Technical Reports on Language Technology
Report No 24, 2017
http://www.njas.helsinki.fi/salama

Anaphora in English to Finnish machine translation[1]

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

Abstract

From the viewpoint of machine translation, anaphoric expressions can be classified as overt and covert anaphora. An overt anaphora is an expression that can refer to virtually any referent, the specific referent being defined by context. A covert anaphora has a similar role, but the expression itself is covert, not represented as an overt expression. The report discusses the problems included into the translation process of both types of anaphoric expressions.

Keywords: *anaphora, machine translation.*

1 Introduction

From the viewpoint of machine translation, anaphoric expressions can be classified as overt and covert anaphora. An overt anaphora is an expression that can refer to virtually any referent, the specific referent being defined by context. A covert anaphora has a similar role, but the expression itself is covert, not represented as an overt expression. The report discusses the problems included into the translation process of both types of anaphoric expressions.

The translation of anaphoric expression from English to Finnish is in most cases unproblematic. Pronouns such as 'this' or 'that', and their plural forms, can be normally translated without problems. However, there are several cases, where the task is not straightforward. For example, Finnish allows the omission of personal pronouns in SG1 and SG2 as well as in PL1 and PL2. It is sometimes hard to decide whether the personal pronoun should be included or not. This problem does not disturb understanding, however. In such cases the question is about style rather than of grammatical correctness. More problematic are such cases, where the source language uses an anaphoric expression, but the target language does not allow it.

A problem of its own is the trend in English to construct the subordinate clause without conjunction and the relative sentence without relative pronoun. The challenge in translation is to add the missing conjunction and the missing relative pronoun in target

language. In the case of the missing conjunction, it is also possible to use a participial phrase construction instead of the subordinate clause.
Note that in the examples below articles are removed, because they are not necessary in displaying the problems.

2 Covert anaphora

First, we discuss the methods of handling covert anaphora, that is, cases where the anaphoric expression is assumed but not expressed overtly.

2.1 Missing conjunction in subordinate clauses

In modern writing and speech of English, the conjunction is frequently omitted in subordinate clauses. However, this does not take place haphazardly. The investigation into the matter shows that certain rules seem to constrain the phenomenon.

The first constraint is the type of verb of the main clause, after which the subordinate clause comes. For example, the verbs such as: (*say, speak, tell, suggest, demand, ask, urge, shout, whisper, think, dream, ensure, guess*) occur as the finite verb of the main clause.

The second constraint is that the verb of the main clause must be close to the end of the clause. In fact, often it is the last word of the main clause, but also an object or indirect object is allowed after the verb. In case the verb is more distantly placed on the left, the conjunction is normally included.

In (1-4) I will demonstrate how the conjunction can be added to Finnish translation.

(1)
The sentence "He told he will return tomorrow" does not have a conjunction. The conjunction *että* is added to the verb of the main clause, so that it starts the subordinate clause in translation (2).

(2)
```
"<*he>"
      "he" { *hän Np9 FRONT } %SUBJ CAPINIT PRON PERS NOM SG3
"<told>"
      "tell" { kertoa V52-K   } %+FMAINV O-ALL V PAST { , että }
@SG
"<he>"
      "he" { hän Np9 FRONT } %SUBJ PRON PERS NOM SG3
"<will>"
      "will" { NOGLOSS } %+FAUXV V AUXMOD @SG
"<return>"
      "return" { palata V73  } %-FMAINV O-ILL V INF @SG
"<tomorrow>"
      "tomorrow" { huomenna } %ADVL ADV
```

The rule for adding the conjunction in the correct place in (2) is based on the following algorithm: To the verb that belongs to the group *COMM*, add the string *'{ , että }*', if on

the right there is a finite verb (do not scan further than to clause boundary), and on the left there is subject or verb imperative (do not scan further than to clause boundary). Immediately on the right there should not be any of the following tags: *COMMA, CC, EN, NEG, OBJ, A, GEN*. After the second cohort to the right there should not be the tag *OBJ* (do not scan further than to clause boundary or verb). In the second cohort to the left there should not be the tag belonging to the set *REL* or *"if"*. In the first cohort to the left there should not be *COMMA* or the tag belonging to the set *REL*.

