

<https://helda.helsinki.fi>

Multi-word verbs in English to Finnish machine translation II : Multi-word verb mapping to single-word and multi-word verbs

Hurskainen, Arvi

University of Helsinki, Institute for Asian and African Studies
2020

Hurskainen , A 2020 ' Multi-word verbs in English to Finnish machine translation II :
Multi-word verb mapping to single-word and multi-word verbs ' Technical reports on
language technology , no. 52 , University of Helsinki, Institute for Asian and African Studies ,
Helsinki . < <http://www.njas.helsinki.fi/salama/multiword-verbs-in-en-to-fi-mt-II.pdf> >

<http://hdl.handle.net/10138/317507>

cc_by_nc
publishedVersion

Downloaded from Helda, University of Helsinki institutional repository.

This is an electronic reprint of the original article.

This reprint may differ from the original in pagination and typographic detail.

Please cite the original version.

Multi-word verbs in English to Finnish machine translation II: Multi-word verb mapping to single-word and multi-word verbs

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

Abstract

The report discusses such cases, where a multi-word English verb maps to a single-word or multi-word Finnish verb. Because the verbs in this group have more than one word in source language (English in this case), they must be handled as single units. This applies especially to multi-word to single-word mapping. If there are two words, for example, on the source side and target side, one could argue that they should be handled as single words. However, the word-by-word translation is often not possible even in these cases. Therefore, it is safe to handle also these cases as multi-word expressions. The isolation rules were implemented so that the second (and possibly third) member of the MWE can be immediately after the verb, is in a restricted but non-defined distance. Below we look into the problems involved in the translation of this verb group.

Key Words: *morphology, multi-words, machine translation.*

1 Introduction

The Salama translation system has transfer lexicons for each language pair. In the current development phase, the English to Finnish translation lexicon has 471 such English verbs, which are composed of two parts. A total of 387 of them are such verbs, where a multi-word verb maps to a single-word verb. A total of 84 of them are such, where a multi-word verb maps to a multi-word verb. That is, both sides have at least two verb components.

The number of multi-word verbs above was extracted from the multi-word lexicon, which contains all types of multi-words. The list is in no way fixed. Many of those listed as multi-words could also be handled as separate words. Yet it has been found convenient to describe them as multi-words, because this method reduces the need to write disambiguation rules.

Many of the English multi-word verbs are composed of the verb and a second member, which often is an adverb or preposition. Also nouns, adjectives, verbs and numerals occur as the second member. These second members are often ambiguous. If they would be handled as separate words, they would require a lot of rule writing. Now when the combination of the verb and the second member is isolated as a single unit, it can be given a gloss together with inflection instructions. Especially in translating into

Finnish this is important, because the verb defines, what kind of inflection structure the modifiers of the verb take.

A further problem with English multi-word verbs is that the second member is not necessarily located immediately after the verb. For example, it is often after the object, adverb or some other verb modifier. This adds a further challenge to the translation system.

The problems inherent in translating multi-word verbs can be divided into two groups, (a) how to identify and handle such multi-word verbs, which are non-contiguous, and (b) how to produce correct surface forms in target language and how to control the right word order.

2 Multi-word verb to single-word verb

First, we look at the verb group, where a multi-word verb maps to a single-word verb. This is the largest group, containing currently 387 verbs. The translation of the verb as such is not a problem, because normal rules can be used. A problem is that the second part of the verb in source language does not necessarily follow immediately the verb itself. Therefore, we must write such rules, which allow words between the verb and the second part. (The second part can contain more than one word.) The identification of a non-contiguous multi-word expression is not a problem. A problem is that it is difficult to control that the second part is not treated as normal word in the further phases of processing. If this is not done, the system adds a gloss also to the other part, although its semantic meaning is already in the new gloss of the multi-word verb.

Now we look at such structures, where the second member of the multi-word verb is a preposition.

2.1 The second member is a preposition

A preposition as the second member could also be implemented so that it is given the tag NOGLOSS in the verb structures. Then we would not need to treat the verb as a MWE. But there is a further problem. The verb complement in target language may have a different role than in the source language, and it is difficult to control such changes without MWE description.

Consider example (1).

```
(1)
(a) "<They>"
    "they" %SUBJ PRON NOM PL3
"<applied>"
    "apply" %+FMAINV V PAST *>MW { anoa V52 O-PAR , hakea V58-D
O-PAR } MW
"<for>"
    "for" %ADVL PREP
"<promotion>"
    "promotion" %<P N SG NOM
"<.>"
    "."
```

```
(b) "<They>"
    "they" %SUBJ PRON NOM PL3 CAPINIT
"<applied>"
    "apply" %+FMAINV V PAST *>MW { anoa V52 O-PAR , hakea V58-D
O-PAR } MW
"<again>"
    "again" %ADVL ADV
"<for>"
    "for" %ADVL PREP
"<promotion>"
    "promotion" %<P N SG NOM
"<.>"
    ". ."
(c) "<They>"
    "they" %SUBJ PRON NOM PL3 CAPINIT
"<applied>"
    "apply" %+FMAINV V PAST *>MW { anoa V52 O-PAR , hakea V58-D
O-PAR } MW
"<for>"
    "for" %ADVL PREP
"<promotion>"
    "promotion" %<P N SG NOM
"<again>"
    "again" %ADVL ADV
"<.>"
    ". "
```

Sentence (a) has the verb *apply* and its post complement (%<P) *promotion*, preceded with the preposition *for*. When this is translated into Finnish, the verb complement becomes an object, and it must have the appropriate object case. Object cases are partitive, genitive accusative and nominative accusative. It is very difficult to choose the correct object case without treating the verb and preposition as a MWE. The verb *apply* has several glosses in Finnish, but in this context it has only two glosses, which are semantically close to each other.

