Technical Reports on Language Technology
Report No 56, 2020
http://www.njas.helsinki.fi/salama

Optimizing the description of multi-word expressions in English[1]

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

Abstract

The description of multi-word expressions (MWE) is a necessary phase in rule-based machine translation. Because the concept MWE contains several types of word clusters, it is not self-evident how they should be described. One approach is that the isolation of multi-words is carried out after the morphological analysis, but before disambiguation. If the POS ambiguity of the language is minimal, this method is suitable, and perhaps also optimal. In case the POS ambiguity of the language is extensive, this method is hardly optimal. English belongs to this type of languages. The more optimal solution is that the isolation of MWEs is carried out in two phases. This method will be discussed and demonstrated in this report.

Key Words: *multi-word expressions, morphological analysis.*

1 Introduction

Multi-word expressions are clusters of two or more words, which together carry a meaning. The members of the MWEs are often consecutive. Also non-consecutive MWEs exist. Some MWEs are frozen. That is, they do not inflect. Other types of MWEs inflect. Depending on the language type, each member of the cluster may inflect, often having even tens of forms.

The basic approach to isolate the MWEs is such, that first the text is analysed morphologically. Then MWEs are isolated, and the disambiguation comes after it. In case the language has only limited POS ambiguity, this method is suitable.

English is morphologically simple, and as such an atypical language. The morphological simplicity has the backside that words are often ambiguous, also in relation to POS. In English, the word-form is very often a verb, a noun, or an adjective. If the isolation of MWEs is left to the phase after morphological analysis, it cannot be done immediately after analysis. The text must be disambiguated first, and only then the isolation of the MWEs can be done.

The method is prone to errors, because if the disambiguation is not accurate, the isolation of the MWEs fails. This danger can be reduced significantly, if we divide the task into two phases.

In the first phase, we isolate part of the MWEs already before analysis. This group of MWEs includes frozen MWEs, and such inflecting MWEs, which can then be described in the morphological analyser as a single unit.

In the second phase, we isolate the MWEs after morphological analysis and disambiguation. These are such cases, where the context must first be checked, before the isolation operation can be performed. The isolation of the MWEs in this phase is complex, because several constraints must be taken into consideration. Such constraints include the immediate context on the left and right, the POS category of one or more members, the identification of the inflecting member, and the possibility that one member can be distantly located.

The method proposed here for a language such as English has many advantages. The majority of MWEs can be isolated directly before morphological analysis.

2. Isolation of MWEs before morphological analysis

When we isolate MWEs before morphological analysis, we must do it in two phases. First, we must mark as single unit those word clusters, which constitute a MWE. For example, the sequence *in front of* is converted to *in_front_of*. This is done after tokenisation, so that each word is clearly separated from punctuation marks and diacritics. Also, the capital letters are converted into such bigraphs, where each capital letter is represented by an asterisk '*' followed by the corresponding lower case letter. For example, A is converted to *a. This is done, so that we can simplify the morphological lexicon without losing the information about capital letters in original text.

In (1) is an example of a sentence, where MWEs are isolated.

(1)
```
*the *prime_*minister stands in_front_of the *state_*house .
```

An underscore is used for keeping the members of the MWE together. The vertical format of the sentence is in (2).

(2)
```
*the
*prime_*minister
stands
in_front_of
the
*state_*house
```

We see that each MWE constitutes technically a single word. When we analyse this sentence, we must also cope with MWEs in some way. One method would be to use a guesser without explicitly including the MWEs also into the lexicon. Such a guesser would work quite reliably in cases such as proper names, but with other POS categories it would be very unreliable.

Therefore, it is advisable that the MWEs are included also into the lexicon, where they will be located in correct places. Ambiguous MWEs will be listed in more than one place. The analysed sentence is in (3).

(3)
```
"<*the>"
      "the" DET CAP DEF
"<*prime_*minister>"
      "prime_*minister" N CAP SG
"<stands>"
      "stand" V PRES SG3
      "stand" N PL
"<in_front_of>"
      "in_front_of" PREP
"<the>"
      "the" DET DEF
"<*state_*house>"
      "state_*house" PROPN CAP Heur
"<.>"
      "." **CLB
```

Note that the cluster **state_*house* was marked as a MWE in the pre-analysis phase, but it was not included into the lexicon. The guesser interpreted it correctly as a proper name. The other two MWEs were listed in the lexicon, and they were given correct interpretation.

