nominative and *Obama* in adessive. Translation is in (27).

(27)
Barack Obamalla on paljon potentiaalia.

4 Multiword expressions as proper names

Many multiword espressions (MWE) are in fact proper names. Also here we encounter the same problem as above with single words. Should the MWE be translated or not? Consider the examples in (28).

(28)
(a)

```
"<It>"
      "it" { se Np11 FRONT OUT } %SUBJ CAPINIT PRON NOM SG3
      "it" { sen } %SUBJ CAPINIT PRON NOM SG3
      "it" { NOGLOSS } %SUBJ CAPINIT PRON NOM SG3
      "it" { itse N8 FRONT } %SUBJ CAPINIT PRON NOM SG3
      "it" { PROP-CAND } %SUBJ CAPINIT PRON NOM SG3
"<is>"
      "be" { olla V67b BE V-4INF-TRA } O-LOC1 %+FMAINV V PRES SG3
      "be" { olla V67b V-3INF-ILL } O-LOC1 %+FMAINV V PRES SG3
      "be" { olla V67b V-3INF-INE } O-LOC1 %+FMAINV V PRES SG3
      "be" { olla V67b BE O-PAR } O-LOC1 %+FMAINV V PRES SG3
      "be" { eivät ole :2 V-4INF-TRA } O-LOC1 %+FMAINV V PRES SG3
      "be" { eivät olleet :3 V-4INF-TRA } O-LOC1 %+FMAINV V PRES
SG3
      "be" { eivät ole :2 O-PAR V-4INF-TRA } O-LOC1 %+FMAINV V
PRES SG3
      "be" { eivät olleet :3 O-PAR V-4INF-TRA } O-LOC1 %+FMAINV V
PRES SG3
      "be" { emme :6 } O-LOC1 %+FMAINV V PRES SG3
      "be" { emme ole V-3INF-INE } O-LOC1 %+FMAINV V PRES SG3
      "be" { emme olleet V-3INF-INE } O-LOC1 %+FMAINV V PRES SG3
      "be" { ei ole V-4INF-TRA } O-LOC1 %+FMAINV V PRES SG3
      "be" { ei ollut V-4INF-TRA } O-LOC1 %+FMAINV V PRES SG3
      "be" { ei ole O-PAR V-4INF-TRA } O-LOC1 %+FMAINV V PRES SG3
      "be" { NOGLOSS } O-LOC1 %+FMAINV V PRES SG3
      "be" { joka Np13 } O-LOC1 %+FMAINV V PRES SG3
      "be" { jotka Np14 } O-LOC1 %+FMAINV V PRES SG3
      "be" { tulla V67 V-3INF-ILL } O-LOC1 %+FMAINV V PRES SG3
"<important>"
      "important" { tärkeä N15 FRONT } %PCOMPL-S A ABS
      "important" { tärkeämpi N16-H FRONT } %PCOMPL-S A ABS
      "important" { tärkein N51 FRONT } %PCOMPL-S A ABS
      "important" { merkittävä VAA N10 FRONT } %PCOMPL-S A ABS
"<to>"
```

```
      "to" { NOGLOSS } %INFMARK> INFMARK>
"<read>"
      "read" { lukea V58-D O-PAR } %-FMAINV V INF
"<daily>"
      "daily" { päivittäinen NEN N38 FRONT } %A> A ABS
      "daily" { jokapäiväinen NEN N38 FRONT } %A> A ABS
"<news>"
      "news" { uutinen N38 } %OBJ N PL
      "news" { uutis-- COMP } %OBJ N PL
"<.>"
      "." { . }
```