The isolated MWE has the tag **>MW*. This means that on the right there must be the second member of the MWE. The rule itself defines that the second member is *for*, but its precise location is not defined. The rule constraints it by stating, that it should not be beyond clause boundary, verb or preposition. When the rule is so formulated, the structure allows non-contiguous MWEs. Examples (b) and (c) show that the MWE can be contiguous or non-contiguous.

In normal MWE rules, the structure of the MWE is precisely defined using angle brackets *<* and *>* to show the location of the other members of the MWE. This code helps in writing rules for controlling that the rules, which add glosses to words, do not touch the words that are part of the MWE. For non-contiguous cases this method is not possible, because we cannot know exactly the place of the other member.

We can try the rule in (2) for replacing the analysis of the preposition with a MWE tag.

(2)

```
REPLACE ( { MW*< } ) TARGET (PREP) (1 (%<P)) (*-1 (*>MW) BARRIER  
CLBV OR (PREP) LINK 0 (PREP>));
```

The rule reads: Replace all tags in the reading with *PREP* with *MW*<*, if in the next cohort to the right there is the tag *%<P*. Somewhere on the left there should be the tag **>MW*, but do not scan beyond clause boundary or verb or the tag *PREP*. On the found reading there should be the tag *PREP>*.

We see in (3) that in all three examples, the preposition got the MWE tag.

(3)

```
"<They>"  
    "they" %SUBJ PRON NOM PL3  
"<applied>"  
    "apply" %+FMAINV MW V PAST *>MW { anoa V52 O-PAR , hakea  
V58-D O-PAR } PREP>  
"<for>"  
    "for" { MW*< }  
"<promotion>"  
    "promotion" %<P N SG NOM  
"<.>"  
    "."  
"<They>"  
    "they" %SUBJ PRON NOM PL3 CAPINIT  
"<applied>"  
    "apply" %+FMAINV MW V PAST *>MW { anoa V52 O-PAR , hakea  
V58-D O-PAR } PREP>  
"<again>"  
    "again" %ADVL ADV  
"<for>"  
    "for" { MW*< }  
"<promotion>"  
    "promotion" %<P N SG NOM  
"<.>"  
    "."  
"<They>"  
    "they" %SUBJ PRON NOM PL3 CAPINIT  
"<applied>"  
    "apply" %+FMAINV MW V PAST *>MW { anoa V52 O-PAR , hakea  
V58-D O-PAR } PREP>  
"<for>"  
    "for" { MW*< }  
"<promotion>"  
    "promotion" %<P N SG NOM  
"<again>"  
    "again" %ADVL ADV  
"<.>"  
    "."
```

Now when we add the Finnish glosses to the readings, we see that all such words, which have the MWE tag, will be without new glosses (4).

```
(4)
"<They>"
  "they" { he Np12 FRONT } %SUBJ PRON PL3 NOM
"<applied>"
  "apply" { anoa V52 } %+FMMAINV O-PAR MW V PAST *>MW PREP> PL
"<for>"
  "for" { MW*< }
"<promotion>"
  "promotion" { ylennys N39 FRONT } %<P N SG PAR
"<.>"
  "." { . }

"<They>"
  "they" { he Np12 FRONT } %SUBJ PRON PL3 CAPINIT NOM
"<applied>"
  "apply" { anoa V52 } %+FMMAINV O-PAR MW V PAST *>MW PREP> PL
"<again>"
  "again" { taas } %ADVL ADV
"<for>"
  "for" { MW*< }
"<promotion>"
  "promotion" { ylennys N39 FRONT } %<P N SG PAR
"<.>"
  "." { . }

"<They>"
  "they" { he Np12 FRONT } %SUBJ PRON PL3 CAPINIT NOM
"<applied>"
  "apply" { anoa V52 } %+FMMAINV O-PAR MW V PAST *>MW PREP> PL
"<for>"
  "for" { MW*< }
"<promotion>"
  "promotion" { ylennys N39 FRONT } %<P N SG PAR
"<again>"
  "again" { taas } %ADVL ADV
"<.>"
  "." { . }
```

The final translation (5) has all the words in place and in correct form.

He anoivat ylennystä.
He anoivat taas ylennystä.
He anoivat ylennystä taas.

We can test the approach by putting more words in between (6).

(6)
"<They>"
 "they" %SUBJ PRON NOM PL3
"<applied>"
 "apply" %+FMAINV MW V PAST *>MW { anoa V52 O-PAR , hakea
V58-D O-PAR } PREP>
"<again and again>"
 "again and again" %ADVL ADV BI
"<for>"
 "for" { MW*< }
"<promotion>"
 "promotion" %<P N SG NOM
"<.>"
 "."

We see that the word sequence *again and again* was already treated in morphological analysis as a MWE. Yet another non-contiguous MWE was identified. The translation is in (7).

(7)
He anoivat aina uudelleen ylennystä.

The verb *apply* has two glosses, *anoa* and *hakea*. So far we have not disambiguated this reading. We have just use the default gloss, that is, the first one. There are cases, when this is not appropriate (8).

(8)
"<They>"
 "they" { he Np12 FRONT } %SUBJ PRON NOM PL3 CAPINIT
"<applied>"
 "apply" { hakea V58-D } %+FMAINV O-PAR MW V PAST *>MW PREP>
"<for>"
 "for" { MW*< }
"<professorship>"
 "professorship" { professuuri N5 } %<P JOB NOPROP N NOM SG
"<.>"
 "." { . }

The translation is in (9)

(9)
He hakivat professuuria.

2.2 The second member is an adverb

If the second member is an adverb and it is located in an unknown distance on the right, it is not easy to control that it will not be translated. It should not, because it is already translated in the MWE (10).