3. Managing the inflection of MWEs

If the MWE does not inflect and it is not dependent on the context, it can be safely described with this method. Inflecting MWEs are more problematic, because, for example, the inflecting noun may have three forms, singular, plural, and the genitive form. The problems are different in marking and in analysing.

In the marking phase, the marking of all three forms can be done with one rule. The left context is the word boundary or asterisk, and the right boundary is a word boundary, or s, or apostrophe. Using these boundary definitions, we can catch the base form, the inflected forms, and also words starting either with asterisk or without. Problematic are such cases, where the inflecting member is not the last one, such as *member of parliament*, where the first member inflects. In such cases, singular and plural forms must be listed separately in rules.

The marking of the MWEs can be done with any suitable method. I have implemented the process using two alternative methods, both of which do the job.

In Perl implementation, the rule types are as in (4).

(4)
```
s/(administrative) (officer)(s|'s)? /$1_$2$3 /gm;
s/(ahead) (of) /$1_$2 /gm;
s/(anybody) (else) /$1_$2 /gm;
```

The first rule catches all three forms of the MWE. The two other ones are frozen and need no alternative forms.

In Beta implementation, the same rules look like in (5).

(5)
```
administrative officer; administrative_officer;
ahead of; ahead_of;
anybody else; anybody_else;
```

In addition, the left and right contexts must be defined accordingly. On the left, the boundary is word boundary or asterisk. On the right, the boundary is word boundary, or s, or '. Also Beta rules catch all needed forms.

In the morphological lexicon, the problems are quite much the same as in the marking phase. All three forms of the multi-word noun, as well as the forms starting with asterisk, can be described using only one lexical entry. Also the base form can be described with this rule. Exceptions are such MWEs, where the first member inflects. Single forms and plural forms must be written as separate entries, and the base form of the plural must be specified separately. (In normal usage, the equal sign '=' is used for marking the stem, if it is the same as the entry.)

The implementation of the above three cases in the morphological lexicon is as in (6).

(6)
```
administrative_officer N "= ";
ahead_of # "= ";
anybody_else # "= ";
```

The first rule is in the lexicon of nouns, and it allows all three forms of the noun. The other two are frozen forms and their execution is terminated instantly. Each of the rules is located in their own sub-lexicons, depending of the POS category.

We test their function by placing each of the MWEs into the same sentence (7).

(7)
```
"<*administrative_officers>"
      "administrative_officer" N CAP PL
"<are>"
      "be" AUXV PRES
"<ahead_of>"
      "ahead_of" PREP
"<time>"
      "time" V vt INF
      "time" V vt IMP
      "time" V vt PRES SG1
      "time" V vt PRES SG2/PL2
      "time" V vt PRES PL1
      "time" V vt PRES PL3
      "time" N PREFR SG
"<more>"
      "many" PRON CMP
      "more" ADV
      "much" ADV CMP
```

```
"<than>"
      "than" CONJ CS
      "than" PREP
"<anybody_else>"
      "anybody_else" PRON
"<.>"
      "." **CLB
```

The first MWE is a noun in plural form. The second one is a preposition. And the third one is a pronoun. When the MWEs are described in this way, they have seldom ambiguity. Also the inflected forms can be described effectively, including genitive, as the example in (8) shows.

(8)

```
"<*the>"
      "the" DET CAP DEF
"<administrative_officer's>"
      "administrative_officer" N SG GEN
"<work>"
      "work" V vt vi INF
      "work" V vt vi IMP
      "work" V vt vi PRES SG1
      "work" V vt vi PRES SG2/PL2
      "work" V vt vi PRES PL1
      "work" V vt vi PRES PL3
      "work" N SG
"<is>"
      "be" AUXV PRES SG3
"<stressful>"
      "stressful" A
"<.>"
      "." **CLB
```

When we disambiguate the sentences in (7) and (8), we get the result as in (9).

(9)

```
"<*administrative_officers>"
      "administrative_officer" N CAP PL
"<are>"
      "be" AUXV PRES
"<ahead_of>"
      "ahead_of" PREP
"<time>"
      "time" N PREFR SG
"<more>"
      "much" ADV CMP
"<than>"
      "than" PREP
"<anybody_else>"
      "anybody_else" PRON
```

```
"<.>"
      "." **CLB
"<*the>"
      "the" DET CAP DEF
"<administrative_officer's>"
      "administrative_officer" N SG GEN
"<work>"
      "work" N SG
"<is>"
      "be" AUXV PRES SG3
"<stressful>"
      "stressful" A
"<.>"
      "." **CLB
```

4. Isolation of MWEs after morphological analysis

It would be ideal that we could isolate MWEs in one place only. This method would give us a good control of the already isolated MWEs, and harmful double isolation could be avoided. In a language such as English, this is hardly the ideal method. Because of the abundant POS ambiguity, there is strong motivation for pre-analysis isolation. However, part of MWEs cannot be isolated in that position, because context control is there not yet available. All such MWEs, which require context control, must be isolated after analysis. This is the case also with non-consecutive MWEs.