(b)

```
"<It>"
      "it" { se Np11 FRONT OUT } %SUBJ CAPINIT PRON NOM SG3
      "it" { sen } %SUBJ CAPINIT PRON NOM SG3
      "it" { NOGLOSS } %SUBJ CAPINIT PRON NOM SG3
      "it" { itse N8 FRONT } %SUBJ CAPINIT PRON NOM SG3
      "it" { PROP-CAND } %SUBJ CAPINIT PRON NOM SG3
"<is>"
      "be" { olla V67b BE V-4INF-TRA } O-LOC1 %+FMAINV V PRES SG3
      "be" { olla V67b V-3INF-ILL } O-LOC1 %+FMAINV V PRES SG3
      "be" { olla V67b V-3INF-INE } O-LOC1 %+FMAINV V PRES SG3
      "be" { olla V67b BE O-PAR } O-LOC1 %+FMAINV V PRES SG3
      "be" { eivät ole :2 V-4INF-TRA } O-LOC1 %+FMAINV V PRES SG3
      "be" { eivät olleet :3 V-4INF-TRA } O-LOC1 %+FMAINV V PRES
SG3
      "be" { eivät ole :2 O-PAR V-4INF-TRA } O-LOC1 %+FMAINV V
PRES SG3
      "be" { eivät olleet :3 O-PAR V-4INF-TRA } O-LOC1 %+FMAINV V
PRES SG3
      "be" { emme :6 } O-LOC1 %+FMAINV V PRES SG3
      "be" { emme ole V-3INF-INE } O-LOC1 %+FMAINV V PRES SG3
      "be" { emme olleet V-3INF-INE } O-LOC1 %+FMAINV V PRES SG3
      "be" { ei ole V-4INF-TRA } O-LOC1 %+FMAINV V PRES SG3
      "be" { ei ollut V-4INF-TRA } O-LOC1 %+FMAINV V PRES SG3
      "be" { ei ole O-PAR V-4INF-TRA } O-LOC1 %+FMAINV V PRES SG3
      "be" { NOGLOSS } O-LOC1 %+FMAINV V PRES SG3
      "be" { joka Np13 } O-LOC1 %+FMAINV V PRES SG3
      "be" { jotka Np14 } O-LOC1 %+FMAINV V PRES SG3
      "be" { tulla V67 V-3INF-ILL } O-LOC1 %+FMAINV V PRES SG3
"<important>"
      "important" { tärkeä N15 FRONT } %PCOMPL-S A ABS
      "important" { tärkeämpi N16-H FRONT } %PCOMPL-S A ABS
      "important" { tärkein N51 FRONT } %PCOMPL-S A ABS
      "important" { merkittävä VAA N10 FRONT } %PCOMPL-S A ABS
"<to>"
      "to" { NOGLOSS } %INFMARK> INFMARK>
"<read>"
      "read" { lukea V58-D O-PAR } %-FMAINV V INF
"<Daily>"
      "daily" { päivittäinen NEN N38 FRONT } %A> CAP A ABS
```

```
      "daily" { jokapäiväinen NEN N38 FRONT } %A> CAP A ABS
      "daily" { PROP-CAND } %A> CAP A ABS
"<News>"
      "news" { uutinen N38 } %OBJ CAP N SG
      "news" { uutis-- COMP } %OBJ CAP N SG
      "news" { PROP-CAND } %OBJ CAP N SG
"<.>"
      "." { . }
```

In (a), we have the words *daily news*, written in lowercase. These are ordinary words and will be translated accordingly. In (b), the same words are written as *Daily News* indicating that these may be proper names each, or that they together may constitute a MWE. If they are proper names, as they in fact are, we have to decide whether they should be translated or not. *Daily News* is a name of a newspaper and it should not be translated. However, it should inflect in Finnish as other words do. When we disambiguate the sentences, we get a result as in (29).

(29)
(a)

```
"<It>"
      "it" { NOGLOSS } %SUBJ CAPINIT PRON NOM SG3
"<is>"
      "be" { olla V67b BE V-4INF-TRA } O-LOC1 %+FMAINV V PRES SG3
"<important>"
      "important" { tärkeä N15 FRONT } %PCOMPL-S A ABS
"<to>"
      "to" { NOGLOSS } %INFMARK> INFMARK>
"<read>"
      "read" { lukea V58-D O-PAR } %-FMAINV V INF
"<daily>"
      "daily" { päivittäinen NEN N38 FRONT } %A> A ABS
"<news>"
      "news" { uutinen N38 } %OBJ N PL
"<.>"
      "." { . }
```

(b)

```
"<It>"
      "it" { NOGLOSS } %SUBJ CAPINIT PRON NOM SG3
"<is>"
      "be" { olla V67b BE V-4INF-TRA } O-LOC1 %+FMAINV V PRES SG3
"<important>"
      "important" { tärkeä N15 FRONT } %PCOMPL-S A ABS
"<to>"
      "to" { NOGLOSS } %INFMARK> INFMARK>
"<read>"
      "read" { lukea V58-D O-PAR } %-FMAINV V INF
"<Daily>"
      "daily" { Daily N1b } %A> CAP PROPNAME A ABS
"<News>"
      "news" { News N1b } %OBJ CAP PROPNAME N SG
```

```
"<.>"
      "." { . }
```

We see that the system adds inflection codes to each member of the MWE *Daily News*. The translation of sentences is in (30).

(30)
(a) *On tärkeää lukea päivittäisiä uutisia.*
(b) *On tärkeää lukea Dailyia Newsia.*

We see that when the members of the MWE are handled separately, all members will inflect, although *Daily* should not inflect. This can be avoided, when *Daily News* is handled in a proper way as an MWE, such as in (31).

(31)

```
"<It>"
      "it" { NOGLOSS } %SUBJ CAPINIT PRON NOM SG3
"<is>"
      "be" { olla V67b BE V-4INF-TRA } O-LOC1 %+FMAINV V PRES SG3
"<important>"
      "important" { tärkeä N15 FRONT } %PCOMPL-S A ABS
"<to>"
      "to" { NOGLOSS } %INFMARK> INFMARK>
"<read>"
      "read" { lukea V58-D O-PAR } %-FMAINV V INF
"<Daily_News>"
      "daily_news" { *daily *news N1b } %OBJ MW CAP N SG
"<.>"
      "." { . }
```

Now we get the correct translation (32).

(32)
On tärkeää lukea Daily Newsia.

In other cases, the MWE proper name must be translated. Consider the example in (33).