(10)
"<He>"
 "he" %SUBJ CAPINIT PRON PERS NOM SG3
"<auctioned>"
 "auction" %+FMAINV MW V PAST *>MW { huutokaupata V73-B O-ACC
}
"<off>"
 "off" %ADVL ADV
"<.>"
 "."
"<He>"
 "he" %SUBJ CAPINIT PRON PERS NOM SG3
"<auctioned>"
 "auction" %+FMAINV MW V PAST *>MW { huutokaupata V73-B O-ACC
}
"<things>"
 "thing" %OBJ DEF N PL N-ING
"<off>"
 "off" %ADVL ADV
"<.>"
 "."
"<He>"
 "he" %SUBJ CAPINIT PRON PERS NOM SG3
"<auctioned>"
 "auction" %+FMAINV MW V PAST *>MW { huutokaupata V73-B O-ACC
}
"<all>"
 "all" %DN> DET
"<his>"
 "he" %A> PRON PERS GEN SG3
"<books>"
 "book" %OBJ DEF N PL
"<off>"
 "off" %ADVL ADV
"<.>"
 "."

We can consider a rule such as in (11).

(11)
REPLACE ({ MW*< }) TARGET (ADV) (*-1 (*>MW) BARRIER CLBV OR (ADV)
LINK 0 (ADV>));

The rule reads: If the reading has the tag *ADV*, replace it with { *MW*<* }, if somewhere on the left there is the tag **>MW*. However, do not scan beyond the clause boundary, or verb, or adverb. On the found reading there should be the tag *ADV>*.

In (12) there are examples of how the rule works.

(12)

```
"<He>"
    "he" %SUBJ CAPINIT PRON PERS NOM SG3
"<auctioned>"
    "auction" %+FMAINV MW V PAST *>MW { huutokaupata V73-B O-ACC
} ADV>
"<off>"
    "off" { MW*< }
"<.>"
    "."
"<He>"
    "he" %SUBJ CAPINIT PRON PERS NOM SG3
"<auctioned>"
    "auction" %+FMAINV MW V PAST *>MW { huutokaupata V73-B O-ACC
} ADV>
"<things>"
    "thing" %OBJ DEF N PL N-ING
"<off>"
    "off" { MW*< }
"<.>"
    "."
"<He>"
    "he" %SUBJ CAPINIT PRON PERS NOM SG3
"<auctioned>"
    "auction" %+FMAINV MW V PAST *>MW { huutokaupata V73-B O-ACC
} ADV>
"<all>"
    "all" %DN> DET
"<his>"
    "he" %A> PRON PERS GEN SG3
"<books>"
    "book" %OBJ DEF N PL
"<off>"
    "off" { MW*< }
"<.>"
    "."
```

Now when we add Finnish tags to readings, we see that the words, which are part of the MWE, are left intact (13)

(13)

```
"<He>"
    "he" { hän Np9 FRONT OUT } HUM MALE %SUBJ CAPINIT PRON PERS
NOM SG3
    "he" { hänen } HUM MALE %SUBJ CAPINIT PRON PERS NOM SG3
    "he" { NOGLOSS } HUM MALE %SUBJ CAPINIT PRON PERS NOM SG3
    "he" { itse N8 FRONT } HUM MALE %SUBJ CAPINIT PRON PERS NOM
SG3
    "he" { PROP-CAND } HUM MALE %SUBJ CAPINIT PRON PERS NOM SG3
"<auctioned>"
```

```
"auction" { huutokaupata V73-B O-ACC } %+FMAINV MW V PAST
*>MW ADV>
"<things>"
  "thing" { asia N12 } %OBJ DEF N PL N-ING
  "thing" { esine N48 FRONT } %OBJ DEF N PL N-ING
  "thing" { kama N9 } %OBJ DEF N PL N-ING
"<off>"
  "off" { MW*< }
"<.>"
  "." { . }
"<He>"
  "he" { hän Np9 FRONT OUT } HUM MALE %SUBJ CAPINIT PRON PERS
NOM SG3
  "he" { hänen } HUM MALE %SUBJ CAPINIT PRON PERS NOM SG3
  "he" { NOGLOSS } HUM MALE %SUBJ CAPINIT PRON PERS NOM SG3
  "he" { itse N8 FRONT } HUM MALE %SUBJ CAPINIT PRON PERS NOM
SG3
  "he" { PROP-CAND } HUM MALE %SUBJ CAPINIT PRON PERS NOM SG3
"<auctioned>"
  "auction" { huutokaupata V73-B O-ACC } %+FMAINV MW V PAST
*>MW ADV>
"<all>"
  "all" { kaikki N7-A } %DN> DET
"<his>"
  "he" { hän Np9 FRONT OUT } HUM MALE %A> PRON PERS GEN SG3
  "he" { hänen } HUM MALE %A> PRON PERS GEN SG3
  "he" { NOGLOSS } HUM MALE %A> PRON PERS GEN SG3
  "he" { itse N8 FRONT } HUM MALE %A> PRON PERS GEN SG3
"<books>"
  "book" { kirja N9 } %OBJ DEF N PL
"<off>"
  "off" { MW*< }
"<.>"
  "." { . }
```

After semantic disambiguation the result is as in (14)

```
(14)
"<He>"
  "he" { hän Np9 FRONT } %SUBJ OUT HUM MALE CAPINIT PRON PERS
NOM SG3
"<auctioned>"
  "auction" { huutokaupata V73-B } %+FMAINV O-ACC MW V PAST
*>MW ADV>
"<off>"
  "off" { MW*< }
"<.>"
  "." { . }
"<He>"
  "he" { hän Np9 FRONT } %SUBJ OUT HUM MALE CAPINIT PRON PERS
NOM SG3
```

```
"<auctioned>"
    "auction" { huutokaupata V73-B } %+FMAINV O-ACC MW V PAST
*>MW ADV>
"<things>"
    "thing" { esine N48 FRONT } %OBJ DEF N PL N-ING
"<off>"
    "off" { MW*< }
"<.>"
    "." { . }
"<He>"
    "he" { hän Np9 FRONT } %SUBJ OUT HUM MALE CAPINIT PRON PERS
NOM SG3
"<auctioned>"
    "auction" { huutokaupata V73-B } %+FMAINV O-ACC MW V PAST
*>MW ADV>
"<all>"
    "all" { kaikki N7-A } %DN> DET
"<his>"
    "he" { NOGLOSS } %A> HUM MALE PRON PERS GEN SG3
"<books>"
    "book" { kirja N9 } %OBJ DEF N PL
"<off>"
    "off" { MW*< }
"<.>"
    "." { . }
```

The correct translation can now be produced (15).