Here we get two criteria for deciding, to which group each MWE belongs. However, the borderline is not strict. Several MWEs can be isolated with either method. One should be careful, that the isolation is done only once.

It is also characteristic to the isolation of MWEs on this point that the decision on whether a cluster of words should be interpreted as a MWE is not self-evident. Whereas the MWEs isolated in the pre-analysis phase are clear cases, the situation is more complex here. At least two criteria must be taken into consideration. First, we should check whether the word cluster is a MWE in this particular context. Second, we must take into consideration the target language. If we, for example, isolate MWEs, keeping in mind translation into Swahili, the group of MWEs will be different compared with isolation for translation into Finnish. In other words, the group of MWEs is not fixed. There are also many cases, where translation can be done using either way, by isolating as a MWE, or by translating individual words.

Compare the two sentences in (10).

(10)
```
He will come this week.
This week will be very rainy.
```

Both sentences contain the word cluster *this week*. When we proceed in translation towards Swahili, one intermediate phase is as in (11).

(11)
```
"<He>"
```

```
      "he" { NOGLOSS } MALE %SUBJ CAPINIT PRON PERS NOM SG3
"<will>"
      "will" { FUT } %+FAUXV ACR V AUXMOD
"<come>"
      "come" { INFMARK+jA } MONOSLB %-FMAINV V INF
"<this>"
      "this" { h } %DN> DET DEM SG
"<week>"
      "week" { 9SG wiki } %ADVL N SG NOM
"<.>"
      "." { . }

"<This>"
      "this" { h } %DN> CAPINIT DET DEM SG
"<week>"
      "week" { 9SG wiki } %SUBJ N SG NOM
"<will>"
      "will" { FUT } %+FAUXV ACR V AUXMOD
"<be>"
      "be" { INFMARK+wA } MONOSLB %-FMAINV V INF
"<very_rainy>"
       "very_rainy" { -enye mvua nyingi } A-MW %PCOMPL-S MW A ABS
A
"<.>"
      "." { . }
```

We see that the gloss of the demonstrative pronoun is *h*, and of the week is *9SG wiki*. We further note that the cluster *very rainy* was isolated as a MWE.

The final translation is in (12).

(12)

```
Atakuja wiki hii.
Wiki hii itakuwa yenye mvua nyingi.
```

The word cluster *wiki hii* is identical in both sentences. We can conclude that there is no need to isolate this word cluster as a MWE.

Now we translate the same sentences into Finnish. One intermediate phase of translation is displayed in (13).

(13)

```
"<He>"
      "he" { hän Np9 FRONT } %SUBJ OUT HUM MALE CAPINIT PRON PERS
NOM SG3
"<will>"
      "will" { NOGLOSS } %+FAUXV V AUXMOD
"<come>"
      "come" { tulla V67 } %-FMAINV O-LOC3 MOVE V INF
"<this_week>"
      "this_week" { tällä viikolla } %ADVL MW N NOM SG
"<.>"
```

```
      "." { . }
"<This>"
      "this" { tämä Np1 FRONT } %DN> CAPINIT DET DEM SG
"<week>"
      "week" { viikko N1-A } %SUBJ TIME N NOM SG
"<will>"
      "will" { NOGLOSS } %+FAUXV V AUXMOD
"<be>"
      "be" { olla V67b } %-FMAINV V-3INF-ILL O-LOC1 V INF
"<very>"
      "very" { hyvin } %AD-A> ADV
"<rainy>"
      "rainy" { sateinen N38 } %PCOMPL-S NEN INDEF A ABS
"<.>"
      "." { . }
```

We see that in the first sentence the cluster *this week* was isolated as a MWE, and in the second sentence not. In the first sentence, the Finnish gloss was directly written as *tällä viikolla*, which is adessive singular. In the second sentence the words were glossed separately in nominative form, provided with inflection class codes. When we add the particular inflection codes, the result is as in (14).