(33)

```
"<We>"
      "we" %SUBJ CAPINIT PRON PERS NOM PL1
"<will>"
      "will" %+FAUXV V AUXMOD
"<meet>"
      "meet" %-FMAINV V INF
"<the>"
      "the" %DN> DET
"<Minister>"
      "minister" %OBJ CAP N SG DEF
"<of>"
```

```
      "of" %<NOM-OF PREP
"<Foreign>"
      "foreign" %A> CAP A ABS
"<Affairs>"
      "affair" %<P CAP N PL NOM
"<.>"
      "."
```

When we add Finnish glosses, the result is as in (34).

(34)

```
"<We>"
      "we" { me Np6 FRONT OUT } HUM %SUBJ CAPINIT PRON PERS NOM
PL1
      "we" { meidän } HUM %SUBJ CAPINIT PRON PERS NOM PL1
      "we" { NOGLOSS } HUM %SUBJ CAPINIT PRON PERS NOM PL1
      "we" { itse N8 FRONT } HUM %SUBJ CAPINIT PRON PERS NOM PL1
      "we" { PROP-CAND } HUM %SUBJ CAPINIT PRON PERS NOM PL1
"<will>"
      "will" { NOGLOSS } %+FAUXV V AUXMOD
      "will" { aikoa V52-D } %+FAUXV V AUXMOD
      "will" { tulla V67 } %+FAUXV V AUXMOD
"<meet>"
      "meet" { kohdata V73-F O-ACC V-1INF-TRA } %-FMAINV V INF
      "meet" { kohdata V73-F V-1INF-TRA } %-FMAINV V INF
      "meet" { täyttää V53-C FRONT O-ACC } %-FMAINV V INF
      "meet" { tavata V73 O-ACC } %-FMAINV V INF
      "meet" { kokoontua V52-J } %-FMAINV V INF
"<Minister>"
      "minister" { ministeri N6 FRONT } HUM TITLE NOPROP %OBJ DEF
CAP N SG
      "minister" { PROP-CAND } HUM TITLE NOPROP %OBJ DEF CAP N SG
"<of>"
      "of" { M-LOC2 } %<NOM-OF PREP
      "of" { NOGLOSS M-GEN } %<NOM-OF PREP
      "of" { NOGLOSS M-ELA } %<NOM-OF PREP
      "of" { NOGLOSS M-ABL } %<NOM-OF PREP
      "of" { NOGLOSS M-ACC-N } %<NOM-OF PREP
      "of" { NOGLOSS M-PAR } %<NOM-OF PREP
      "of" { NOGLOSS :2 } %<NOM-OF PREP
"<Foreign_Affairs>"
      "foreign_affair" { ulkoasia N12 } %<P MW CAP N PL NOM
      "foreign_affair" { PROP-CAND } %<P MW CAP N PL NOM
"<.>"
      "." { . }
```

The word *Minister* has a tag NOPROP indicating that this should be translated directly and not treated as a proper name. The words *Foreign* and *Affairs* have aleady been joined as a MWE. This cluster has two interpretations, *ulkoasia* and PROP-CAND. The disambiguated result is in (35).

(35)

```
"<We>"
      "we" { me Np6 FRONT OUT } HUM %SUBJ CAPINIT PRON PERS NOM
PL1
"<will>"
      "will" { NOGLOSS } %+FAUXV V AUXMOD
"<meet>"
      "meet" { kohdata V73-F O-ACC V-1INF-TRA } %-FMAINV V INF
"<Minister>"
      "minister" { ministeri N6 FRONT } HUM TITLE NOPROP %OBJ DEF
CAP N SG
"<of>"
      "of" { NOGLOSS M-GEN } %<NOM-OF PREP
"<Foreign_Affairs>"
      "foreign_affair" { ulkoasia N12 } %<P MW CAP N PL NOM
"<.>"
      "." { . }
```

There was no rule for selecting between the two interpretations of *Foreign Affairs*. The system uses defaults, and the default is that if the word has the normal interpretation and the PROP-CAND interpretation, and no rule selects the latter interpretation, choose the first one. The translation is in (36).

(36)
Me kohtaamme Ulkoasioiden Ministerin.

Although the translation in (36) is acceptable, it does follow the current othographic practice. The proper name *Ulkoasioiden Ministeri* is preferably glossed as *ulkoministeri*, as a single word and in lowercase. This can be done in two ways. In one method, we provide the words *Minister* and *Foreign_Affairs* with the tag NOCAP. When we do this, the translation is as in (37).

(37)
Me kohtaamme ulkoasioiden ministerin.

Now the proper name in English is in lowercase in Finnish, but not yet in the form we wish. When we add for the MWE *Foreign_Affais* the possibility of being part in compound (code: ulko-- COMP), we get the desired result (38).

(38)
Me kohtaamme ulkoministerin.

5 Conclusion

The discussion shows that handling the uppercase problem in MT is far from simple. In this report I have not handled such questions as the names of months ad week days, which in English are written with capital initial but in Finnish in lowercase. These are closed sets and can be handled normally without difficulties in this translation task. An exception is the Swahili to English translation, where weekdays are often used also as person names. This question is dicussed earlier in this report series.

Although we have seen how various problems can ba handled in a rule-based translation system, we still frequently encounter cases, which require new rules and their testing.