(15)

Hän huutokauppasi.

Hän huutokauppasi esineet.

Hän huutokauppasi kaikki kirjansa.

2.3 The second member is a noun

The second member of the multi-word verb can also be a noun. It is often an object, but not always. The examples in (16) demonstrate this.

(16)

```
"<He>"
    "he" %SUBJ CAPINIT PRON PERS NOM SG3
"<took>"
    "take" %+FMAINV V PAST *>MW { kotiuttaa V53-C O-ACC } N> MW
"<all>"
    "all" %DN> DET
"<profits>"
    "profit" %OBJ N PL
"<home>"
    "home" %ADVL N SG NOM
"<.>"
```

```
    "."
"<<s>>"
    "<s>"
"<plan>"
    "plan" %SUBJ DEF N SG NOM
"<took>"
    "take" %+FMAINV V PAST >MW { hahmottua V52-C , muotoutua
V52-F } OBJ> MW >MW
"<shape>"
    "shape" %OBJ INDEF N SG
"<.>"
    "."
"<tree>"
    "tree" %SUBJ DEF N SG NOM
"<took>"
    "take" %+FMAINV V PAST *>MW { juurtua V52-K O-ILL } OBJ> MW
"<root>"
    "root" %OBJ INDEF N SG
"<.>"
    "."
```

In order to be able to mark the second member as part of the MWE, the MWE definition has a specific tag to indicate the type of the second member. Such added tags above are *N>* and *OBJ>*.

The rules for controlling the second member are in (17).

```
(17)
REPLACE ({ MW*< }) TARGET OBJ (*-1 (\*>MW) BARRIER CLBV LINK 0
(OBJ>));
REPLACE ({ MW*< }) TARGET N (NOT 0 OBJ) (*-1 (\*>MW) BARRIER CLBV
LINK 0 (N>));
```

The first rule is for cases, where the target is an object. The second rule is for non-object cases.

The effect of the rules is demonstrated in (18).

```
(18)
"<He>"
    "he" %SUBJ CAPINIT PRON PERS NOM SG3
"<took>"
    "take" %+FMAINV MW V PAST *>MW { kotiuttaa V53-C O-ACC } N>
"<all>"
    "all" %DN> DET
"<profits>"
    "profit" %OBJ N PL
"<home>"
    "home" { MW*< }
"<.>"
    "."
"<plan>"
```

```
"plan" %SUBJ DEF N SG NOM
"<took>"
  "take" { hahmottua V52-C , muotoutua V52-F } %+FMAINV MW V
PAST >MW OBJ>
"<shape>"
  "shape" MW<
"<.>"
  "."
"<tree>"
  "tree" %SUBJ DEF N SG NOM
"<took>"
  "take" %+FMAINV MW V PAST *>MW { juurtua V52-K O-ILL } OBJ>
"<root>"
  "root" { MW*< }
"<.>"
  "."
```

After adding Finnish tags and after semantic disambiguation the result is in (19).

```
(19)
"<He>"
  "he" { hän Np9 FRONT OUT } HUM MALE %SUBJ CAPINIT PRON PERS
NOM SG3
"<took>"
  "take" { kotiuttaa V53-C O-ACC } %+FMAINV MW V PAST *>MW N>
"<all>"
  "all" { kaikki N7-A } %DN> DET
"<profits>"
  "profit" { voitto N1-C } %OBJ N PL
"<home>"
  "home" { MW*< }
"<.>"
  "." { . }
"<plan>"
  "plan" { suunnitelma N10 } %SUBJ DEF N NOM SG
"<took_shape>"
  "take_shape" { hahmottua V52-C } %+FMAINV MW MW V PAST OBJ>
"<.>"
  "." { . }
"<tree>"
  "tree" { puu N18 } %SUBJ DEF N NOM SG
"<took>"
  "take" { juurtua V52-K O-ILL } %+FMAINV MW V PAST *>MW OBJ>
"<root>"
  "root" { MW*< }
"<.>"
  "." { . }
```

We see that the readings have only those glosses, which we need for translation. The result is in (20).

(20)
Hän kotiutti kaikki voitot.
Suunnitelma hahmottui.
Puu juurtui.

2.4 The second member is an adjective

Most verbs, which have an adjective modifier, can be translated normally. There are, however, verbs, which in source language are MWEs, but in target language single-word verbs. The meaning of the verb modifier is already included into the verb gloss. Therefore, the modifier should not have a gloss.

In (21) are examples of this type of MW verbs.