(14)

```
"<He>"
      "he" { h:än Np9 FRONT } %SUBJ OUT HUM MALE CAPINIT PRON PERS
SG3 NOM
"<will>"
      "will" { NOGLOSS } %+FAUXV V AUXMOD SG PRES
"<come>"
      "come" { tul:la V67 } %-FMAINV O-LOC3 MOVE V SG PRES
"<this_week>"
      "this_week" { tällä viikolla } %ADVL MW MW N SG NOM
"<.>"
      "." { . }
"<This>"
      "this" { tä:mä Np1 FRONT } %DN> CAPINIT DET DEM SG NOM
"<week>"
      "week" { viikk:o N1-A } %SUBJ TIME N SG NOM
"<will>"
      "will" { NOGLOSS } %+FAUXV V AUXMOD SG PRES
"<be>"
      "be" { o:lla V67b } %-FMAINV V-3INF-ILL O-LOC1 V SG PRES
"<very>"
      "very" { hyvin } %AD-A> ADV
"<rainy>"
      "rainy" { satei:nen N38 } %PCOMPL-S NEN INDEF A ABS SG NOM
"<.>"
      "." { . }
```

Now we have all the information needed, so that we can make the final translation (15).

(15)
Hän tulee tällä viikolla.
Tämä viikko on hyvin sateinen.

When we compare the Swahili and Finnish translations, we see that the need to isolate MWEs was different. For Swahili, the cluster *this week* needed no isolation, while for Finnish it was justified. For Finnish, the cluster *very rainy* needed no isolation, but for Swahili it was necessary.

5. Clashing structures and non-consecutive MWEs

Translation requires sometimes careful designing of rules for handling MWEs. Consider the examples in (16). We consider these sentences from the viewpoint of translating them into Finnish.

(16)
They had to bail them out.
They had to be bailed out.
They have to bail them out.
They have to be bailed out.

The first and third sentence are in active mood, and the second and fourth sentence are in passive mood. After analysis and semantic disambiguation, we isolate MWEs (17).

(17)
```
"<They>"
      "they" %SUBJ PRON NOM PL3
"<had>"
      "have" %+FMAINV V PAST >MW { täytyy S-GEN , täytyi S-GEN :2
, täytyy S-ACC , täytyi S-ACC :3 , täydy , täytynyt } MW
"<to>"
      "to" %INFMARK> INFMARK>
"<bail>"
      "bail" %-FMAINV V INF *>MW { vapauttaa V53-C TRV || takuita
vastaan } ADV> MW
"<them>"
      "they" %OBJ PRON PERS PL3
"<out>"
      "out" %ADVL ADV
"<.>"
      "."
"<They>"
      "they" %SUBJ PRON NOM PL3 CAPINIT
"<had>"
      "have" %+FMAINV V PAST >MW { täytyy S-GEN , täytyi S-GEN :2
, täytyy S-ACC , täytyi S-ACC :3 , täydy , täytynyt } MW
"<to>"
```

```
      "to" %INFMARK> INFMARK>
"<be>"
      "be" %-FAUXV V INF
"<bailed>"
      "bail" %-FMAINV V EN *>MW { vapauttaa V53-C TRV || takuita
vastaan } ADV> MW
"<out>"
      "out" %ADVL ADV
"<.>"
      "."
"<They>"
      "they" %SUBJ PRON NOM PL3 CAPINIT
"<have>"
      "have" %+FMAINV V PRES >MW { täytyy S-GEN , täytyi S-GEN :2
, täytyy S-ACC , täytyi S-ACC :3 , täydy , täytynyt } MW
"<to>"
      "to" %INFMARK> INFMARK>
"<bail>"
      "bail" %-FMAINV V INF *>MW { vapauttaa V53-C TRV || takuita
vastaan } ADV> MW
"<them>"
      "they" %OBJ PRON PERS PL3
"<out>"
      "out" %ADVL ADV
"<.>"
      "."
"<They>"
      "they" %SUBJ PRON NOM PL3 CAPINIT
"<have>"
      "have" %+FMAINV V PRES >MW { täytyy S-GEN , täytyi S-GEN :2
, täytyy S-ACC , täytyi S-ACC :3 , täydy , täytynyt } MW
"<to>"
      "to" %INFMARK> INFMARK>
"<be>"
      "be" %-FAUXV V INF
"<bailed>"
      "bail" %-FMAINV V EN *>MW { vapauttaa V53-C TRV || takuita
vastaan } ADV> MW
"<out>"
      "out" %ADVL ADV
"<.>"
      "."
```

We see that the method of isolating the MWEs here is different than what we saw earlier. Here we have used Constraint Grammar rules for marking the head of the MWE, for example >MW. The tag means that this word is part of the MWE, and so is also the next word to the right.