Final translation is in (3).

(3)
Hän kertoi, että hän palaa huomenna

If we add to the sentence an indirect object, we will get the form as in (3).

(4)
He told his neighbours he will return tomorrow

In this case, the conjunction is added to the indirect object, and not to the verb (5).

(5)
```
"<*he>"
      "he" { *hän Np9 FRONT } %SUBJ CAPINIT PRON PERS NOM SG3
"<told>"
      "tell" { kertoa V52-K  } %+FMAINV O-ALL V PAST @SG
"<his>"
      "he" { hän Np9 FRONT } %A> PRON PERS GEN SG3 @PL
"<neighbours>"
      "neighbour" { naapuri N6 } %I-OBJ HUM N PL NOM { , että }
"<he>"
      "he" { hän Np9 FRONT } %SUBJ PRON PERS NOM SG3
"<will>"
      "will" { NOGLOSS } %+FAUXV V AUXMOD @SG
"<return>"
      "return" { palata V73  } %-FMAINV O-ILL V INF @SG
"<tomorrow>"
      "tomorrow" { huomenna } %ADVL ADV
```

The rule for achieving this is formulated according to the following algorithm: To the object (*%OBJ*) or indirect object (*%I-OBJ*) add the string '*{ , että }*', if on the left there is a finite verb belonging to the set COMM (do not scan further than to the clause boundary or verb), and to the right is a finite verb (do not scan further than the clause boundary). The immediate cohort to the right should not have *COMMA* or *CC*.

The final translation is in (6).

(6)
Hän kertoi naapureilleen, että hän palaa huomenna

2.2 Missing relative pronoun

Another instance of covert anaphora is the omission of the relative pronoun, whereby the expression is constructed using the participial or gerund form of the verb. In Finnish, such constructions should normally be translated using a relative clause, and the relative pronoun should be added. The challenge is how to insert the pronoun into the text, where there is no definite point to show where it should be inserted.

Perhaps the most common case is that the relative pronoun is omitted after the verb object (7).

(7)

```
"<*mr>"
      "mr" { *herra N9 } %A> HUM TITLE CAP ABBR NOM SG
"<*putin>"
      "putin" { *putin N1b } %SUBJ HUM CAP N NOM SG
"<got>"
      "get" { saada V63  } %+FMAINV O-ACC O-TRA V-3INF-ILL V PAST
SG
"<apology>"
      "apology" { anteeksipyyntö N1-J FRONT } %OBJ N NOM SG { ,
joka Np13 } <REL> ACC
"<he>"
      "he" { hän Np9 FRONT } %SUBJ PRON PERS NOM SG3
"<demanded>"
      "demand" { vaatia V61-F  } %+FMAINV O-PAR V-3INF-ILL V PAST
SG
"<from>"
      "from" { NOGLOSS  } %ADVL M-ABL PREP
"<*president>"
      "president" { *presidentti N5-C FRONT } %A> HUM CAP N NOM SG
"<*erdogan>"
      "erdogan" { *erdogan N1b } %<P CAP N NOM SG
```

The string '*{ , joka Np13 }*' was added in (7) into the analysis of the verb object *"apology"*. Note that the gloss *joka* has the inflection code Np13, because the pronoun may inflect.

The algorithm for the rule that inserts the relative pronoun is: Into the object of the sentence add the string '*{ , joka Np13 } <REL> ACC*', if immediately on the right is the subject, and in the second cohort to the right is a finite verb. The target should not include the tag *<Rel>* or *"what"*. The first cohort to the right should not include the tag *<Rel>*.

The translation is in (8). Note that the relative pronoun is in singular accusative (genitive version).

(8)
Herra Putin sai anteeksipyynnön, jonka hän vaati Presidentti Erdoganilta

A slightly different example is in (9).

(9)
```
"<*erdogan>"
      "erdogan" { *erdogan N1b } %SUBJ CAP N NOM SG
"<needs>"
      "need" { tarvita V69  } %+FMAINV O-PAR V-3INF-ILL V PRES SG3
"<all>"
      "all" { kaikki N7-A } %DN> DET PL
"<friends>"
      "friend" { ystävä N10 FRONT } %OBJ HUM N PL NOM { , joka
Np13 } <REL> ACC
"<he>"
      "he" { hän Np9 FRONT } %SUBJ PRON PERS NOM SG3
"<can>"
      "can" { voida V62 } %+FAUXV V AUXMOD SG
"<get>"
      "get" { saada V63  } %-FMAINV O-ACC O-TRA V-3INF-ILL V INF
SG
```

The translation is in (10). Here again, the pronoun is in accusative, but because it is in plural, it takes a nominative form!