(21)
"<He>"
 "he" %SUBJ CAPINIT PRON PERS NOM SG3
"<got>"
 "get" %+FMAINV MW V PAST *>MW { harmistua V52 } EN>
"<annoyed>"
 "annoy" %-FMAINV V EN
"<.>"
 "."
"<He>"
 "he" %SUBJ CAPINIT PRON PERS NOM SG3
"<got>"
 "get" %+FMAINV MW V PAST *>MW { hämmentyä V52-J FRONT } EN>
"<confused>"
 "confuse" %-FMAINV V EN
"<.>"
 "."
"<weather>"
 "weather" %SUBJ DEF N SG NOM
"<got>"
 "get" %+FMAINV MW V PAST *>MW { kirkastua V52 }
"<brighter>"
 "bright" %PCOMPL-S INDEF A CMP
"<.>"
 "."
"<publisher>"
 "publisher" %SUBJ DEF N SG NOM
"<made>"
 "make" %+FMAINV MW V PAST *>MW { julkistaa V53 O-ACC } A>
"<his>"
 "he" %A> PRON PERS GEN SG3
"<book>"
 "book" %OBJ DEF N SG
"<public>"
 "public" %<A A
"<.>"

". "

Note that in structures *got annoyed* and *got confused* the new interpretation was attached to the first verb. This was done, because this is the verb, which carries the information of verb inflection.

Also note that the MW description has a tag *A>* or *EN>*, indicating that this type of modifier should be marked as part of the MWE.

The rule, which does this, is in (22).

(22)

```
REPLACE ( { MW*< } ) TARGET A OR EN (*-1 (\*>MW) BARRIER CLBV LINK 0  
(A>) OR (EN>));
```

The rule reads: If there is a reading with the tag *A* or *EN*, replace the whole reading with *{ MW*< }*. Somewhere on the left there should be the tag **>MW*. You should not scan beyond the clause boundary or verb. On the found reading there should be the tag *A>* or *EN>*.

After applying this rule, the result is in (23).

(23)

```
"<He>"  
  "he" %SUBJ CAPINIT PRON PERS NOM SG3  
"<got>"  
  "get" %+FMAINV MW V PAST *>MW { harmistua V52 } EN>  
"<annoyed>"  
  "annoy" { MW*< }  
"<.>"  
  ". "  
"<He>"  
  "he" %SUBJ CAPINIT PRON PERS NOM SG3  
"<got>"  
  "get" %+FMAINV MW V PAST *>MW { hämmentyä V52-J FRONT } EN>  
"<confused>"  
  "confuse" { MW*< }  
"<.>"  
  ". "  
"<weather>"  
  "weather" %SUBJ DEF N SG NOM  
"<got>"  
  "get" %+FMAINV MW V PAST *>MW { kirkastua V52 } A>  
"<brighter>"  
  "bright" { MW*< }  
"<.>"  
  ". "  
"<publisher>"  
  "publisher" %SUBJ DEF N SG NOM  
"<made>"  
  "make" %+FMAINV MW V PAST *>MW { julkistaa V53 O-ACC } A>
```

```
"<his>"
    "he" %A> PRON PERS GEN SG3
"<book>"
    "book" %OBJ DEF N SG
"<public>"
    "public" { MW*< }
"<.>"
    ". "
```

The sentences can now be translated (24).

(24)
Hän harmistui.
Hän hämmentyi.
Sää kirkastui.
Kustantaja julkisti hänen kirjansa.

Let us test what happens, if the second member is not immediately after the verb (25).

```
(25)
"<He>"
    "he" %SUBJ CAPINIT PRON PERS NOM SG3
"<got>"
    "get" %+FMAINV MW V PAST *>MW { harmistua V52 } EN>
"<constantly>"
    "constantly" %ADVL ADV
"<annoyed>"
    "annoy" { MW*< }
"<.>"
    ". "
"<He>"
    "he" %SUBJ CAPINIT PRON PERS NOM SG3
"<got>"
    "get" %+FMAINV MW V PAST *>MW { hämmentyä V52-J FRONT } EN>
"<constantly>"
    "constantly" %ADVL ADV
"<confused>"
    "confuse" { MW*< }
"<.>"
    ". "
"<weather>"
    "weather" %SUBJ DEF N SG NOM
"<got>"
    "get" %+FMAINV MW V PAST *>MW { kirkastua V52 } A>
"<repeatedly>"
    "repeatedly" %ADVL ADV
"<brighter>"
    "bright" { MW*< }
"<.>"
    ". "
```



```
"<publisher>"
    "publisher" %SUBJ DEF N SG NOM
"<made>"
    "make" %+FMAINV MW V PAST *>MW { julkistaa V53 O-ACC } A>
"<his>"
    "he" %A> PRON PERS GEN SG3
"<book>"
    "book" %OBJ DEF N SG
"<public>"
    "public" { MW*< }
"<.>"
    "."
```

We see that the second member was correctly marked as part of the MWE, and other words got the Finnish gloss.

The translation is in (26).

(26)

Hän harmistui jatkuvasti.

Hän hämmentyi jatkuvasti.

Sää kirkastui toistuvasti.

Kustantaja julkisti hänen kirjansa.

3 The verb is a MWE in source language and target language

There are cases, where the verb is a MWE on both sides. Yet they cannot be translated as separate words. On the source side, the cluster of words is isolated as a MWE, and on the target side it is given a gloss, which also is a MWE.

3.1 MWE for avoiding unnecessary ambiguity

There are cases, where the structure could be translated directly without resorting to the MW construction, but where the ambiguity of both words would require disambiguation, which in any case would be unsecure. When the word pair is defined as a MWE, no disambiguation is needed.

Consider the example in (27), where no MWE definition is used.