Only the MWEs are given glosses in this phase. Next we mark also the other members as part of the MWE, so that they will be immune when glosses are marked to the rest of words. This phase is described in (18).

(18)

```
"<They>"
      "they" %SUBJ PRON NOM PL3
"<had>"
      "have" { täytyy S-GEN , täytyi S-GEN :2 , täytyy S-ACC ,
täytyi S-ACC :3 , täydy , täytynyt } %+FMAINV MW V PAST >MW
"<to>"
      "to" MW<
"<bail>"
      "bail" %-FMAINV MW V INF *>MW { vapauttaa V53-C TRV ||
takuita vastaan } ADV>
"<them>"
      "they" %OBJ PRON PERS PL3
"<out>"
      "out" { MW*< }
"<.>"
      "."
"<They>"
      "they" %SUBJ PRON NOM PL3 CAPINIT
"<had>"
      "have" { täytyy S-GEN , täytyi S-GEN :2 , täytyy S-ACC ,
täytyi S-ACC :3 , täydy , täytynyt } %+FMAINV MW V PAST >MW
"<to>"
      "to" MW<
"<be>"
      "be" %-FAUXV V INF
"<bailed>"
      "bail" %-FMAINV MW V EN *>MW { vapauttaa V53-C TRV ||
takuita vastaan } ADV>
"<out>"
      "out" { MW*< }
"<.>"
      "."
"<They>"
      "they" %SUBJ PRON NOM PL3 CAPINIT
"<have>"
      "have" { täytyy S-GEN , täytyi S-GEN :2 , täytyy S-ACC ,
täytyi S-ACC :3 , täydy , täytynyt } %+FMAINV MW V PRES >MW
"<to>"
      "to" MW<
"<bail>"
      "bail" %-FMAINV MW V INF *>MW { vapauttaa V53-C TRV ||
takuita vastaan } ADV>
"<them>"
      "they" %OBJ PRON PERS PL3
"<out>"
      "out" { MW*< }
"<.>"
      "."
"<They>"
```

```
      "they" %SUBJ PRON NOM PL3 CAPINIT
"<have>"
      "have" { täytyy S-GEN , täytyi S-GEN :2 , täytyy S-ACC ,
täytyi S-ACC :3 , täydy , täytynyt } %+FMAINV MW V PRES >MW
"<to>"
      "to" MW<
"<be>"
      "be" %-FAUXV V INF
"<bailed>"
      "bail" %-FMAINV MW V EN *>MW { vapauttaa V53-C TRV ||
takuita vastaan } ADV>
"<out>"
      "out" { MW*< }
"<.>"
      "."
```

Now when all members of the MWE are marked, and also their position in the MWE is marked, we can join the members together (19).

(19)

```
"<They>"
      "they" %SUBJ PRON NOM PL3
"<had_to>"
      "have_to" { täytyy S-GEN , täytyi S-GEN :2 , täytyy S-ACC ,
täytyi S-ACC :3 , täydy , täytynyt } %+FMAINV MW MW V PAST
"<bail_out>"
      "bail_out" { vapauttaa V53-C TRV || takuita vastaan } %-
FMAINV MW V INF *>MW ADV>
"<them>"
      "they" %OBJ PRON PERS PL3
"<.>"
      "."
"<They>"
      "they" %SUBJ PRON NOM PL3 CAPINIT
"<had_to>"
      "have_to" { täytyy S-GEN , täytyi S-GEN :2 , täytyy S-ACC ,
täytyi S-ACC :3 , täydy , täytynyt } %+FMAINV MW MW V PAST
"<be>"
      "be" %-FAUXV V INF
"<bailed_out>"
      "bail_out" { vapauttaa V53-C TRV || takuita vastaan } %-
FMAINV MW V EN *>MW ADV>
"<.>"
      "."
"<They>"
      "they" %SUBJ PRON NOM PL3 CAPINIT
"<have_to>"
      "have_to" { täytyy S-GEN , täytyi S-GEN :2 , täytyy S-ACC ,
täytyi S-ACC :3 , täydy , täytynyt } %+FMAINV MW MW V PRES
"<bail_out>"
```

```
      "bail_out" { vapauttaa V53-C TRV || takuita vastaan } %-
FMAINV MW V INF *>MW ADV>
"<them>"
      "they" %OBJ PRON PERS PL3
"<.>"
      "."
"<They>"
      "they" %SUBJ PRON NOM PL3 CAPINIT
"<have_to>"
      "have_to" { täytyy S-GEN , täytyi S-GEN :2 , täytyy S-ACC ,
täytyi S-ACC :3 , täydy , täytynyt } %+FMAINV MW MW V PRES
"<be>"
      "be" %-FAUXV V INF
"<bailed_out>"
      "bail_out" { vapauttaa V53-C TRV || takuita vastaan } %-
FMAINV MW V EN *>MW ADV>
"<.>"
      "."
```