(10)
Erdogan tarvitsee kaikkia ystäviä, jotka hän voi saada

2.3 Relative pronoun after participial verb form

In (11) is an example, where the relative pronoun is omitted after the participial verb form. The verb is in a neutropassive role, because it does not have a subject or agent. The translation of such forms is problematic, because the verb form does not have a time referent. It could be translated with present or past tense, or with a participial form. None of these is ideal, because they include a time referent, which is missing in the source text. The best translation would be to treat the structure as a participial phrase construction (*kolme veteen heitettyä miestä*), but the implementation of such structures is problematic. Here we translate it with a relative structure.

(11)
```
"<three>"
      "three" { kolme N8 } %A> NUM-PL CARD NUM SG
"<men>"
      "man" { mies N42 FRONT } %SUBJ HUM N PL NOM SG
"<thrown_in>"
      "throw_in" { heittää V53-C FRONT  } %-FMAINV O-ACC O-ILL V
EN { , joka Np13 } <REL> ACC SG
"<water>"
```

```
      "water" { vesi N27 FRONT } %<P N NOM SG
"<are>"
      "be" { olla V67b BE  } %+FMAINV V-4INF-TRA V PRES SG
"<lucky>"
      "lucky" { onnekas N41-A } %PCOMPL-S A ABS PL
```

Translation:
Kolme miestä, jotka on heitetty veteen, ovat onnekkaita

2.4 Passive constructions with agent

Finnish does not have passive constructions with agent. Therefore, this type of sentences must be translated using relative constructions, and the passive structure must be converted to the active structure (12). The process involves the conversion of the participial verb form into perfect tense. Also, the word order must be reorganised.

(12)

```
"<*he>"
      "he" { *hän Np9 FRONT } %SUBJ CAPINIT PRON PERS NOM SG3
CAPINIT
"<calls_for>"
      "call_for" { vaatia V61-F  } %+FMAINV O-PAR V PRES SG3 SG
"<*press>"
      "press" { *press N1b } %<P PROPNAME N NOM SG
"<to>"
      "to" { NOGLOSS } %INFMARK> INFMARK>
"<be>"
      "be" { NOGLOSS } %-FAUXV V INF SG3
"<given>"
      "give" { antaa V56-J  } %-FMAINV O-ACC V EN SG3
"<greater>"
      "great" { suurempi N16-H } %A> A CMP SG
"<access_to>"
      "access_to" { pääsy N1 FRONT  } %OBJ M-ILL MW N NOM SG
"<council>"
      "council" { neuvosto N2 } %A> N NOM
"<papers>"
      "paper" { paperi N6 } %<P N PL NOM
"<and>"
      "and" { ja } %CC CC
"<decisions>"
      "decision" { päätös N39 FRONT } %<P N PL NOM
"<taken>"
      "take" { tehdä V71 FRONT :2  } %-FMAINV O-ACC V EN-AG { ovat
} { , joka Np13 } <REL> ACC PL PL
"<by>"
      "by" { NOGLOSS  } %ADVL AG-PART PREP
"<unofficial>"
      "unofficial" { epävirallinen N38 } %A> A ABS PL
"<committees>"
```

```
      "committee" { komitea N12 } %<P HUM N PL NOM
"<or>"
      "or" { tai } %CC CC
"<working_groups>"
      "working_group" { työryhmä N10 FRONT } %<P MW N PL NOM
```

Translation:
Hän vaatii Pressia annettavaksi suuremman pääsyn neuvoston papereihin ja päätöksiin, jotka epäviralliset komiteat tai työryhmät ovat tehneet.

2.5 Two covert anaphora

The sentence may have two or more hidden anaphora, each of which must be handled accordingly (13).