(27)

```
"<They>"
    "they" { he Np10 FRONT OUT HUM } %SUBJ PRON NOM PL3
    "they" { ne Np12 FRONT } %SUBJ PRON NOM PL3
    "they" { NOGLOSS } %SUBJ PRON NOM PL3
    "they" { itse N8 FRONT } %SUBJ PRON NOM PL3
    "they" { niiden } %SUBJ PRON NOM PL3
    "they" { heidän HUM } %SUBJ PRON NOM PL3
"<fought>"
    "fight" { taistella V67 } { vastaan+POST } %+FMAINV V PAST
    "fight" { tapella V67-B } { vastaan+POST } %+FMAINV V PAST
```

```
"fight" { taistella V67 O-PAR } { vastaan+POST } %+FMAINV V
PAST
"<fire>"
  "fire" { tuli N23 } %OBJ INDEF N SG
  "fire" { tulipalo N1 } %OBJ INDEF N SG
  "fire" { metsäpalo N1 } %OBJ INDEF N SG
  "fire" { palo N1 } %OBJ INDEF N SG
"<.>"
  "." { . }
```

The sentence has a lot of ambiguity. Now, when we use the MWE definition, we get a much simpler reading (28).

```
(28)
"<They>"
  "they" { he Np10 FRONT OUT HUM } %SUBJ PRON NOM PL3
  "they" { ne Np12 FRONT } %SUBJ PRON NOM PL3
  "they" { NOGLOSS } %SUBJ PRON NOM PL3
  "they" { itse N8 FRONT } %SUBJ PRON NOM PL3
  "they" { niiden } %SUBJ PRON NOM PL3
  "they" { heidän HUM } %SUBJ PRON NOM PL3
"<fought>"
  "fight" { taistella V67 || tulta vastaan } %+FMAINV MW V
PAST *>MW
"<fire>"
  "fire" { MW*< }
"<.>"
  "." { . }
```

The second part of the multi-word verb is moved to a separate line (29).

```
(29)
"<They>"
  "they" { he Np10 FRONT } %SUBJ OUT HUM PRON NOM PL3
"<fought>"
  "fight" { taistella V67 } %+FMAINV MW V PAST *>MW
"<tulta vastaan>"
  "tulta vastaan" { +tulta vastaan } X
"<fire>"
  "fire" { MW*< }
"<.>"
  "." { . }
```

Now we can translate it (30).

```
(30)
He taistelivat tulta vastaan.
```

3.2 Various types of MW verbs and their implementation

There are several types of MW verbs with a number of problems in translation. One thing is the correct handling of the MW verb members. Another problem is the handling of postmodifiers of the verb, when the default handling is not appropriate. Below we discuss a number of such cases.

Let us see what happens when the verb constituents are not contiguous (31).

```
(31)
"<I>"
    "i" %SUBJ PRON PERS NOM SG1
"<knew>"
    "know" %+FMAINV MW V PAST *>MW { tuntea V59-J TRV || nimellä
} AS>
"<him>"
    "he" %OBJ PRON PERS SG3
"<as>"
    "as" { MW*< }
"<Kalle>"
    "kalle" %<P CAP N SG NOM
"<.>"
    "."
```

We see that the MW verb *know* was glossed as *tuntea nimeltä*. Both constituents were marked as part of the MWE. Now the Finnish glosses can be added (32).

```
(32)
"<I>"
    "i" { minä Np5 } HUM OUT %SUBJ PRON PERS NOM SG1
    "i" { NOGLOSS } HUM OUT %SUBJ PRON PERS NOM SG1
    "i" { itse N8 FRONT } HUM OUT %SUBJ PRON PERS NOM SG1
"<knew>"
    "know" { tuntea V59-J TRV || nimellä } %+FMAINV MW V PAST
*>MW AS>
"<him>"
    "he" { hän Np9 FRONT OUT } HUM MALE %OBJ PRON PERS SG3
    "he" { hänen } HUM MALE %OBJ PRON PERS SG3
    "he" { NOGLOSS } HUM MALE %OBJ PRON PERS SG3
    "he" { itse N8 FRONT } HUM MALE %OBJ PRON PERS SG3
"<as>"
    "as" { MW*< }
"<Kalle>"
    "kalle" { kalle N8 } HUM %<P CAP N NOM SG
    "kalle" { PROP-CAND } HUM %<P CAP N NOM SG
"<.>"
    "." { . }
```

Semantic disambiguation removes wrong readings (33).

(33)
"<I>"
 "i" { minä Np5 } %SUBJ HUM OUT PRON PERS NOM SG1
"<knew>"
 "know" { tuntea V59-J } %+FMAINV TRV MW V PAST *>MW AS>
"<nimellä>"
 "nimellä" { +nimellä } X
"<him>"
 "he" { hän Np9 FRONT } %OBJ OUT HUM MALE PRON PERS SG3
"<as>"
 "as" { MW*< }
"<Kalle>"
 "kalle" { kalle N8 } %<P HUM CAP N NOM SG
"<.>"
 "." { . }

The translation is in (34).

(34)
Minä tunsin hänet nimellä Kalle.

There are also such MW verbs, which have three constituents, such as *take into account*. The verb can be detached from the other two words. In (35) we have an example of this.

(35)
"<He>"
 "he" %SUBJ CAPINIT PRON PERS NOM SG3
"<took>"
 "take" %+FMAINV MW V PAST *>MW { ottaa V53-C || huomioon }
%<P>
"<all>"
 "all" %DN> DET
"<complaints>"
 "complaint" %OBJ N PL
"<into>"
 "into" %ADVL PREP
"<account>"
 "account" { MW*< }
"<.>"
 "."

Note that only the word *account* was marked as a member of MWE. The preposition *into* does not need marking, because it itself has no word gloss, and therefore cannot disturb translation.

The rule for finding and marking this MWE is in (36).

(36)
ADD (*>MW { ottaa V53-C || huomioon } %<P> MW) TARGET ("take") +
V (*1 ("into") BARRIER CLBV LINK 1 ("account")) (NOT 0 MW);

The rule reads: Add the string **>MW { ottaa V53-C || huomioon } %<P> MW* to the reading, which has the tag "take" plus *V*. Also, somewhere on the right there should be the tag "into", and immediately after it the tag "account". Do not scan beyond the clause boundary or verb.