The rule for isolating the cluster *bail out* is such, that it allows words in between. In sentences with active mood this becomes visible, when in the original sentences the words are not immediately after each other.

Then we add glosses to other words and cascade all glosses (20).

(20)

```
"<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
"<had_to>"
      "have_to" { täytyy S-GEN } %+FMAINV MW MW V PAST
      "have_to" { täytyi S-GEN :2 } %+FMAINV MW MW V PAST
      "have_to" { täytyy S-ACC } %+FMAINV MW MW V PAST
      "have_to" { täytyi S-ACC :3 } %+FMAINV MW MW V PAST
      "have_to" { täydy } %+FMAINV MW MW V PAST
      "have_to" { täytynyt } %+FMAINV MW MW V PAST
"<bail_out>"
      "bail_out" { vapauttaa V53-C TRV || takuita vastaan } %-
FMAINV MW V INF *>MW ADV>
"<them>"
      "they" { he Np10 FRONT OUT HUM } %OBJ PRON PERS PL3
      "they" { ne Np12 FRONT } %OBJ PRON PERS PL3
      "they" { NOGLOSS } %OBJ PRON PERS PL3
      "they" { itse N8 FRONT } %OBJ PRON PERS PL3
      "they" { niiden } %OBJ PRON PERS PL3
      "they" { heidän HUM } %OBJ PRON PERS PL3
"<.>"
```

```
      "." { . }
"<They>"
      "they" { he Np10 FRONT OUT HUM } %SUBJ PRON NOM PL3 CAPINIT
      "they" { ne Np12 FRONT } %SUBJ PRON NOM PL3 CAPINIT
      "they" { NOGLOSS } %SUBJ PRON NOM PL3 CAPINIT
      "they" { itse N8 FRONT } %SUBJ PRON NOM PL3 CAPINIT
      "they" { niiden } %SUBJ PRON NOM PL3 CAPINIT
      "they" { heidän HUM } %SUBJ PRON NOM PL3 CAPINIT
"<had_to>"
      "have_to" { täytyy S-GEN } %+FMAINV MW MW V PAST
      "have_to" { täytyi S-GEN :2 } %+FMAINV MW MW V PAST
      "have_to" { täytyy S-ACC } %+FMAINV MW MW V PAST
      "have_to" { täytyi S-ACC :3 } %+FMAINV MW MW V PAST
      "have_to" { täydy } %+FMAINV MW MW V PAST
      "have_to" { täytynyt } %+FMAINV MW MW V PAST
"<be>"
      "be" { olla V67b BE TRV-N V-4INF-TRA } O-LOC1 %-FAUXV V INF
      "be" { olla V67b V-3INF-ILL } O-LOC1 %-FAUXV V INF
      "be" { olla V67b V-3INF-INE } O-LOC1 %-FAUXV V INF
      "be" { olla V67b BE O-PAR } O-LOC1 %-FAUXV V INF
      "be" { eivät ole O-PAR V-4INF-TRA :2 } O-LOC1 %-FAUXV V INF
      "be" { eivät olleet O-PAR V-4INF-TRA :3 } O-LOC1 %-FAUXV V
INF
      "be" { emme :6 } O-LOC1 %-FAUXV V INF
      "be" { emme ole V-3INF-INE } O-LOC1 %-FAUXV V INF
      "be" { emme olleet V-3INF-INE } O-LOC1 %-FAUXV V INF
      "be" { ei ollut V-4INF-TRA } O-LOC1 %-FAUXV V INF
      "be" { ei ole O-PAR V-4INF-TRA } O-LOC1 %-FAUXV V INF
      "be" { NOGLOSS } O-LOC1 %-FAUXV V INF
      "be" { joka Np13 } O-LOC1 %-FAUXV V INF
      "be" { jotka Np14 } O-LOC1 %-FAUXV V INF
      "be" { tulla V67 V-3INF-ILL } O-LOC1 %-FAUXV V INF
"<bailed_out>"
      "bail_out" { vapauttaa V53-C TRV || takuita vastaan } %-
FMAINV MW V EN *>MW ADV>
"<.>"
      "." { . }
"<They>"