(13)

```
"<*robby>"
      "robby" { *robby N1b } %A> CAP N NOM SG
"<*miller>"
      "miller" { *miller N1b } %SUBJ CAP N NOM SG
"<said>"
      "say" { sanoa V52  } %+FMAINV O-ALL HUM V PAST { , että } SG
"<authorities>"
      "authority" { viranomainen N38 } %SUBJ N PL NOM
"<rescued>"
      "rescue" { pelastaa V53  } %+FMAINV O-ACC V PAST PL
"<72>"
      "72" { 72 } %QN> CARD NUM NUM-PL SG
"<people>"
      "people" { ihminen N38 FRONT } %OBJ HUM N NOM SG
"<and>"
      "and" { ja } %CC CC
"<seven>"
      "seven" { seitsemän N10b FRONT } %QN> NUM-PL CARD NUM SG
"<pets>"
      "pet" { lemmikki N5-A FRONT } %OBJ AN N PL NOM SG
"<stranded>"
      "strand" { jumiuttaa V53-C  } %-FMAINV O-ACC V EN-AG { on }
{ , joka Np13 } <REL> ACC SG
"<by>"
      "by" { NOGLOSS  } %ADVL AG-PART PREP
"<high>"
      "high" { korkea N15 } %A> A ABS SG
"<water>"
      "water" { vesi N27 FRONT } %<P N NOM SG
```

Translation:
Robby Miller sanoi, että viranomaiset pelastivat 72 ihmistä ja seitsemän lemmikkiä, jotka korkea vesi on jumiuttanut.

Sometimes two covert anaphora are congested, due to MWE isolation (14).

(14)
```
"<police>"
      "police" { poliisi N6 } %SUBJ HUM N NOM SG
"<say>"
      "say" { sanoa V52  } %+FMAINV O-ACC HUM V PRES SG
"<reign_of_terror>"
      "reign_of_terror" { hirmuvalta N9-I } %OBJ N NOM SG { , että
} { , joka Np13 } <REL> ACC
"<family>"
      "family" { perhe N48 FRONT } %SUBJ HUM N NOM SG
"<inflicted>"
      "inflict" { aiheuttaa V53-C  } %+FMAINV O-PAR V PAST SG
"<included>"
      "included" { sisällytetty N1-C FRONT } %A> A ABS SG
"<fighting>"
      "fighting" { taistelu N2 } %OBJ N NOM N-ING SG
"<in>"
      "in" { NOGLOSS  } %ADVL M-ADE PREP
"<street>"
      "street" { katu N1-F } %<P N NOM SG
```

Translation:
Poliisi sanoo, että hirmuvaltaan, jonka perhe aiheutti, sisältyi taistelua kadulla

2.6 Ambiguous covert anaphora

The omission of anaphora sometimes creates situations, where it is impossible to decide, whether the covert anaphora should be translated with a relative pronoun (*joka*) or subordinate conjunction (*että*). In (14) is such an example.

(14)
```
"<*family>"
      "family" { *perhe N48 FRONT } %SUBJ HUM CAPINIT N NOM
CAPINIT SG
"<tells>"
      "tell" { kertoa V52-K  } %+FMAINV O-ALL V PRES SG3
"<station>"
      "station" { asema N10 } %OBJ N NOM SG { , että } { , joka
Np13 } <REL> ACC
"<he>"
      "he" { hän Np9 FRONT } %SUBJ PRON PERS NOM SG3
"<is>"
      "be" { NOGLOSS } %+FAUXV V PRES SG3
"<expected>"
      "expect" { odottaa V53-C  } %-FMAINV O-PAR V EN SG3
```

```
"<to>"
      "to" { NOGLOSS } %INFMARK> INFMARK>
"<recover>"
      "recover" { palautua V52-F } %-FMAINV V INF SG3
```

In (14), rules have added both interpretations, one of which should be chosen. The alternative translations are:
Perhe kertoo asemalle, että hänen odotetaan palautuvan.
Perhe kertoo asemalle, jonka hänen odotetaan palauttavan.

2.7 Relative pronoun after gerund verb form

Structures with gerund verb form can be translated using relative structures or participial phrase constructions. The former ones are easier to implement without the risk of misconstructions. So, we adopt this option. The example in (15) has two types of covert anaphora.