The rule for marking the third member of the MWE is in (37).

(37)
REPLACE ({ MW*< }) TARGET (%<P> (*-1 (*>MW) BARRIER CLBV LINK 0 (%<P>)));

The rule reads: If there is the tag *%<P>* on the reading, replace the whole reading with *{ MW*< }*. Somewhere on the left there should be the tag **>MW*, and also the tag *%<P>*. Do not scan beyond the clause boundary of verb.

We get the surface forms as in (38).

(38)
"<He>"
"he" { h:än } %SUBJ OUT HUM MALE CAPINIT PRON PERS SG3 NOM
"<took>"
"take" { ott+i } %+FMMAINV MW V PAST *>MW %<P> SG
"<huomioon>"
"huomioon" { +huomioon } X PL
"<all>"
"all" { kaik+ki } %DN> DET PL NOM
"<complaints>"
"complaint" { valitu+kset } %OBJ N PL NOM
"<into>"
"into" { NOGLOSS M-ILL } %ADVL PREP
"<account>"
"account" { MW*< }
"<.>"
"." { . }

When the word order is changed, we get the translation (39).

(39)
Hän otti kaikki valitukset huomioon.

It is possible that more than one word in the MW verb construction must be marked as part of MWE. Consider the example in (40).

(40)
"<He>"
"he" %SUBJ CAPINIT PRON PERS NOM SG3
"<was>"
"be" %+FAUXV V PAST SG *>MW { altistua V52 O-ALL , paljastua V52 } EN> PREP> MW

```
"<exposed>"
  "expose" %-FMAINV V EN
"<to>"
  "to" %ADVL PREP
"<light>"
  "light" %<P N SG NOM
"<.>"
  "."
```

Here the MW verb is *be exposed*. The definition is bound to the auxiliary verb *be*, because it carries the information on inflection. Now there are the words *exposed* and *to*, both of which have their default interpretation, which now should be blocked. The MWE definition has such tags as *EN>* and *PREP>*. These tags trigger the rules, which remove the original readings from those two words, as we see in (41).

```
(41)
"<He>"
  "he" %SUBJ CAPINIT PRON PERS NOM SG3
"<was>"
  "be" %+FAUXV MW V PAST SG *>MW { altistua V52 O-ALL ,
paljastua V52 } EN> PREP>
"<exposed>"
  "expose" { MW*< }
"<to>"
  "to" { MW*< }
"<light>"
  "light" %<P N SG NOM
"<.>"
  "."
```

We see that the case tag for the word *light* is *NOM* by default. The preposition *to* would convert it too *ILL*, again by default. But now when we have removed all instructions of the preposition *to*, we can use the information in the MW verb. There we have the tag *O-ALL*, which helps in assigning the tag *ALL* to the word *light*, as we see in (42).

```
(42)
"<He>"
  "he" { hän Np9 FRONT } %SUBJ OUT HUM MALE CAPINIT PRON PERS
SG3 NOM
"<was>"
  "be" { altistua V52 } %+FAUXV M-ALL MW V PAST *>MW EN> PREP>
SG
"<exposed>"
  "expose" { MW*< }
"<to>"
  "to" { MW*< }
"<light>"
  "light" { valo N1 } %<P N SG ALL
"<.>"
  "." { . }
```

The correct translation can now be produced (43).

(43)
Hän altistui valolle.

Another example of a complex reformulation of the clause is in (44).

(44)
"<He>"
 "he" %SUBJ CAPINIT PRON PERS NOM SG3
"<warned>"
 "warn" %+FMAINV V PAST *>MW { varoittaa V53-C O-PAR P-ELA }
PREP> MW
"<against>"
 "against" %ADVL PREP
"<danger>"
 "danger" %<P N SG NOM
"<.>"
 "."
"<He>"
 "he" %SUBJ CAPINIT PRON PERS NOM SG3
"<warned>"
 "warn" %+FMAINV V PAST *>MW { varoittaa V53-C O-PAR P-ELA }
PREP> MW
"<them>"
 "they" %OBJ PRON PERS PL3
"<against>"
 "against" %ADVL PREP
"<danger>"
 "danger" %<P N SG NOM
"<.>"
 "."

The rule for marking the MWE is in (45)

(45)
ADD (*>MW { varoittaa V53-C O-PAR P-ELA } PREP> MW) TARGET
("warn") + V (*1 ("against") BARRIER CLB) (NOT 0 MW);

The rule adds to the reading, which has the tag "warn" (it also should be a verb) The string **>MW { varoittaa V53-C O-PAR P-ELA } PREP> MW*, if somewhere on the right there is the tag "against". Do not scan beyond clause boundary or verb. The reading should not have been marked as MWE by an earlier rule.

The object of the verb should be in partitive case (*O-PAR*). The postmodifier of the verb should be in relative case (*P-ELA*).

There is also the tag *PREP>*, which helps later in marking the preposition as part of the MWE.

In (46) we mark the other member of the MWE.

(46)
"<He>"
"he" %SUBJ CAPINIT PRON PERS NOM SG3
"<warned>"
"warn" %+FMAINV MW V PAST *>MW { varoittaa V53-C O-PAR P-ELA
} PREP>
"<against>"
"against" { MW*< }
"<danger>"
"danger" %<P N SG NOM
"<.>"
" ."
"<He>"
"he" %SUBJ CAPINIT PRON PERS NOM SG3
"<warned>"
"warn" %+FMAINV MW V PAST *>MW { varoittaa V53-C O-PAR P-ELA
} PREP>
"<them>"
"they" %OBJ PRON PERS PL3
"<against>"
"against" { MW*< }
"<danger>"
"danger" %<P N SG NOM
"<.>"
" ."