      "they" { he Np10 FRONT OUT HUM } %SUBJ PRON NOM PL3 CAPINIT
      "they" { ne Np12 FRONT } %SUBJ PRON NOM PL3 CAPINIT
      "they" { NOGLOSS } %SUBJ PRON NOM PL3 CAPINIT
      "they" { itse N8 FRONT } %SUBJ PRON NOM PL3 CAPINIT
      "they" { niiden } %SUBJ PRON NOM PL3 CAPINIT
      "they" { heidän HUM } %SUBJ PRON NOM PL3 CAPINIT
"<have_to>"
      "have_to" { täytyy S-GEN } %+FMAINV MW MW V PRES
      "have_to" { täytyi S-GEN :2 } %+FMAINV MW MW V PRES
      "have_to" { täytyy S-ACC } %+FMAINV MW MW V PRES
      "have_to" { täytyi S-ACC :3 } %+FMAINV MW MW V PRES
      "have_to" { täydy } %+FMAINV MW MW V PRES
      "have_to" { täytynyt } %+FMAINV MW MW V PRES
```

```
"<bail_out>"
      "bail_out" { vapauttaa V53-C TRV || takuita vastaan } %-
FMAINV MW V INF *>MW ADV>
"<them>"
      "they" { he Np10 FRONT OUT HUM } %OBJ PRON PERS PL3
      "they" { ne Np12 FRONT } %OBJ PRON PERS PL3
      "they" { NOGLOSS } %OBJ PRON PERS PL3
      "they" { itse N8 FRONT } %OBJ PRON PERS PL3
      "they" { niiden } %OBJ PRON PERS PL3
      "they" { heidän HUM } %OBJ PRON PERS PL3
"<.>"
      "." { . }
"<They>"
      "they" { he Np10 FRONT OUT HUM } %SUBJ PRON NOM PL3 CAPINIT
      "they" { ne Np12 FRONT } %SUBJ PRON NOM PL3 CAPINIT
      "they" { NOGLOSS } %SUBJ PRON NOM PL3 CAPINIT
      "they" { itse N8 FRONT } %SUBJ PRON NOM PL3 CAPINIT
      "they" { niiden } %SUBJ PRON NOM PL3 CAPINIT
      "they" { heidän HUM } %SUBJ PRON NOM PL3 CAPINIT
"<have_to>"
      "have_to" { täytyy S-GEN } %+FMAINV MW MW V PRES
      "have_to" { täytyi S-GEN :2 } %+FMAINV MW MW V PRES
      "have_to" { täytyy S-ACC } %+FMAINV MW MW V PRES
      "have_to" { täytyi S-ACC :3 } %+FMAINV MW MW V PRES
      "have_to" { täydy } %+FMAINV MW MW V PRES
      "have_to" { täytynyt } %+FMAINV MW MW V PRES
"<be>"
      "be" { olla V67b BE TRV-N V-4INF-TRA } O-LOC1 %-FAUXV V INF
      "be" { olla V67b V-3INF-ILL } O-LOC1 %-FAUXV V INF
      "be" { olla V67b V-3INF-INE } O-LOC1 %-FAUXV V INF
      "be" { olla V67b BE O-PAR } O-LOC1 %-FAUXV V INF
      "be" { eivät ole O-PAR V-4INF-TRA :2 } O-LOC1 %-FAUXV V INF
      "be" { eivät olleet O-PAR V-4INF-TRA :3 } O-LOC1 %-FAUXV V
INF
      "be" { emme :6 } O-LOC1 %-FAUXV V INF
      "be" { emme ole V-3INF-INE } O-LOC1 %-FAUXV V INF
      "be" { emme olleet V-3INF-INE } O-LOC1 %-FAUXV V INF
      "be" { ei ollut V-4INF-TRA } O-LOC1 %-FAUXV V INF
      "be" { ei ole O-PAR V-4INF-TRA } O-LOC1 %-FAUXV V INF
      "be" { NOGLOSS } O-LOC1 %-FAUXV V INF
      "be" { joka Np13 } O-LOC1 %-FAUXV V INF
      "be" { jotka Np14 } O-LOC1 %-FAUXV V INF
      "be" { tulla V67 V-3INF-ILL } O-LOC1 %-FAUXV V INF
"<bailed_out>"
      "bail_out" { vapauttaa V53-C TRV || takuita vastaan } %-
FMAINV MW V EN *>MW ADV>
"<.>"
      "." { . }
```

The reading is then disambiguated semantically (21).