(15)

```
"<*lille>"
      "lille" { *lille N7 } %A> CAP N NOM SG
"<court>"
      "court" { oikeusistuin N33 } %SUBJ N SG
"<ruled>"
      "rule" { päättää V53-C FRONT :2 } %+FMAINV V PAST { , että }
SG
"<there>"
      "there" { NOGLOSS } %F-SUBJ LOC <Ex> ADV @SG
"<was>"
      "be" { ei ollut  } %+FMAINV V-4INF-TRA V PAST SG
"<no>"
      "no" { NOGLOSS } %DN> DET
"<legal_basis_for>"
      "legal_basis_for" { laillinen N38 perusta N13  } %<NOM M-ALL
MW PREP
"<expelling>"
      "expel" { karkottaa V53-C  } %<P-FMAINV O-ACC V ING SG
"<people>"
      "people" { ihminen N38 FRONT } %OBJ HUM N NOM PL
"<running>"
      "run" { ylläpitää V53-F FRONT  } %-FMAINV O-PAR V ING { ,
joka Np13 } <REL> NOM PL
"<72>"
      "72" { 72 } %QN> CARD NUM NUM-PL SG
"<makeshift>"
      "makeshift" { hätävara N9 } %A> N NOM SG SG
"<shops>"
      "shop" { kauppa N9-B } %OBJ N PL NOM SG
```

Translation:
Lillen oikeusistuin päätti, että ei ollut laillista perustaa karkottaa ihmisiä, jotka ylläpitävät 72 hätävarakauppaa

3 Overt anaphora

Here we discuss such anaphora, where the anaphoric expression is overtly spelled out, but where it cannot be directly translated with a corresponding anaphora. The most outstanding example is the anaphora 'that of' and its plural 'those of'. Note that in the examples below the anaphora 'that' is joined to the following preposition 'of', to form a MWE. The example in (16) shows that the referent of the anaphora can be far away from the anaphoric expression. Yet, in order to get proper translation, the anaphoric expression must be replaced with its referent.
It is important to note that the replacement must be done on the lexical form, and not on the surface form. By doing so we ensure that the copied referent can be inflected as needed in this position.
In (16) the referent is the subject 'skin', and it is copied to the slot of the lexical form of the anaphoric expression '*that_of*' and replaced with it. When only the lexical gloss is replaced, the noun so copied can be inflected as the tags in that reading require.

(16)

```
"<results>"
      "result" %SUBJ N PL NOM
"<show>"
      "show" %+FMAINV V PRES
"<that>"
      "that" %CS CS
"<skin>"
      "skin" %SUBJ N SG NOM
"<of>"
      "of" %<NOM-OF PREP
"<many>"
      "many" %QN> DET ABS PL
"<senior>"
      "senior" %A> N SG NOM
"<*london>"
      "london" %A> CAP N SG NOM
"<citizens>"
      "citizen" %<P N PL NOM
"<will>"
      "will" %+FAUXV V AUXMOD
"<be>"
      "be" %-FMAINV V INF
"<two-and-a-half>"
      "two-and-a-half" %PCOMPL-S N SG NOM
"<years>"
      "year" %ADVL N PL NOM
"<older>"
      "old" %<NOM A CMP
```

```
"<than>"
      "than" %ADVL PREP
"<that_of>"
      "skin" %<P SG NOGLOSS-ANA N M-GEN
"<someone>"
      "someone" %<P PRON NOM SG
"<in>"
      "in" %ADVL PREP
"<countryside>"
      "countryside" %<P N SG NOM
```

The example (16) has no object but two subjects, which causes the problem of which subject should be copied. The rule chooses the nearest subject.
The selection rules are ordered in the following order:
a. Choose the nearest object on the left, if more objects are on the left.
b. Choose the nearest subject on the left, if more subjects are on the left and there are no objects.
c. Copy the object on the left (also if there is a subject on the left).
d. Copy the subject on the left (all other cases have already been treated).

Now when the anaphora has been replaced by its referent, we can add lexical glosses in Finnish (17).