The preposition *against* has now been marked as part of MWE, and it cannot get a gloss. It also has lost its default inflection instructions. When it is now neutral, the post-modifier can get the inflection code from the verb (*P-ELA*).

The inflection codes are added in (47).

(47)
"<He>"
"he" { hän Np9 FRONT } %SUBJ OUT HUM MALE CAPINIT PRON PERS
SG3 NOM
"<warned>"
"warn" { varoittaa V53-C } %+FMAINV O-PAR P-ELA MW V PAST
*>MW PREP> SG
"<against>"
"against" { MW*< }
"<danger>"
"danger" { vaara N9 } %<P N SG ELA
"<.>"
" ." { . }
"<He>"
"he" { hän Np9 FRONT } %SUBJ OUT HUM MALE CAPINIT PRON PERS
SG3 NOM
"<warned>"


```
"warn" { varoittaa V53-C } %+FMAINV O-PAR P-ELA MW V PAST
*>MW PREP> SG
"<them>"
  "they" { he Np10 FRONT } %OBJ OUT HUM PRON PERS PL3 PAR
"<against>"
  "against" { MW*< }
"<danger>"
  "danger" { vaara N9 } %<P N SG ELA
"<.>"
  "." { . }
```

Now we can translate the sentences (48).

(48)
Hän varoitti vaarasta.
Hän varoitti heitä vaarasta.

5 Discussion

Above I have described various types of multi-word verbs and demonstrated methods for handling them, so that correct translation can be produced. The isolation rules were formulated in such a way that the second member of the MWE can be either immediately after the verb, or in a non-defined distance, yet restricted by the clause boundary or verb.

The marking of the second member is implemented so, that the MWE reading contains one or more codes, such as *ADV>*, *PREP>*, *EN>* etc., to show that if there is an adverb, preposition or past participle somewhere between the verb and clause boundary or verb, it should be marked as part of MWE, thus removing the possibility that a reading with a MWE code would be treated as normal word with Finnish gloss.

There are such cases, for example, where two adverbs follow each other, and only one of them is part of MWE. An example is in (50).

```
(50)
"<He>"
  "he" %SUBJ CAPINIT PRON PERS NOM SG3
"<locked>"
  "lock" %+FMAINV MW V PAST *>MW { lukita V69 TRV || sisälle }
ADV>
"<dog>"
  "dog" %OBJ DEF N SG
"<in>"
  "in" { MW*< }
"<.>"
  "."
"<He>"
  "he" %SUBJ CAPINIT PRON PERS NOM SG3
"<locked>"
  "lock" %+FMAINV MW V PAST *>MW { lukita V69 TRV || sisälle }
ADV>
"<dog>"
```

```
"dog" %OBJ DEF N SG
"<in>"
  "in" { MW*< }
"<again>"
  "again" { MW*< }
"<.>"
  "."
"<He>"
  "he" %SUBJ CAPINIT PRON PERS NOM SG3
"<locked>"
  "lock" %+FMAINV MW V PAST *>MW { lukita V69 TRV || sisälle }
ADV>
"<dog>"
  "dog" %OBJ DEF N SG
"<again>"
  "again" { MW*< }
"<in>"
  "in" { MW*< }
"<.>"
  "."
```

We see that the code *ADV>* triggers the rule to replace the original reading of the adverbs *again* and *in* with a MWE code, although the adverb *again* should be left intact.

This can be implemented so that we restrict the rule to apply only to such adverbs, which appear as part of a multi-word verb. We can group such adverbs as a set, and then use the set name in the rule.

The reformulated marking rule for this case is in (51).

(51)
REPLACE ({ MW*< }) TARGET ADV-MW (NOT 1 N) (*-1 (*>MW) BARRIER
CLBV LINK 0 (ADV>));

The rule says: If the reading has a tag, which belongs to the set *ADV-MW*, replace the original reading with *{ MW*< }*. The following word should not be a noun. Somewhere on the left, there should be the tag **>MW*, and on the same line there should also be the tag *ADV>*. Do not scan beyond the clause boundary or verb.

When we use this rule, we get the result as in (52).

```
(52)
"<He>"
  "he" %SUBJ CAPINIT PRON PERS NOM SG3
"<locked>"
  "lock" %+FMAINV MW V PAST *>MW { lukita V69 TRV || sisälle }
ADV>
"<dog>"
  "dog" %OBJ DEF N SG
"<in>"
  "in" { MW*< }
"<.>"
```

```
    "."
"<He>"
    "he" %SUBJ CAPINIT PRON PERS NOM SG3
"<locked>"
    "lock" %+FMAINV MW V PAST *>MW { lukita V69 TRV || sisälle }
ADV>
"<dog>"
    "dog" %OBJ DEF N SG
"<in>"
    "in" { MW*< }
"<again>"
    "again" %ADVL ADV
"<.>"
    "."
"<He>"
    "he" %SUBJ CAPINIT PRON PERS NOM SG3
"<locked>"
    "lock" %+FMAINV MW V PAST *>MW { lukita V69 TRV || sisälle }
ADV>
"<dog>"
    "dog" %OBJ DEF N SG
"<again>"
    "again" %ADVL ADV
"<in>"
    "in" { MW*< }
"<.>"
    "."
```

Now the adverb *again* is left intact, and we can translate the sentences (53).

(53)
Hän lukitsi koiran sisälle.
Hän lukitsi koiran sisälle taas.
Hän lukitsi koiran sisälle taas.

The same method can be used also to other cases, where there is a dancer that the wrong word will be marked as part of MWE.

6 Summary

In the report I have demonstrated that all types of multi-word verbs can be handled and translated from English to Finnish. Only a few examples were used for demonstrating the methods. However, all so far known multi-word verbs were implemented and their correct behaviour tested. No nonresolvable problems were encountered.