(17)

```
"<results>"
      "result" { tulos N39 } %SUBJ N PL NOM
"<show>"
      "show" { osoittaa V53-C  } %+FMAINV O-PAR V PRES PL
"<that>"
      "that" { , että } %CS CS
"<skin>"
      "skin" { iho N1 } %SUBJ N NOM SG
"<of>"
      "of" { NOGLOSS  } %<NOM-OF M-GEN PREP
"<many>"
      "many" { moni N23 } %QN> DET ABS PL GEN
"<senior>"
      "senior" { vanhempi N16-H } %A> HUM N SG GEN
"<*london>"
      "london" { *lontoo N17 } %A> MAA CAP N SG GEN
"<citizens>"
      "citizen" { kansalainen N38 } %<P HUM N PL NOM
"<will>"
      "will" { NOGLOSS } %+FAUXV V AUXMOD SG
"<be>"
      "be" { olla V67b BE  } %-FMAINV V-4INF-TRA V INF SG PRES
"<two-and-a-half>"
      "two-and-a-half" { kaksi N31 ja puoli N26 } %PCOMPL-S N NOM
SG
"<years>"
```

```
      "year" { vuosi N27 } %ADVL TIME N PL NOM
"<older>"
      "old" { vanhempi N16-H } %<NOM A CMP NOM
"<than>"
      "than" { kuin } %ADVL PREP
"<that_of>"
      "skin" { iho N1 } %<P NOGLOSS-ANA N M-GEN SG
"<someone>"
      "someone" { joku N101 } %<P PRON SG GEN
"<in>"
      "in" { NOGLOSS  } %ADVL M-INE PREP
"<countryside>"
      "countryside" { maaseutu N1-F } %<P N SG INE
```

Translation:
Tulokset osoittavat, että monien vanhempien Lontoon kansalaisten iho on kaksi ja puoli vuotta vanhempi kuin jonkun iho maaseudulla

In example (18), the copied anaphora is inflected according to the instructions given in the reading.

(18)

```
"<*but>"
      "but" %CC CAP CC
"<his>"
      "he" %A> PRON PERS GEN SG3
"<vision>"
      "vision" %SUBJ N SG NOM
"<of>"
      "of" %<NOM-OF PREP
"<autonomy>"
      "autonomy" %<P N SG NOM
"<appears>"
      "appear" %+FMAINV V PRES SG3
"<to>"
      "to" %INFMARK> INFMARK>
"<differ>"
      "differ" %-FMAINV V INF
"<sharply>"
      "sharply" %ADVL ADV
"<from>"
      "from" %ADVL PREP
"<that_of>"
      "vision" %<P SG NOGLOSS-ANA N M-GEN
"<*kremlin>"
      "kremlin" %<P CAP N SG NOM
```

Glosses are added to the example in (19)

(19)

```
"<*but>"
      "but" { *mutta } %CC CAP CC
"<his>"
      "he" { hän Np9 FRONT } %A> PRON PERS SG3 GEN
"<vision>"
      "vision" { visio N3 } %SUBJ N NOM SG POS-SGPL
"<of>"
      "of" { NOGLOSS  } %<NOM-OF M-GEN PREP
"<autonomy>"
      "autonomy" { autonomia N12 } %<P N SG GEN POS-SGPL
"<appears>"
      "appear" { näyttää V53-C FRONT } %+FMAINV V PRES SG3
"<to>"
      "to" { NOGLOSS } %INFMARK> INFMARK>
"<differ>"
      "differ" { erottua V52-F } %-FMAINV V INF SG
"<sharply>"
      "sharply" { jyrkästi } %ADVL ADV
"<from>"
      "from" { NOGLOSS  } %ADVL M-ABL PREP
"<that_of>"
      "vision" { visio N3 } %<P NOGLOSS-ANA N M-GEN SG
"<*kremlin>"
      "kremlin" { *kreml N1b FRONT } %<P MAA CAP N SG GEN
```

Translation:
Mutta hänen visionsa autonomiasta näyttää erottuvan jyrkästi Kremlin visiosta

4 Discussion

Anaphora are a complex phenomenon from the viewpoint of machine translation. The discussion above shows, however, that using various computational methods it is possible to handle the problems. Part of the anaphora can be handled using context sensitive constraint grammar rules. Other types of anaphora, such as 'that of', can be managed by using regular expressions. The current implementation handles most anaphora correctly, but more testing and tuning of rules is needed.