(21)

```
"<They>"
      "they" { he Np10 FRONT OUT HUM } %SUBJ PRON NOM PL3
"<had_to>"
      "have_to" { täytyi S-ACC :3 } %+FMAINV MW MW V PAST
"<bail_out>"
      "bail_out" { vapauttaa V53-C TRV || takuita vastaan } %-
FMAINV MW V INF *>MW ADV>
"<them>"
      "they" { he Np10 FRONT OUT HUM } %OBJ PRON PERS PL3
"<.>"
      "." { . }
"<They>"
      "they" { he Np10 FRONT OUT HUM } %SUBJ PRON NOM PL3 CAPINIT
"<had_to>"
      "have_to" { täytyi S-ACC :3 } %+FMAINV MW MW V PAST
"<be>"
      "be" { olla V67b V-3INF-ILL } O-LOC1 %-FAUXV V INF
"<bailed_out>"
      "bail_out" { vapauttaa V53-C TRV || takuita vastaan } %-
FMAINV MW V EN *>MW ADV>
"<.>"
      "." { . }
"<They>"
      "they" { he Np10 FRONT OUT HUM } %SUBJ PRON NOM PL3 CAPINIT
"<have_to>"
      "have_to" { täytyy S-GEN } %+FMAINV MW MW V PRES
"<bail_out>"
      "bail_out" { vapauttaa V53-C TRV || takuita vastaan } %-
FMAINV MW V INF *>MW ADV>
"<them>"
      "they" { he Np10 FRONT OUT HUM } %OBJ PRON PERS PL3
"<.>"
      "." { . }
"<They>"
      "they" { he Np10 FRONT OUT HUM } %SUBJ PRON NOM PL3 CAPINIT
"<have_to>"
      "have_to" { täytyy S-ACC } %+FMAINV MW MW V PRES
"<be>"
      "be" { olla V67b V-3INF-ILL } O-LOC1 %-FAUXV V INF
"<bailed_out>"
      "bail_out" { vapauttaa V53-C TRV || takuita vastaan } %-
FMAINV MW V EN *>MW ADV>
"<.>"
      "." { . }
```

Inflection codes are added, so that each word cold be provided with its appropriate inflection (22).

```
(22)
"<They>"
```

```
      "they" { he Np10 FRONT } %SUBJ OUT HUM PRON PL3 GEN
"<had_to>"
      "have_to" { täytyi } %+FMAINV S-ACC MW MW V PAST PL
"<bail_out>"
      "bail_out" { vapauttaa V53-C } %-FMAINV TRV MW V INF *>MW
ADV> SG
"<takuita vastaan>"
      "takuita vastaan" { +takuita vastaan } X PL
"<them>"
      "they" { he Np10 FRONT } %OBJ OUT HUM PRON PERS  PL ACC
"<.>"
      "." { . }
"<They>"
      "they" { he Np10 FRONT } %SUBJ OUT HUM PRON PL3 CAPINIT ACC-
N
"<had_to>"
      "have_to" { täytyi } %+FMAINV S-ACC MW MW V PAST PL
"<be>"
      "be" { olla V67b } %-FAUXV V-3INF-ILL O-LOC1 V INF SG
"<bailed_out>"
      "bail_out" { vapauttaa V53-C } %-FMAINV TRV MW V EN *>MW
ADV> PL
"<takuita vastaan>"
      "takuita vastaan" { +takuita vastaan } X PL
"<.>"
      "." { . }
"<They>"
      "they" { he Np10 FRONT } %SUBJ OUT HUM PRON PL3 CAPINIT GEN
"<have_to>"
      "have_to" { täytyy } %+FMAINV S-GEN MW MW V PRES PL
"<bail_out>"
      "bail_out" { vapauttaa V53-C } %-FMAINV TRV MW V INF *>MW
ADV> SG
"<takuita vastaan>"
      "takuita vastaan" { +takuita vastaan } X PL
"<them>"
      "they" { he Np10 FRONT } %OBJ OUT HUM PRON PERS  PL ACC
"<.>"
      "." { . }
"<They>"
      "they" { he Np10 FRONT } %SUBJ OUT HUM PRON PL3 CAPINIT ACC-
N
"<have_to>"
      "have_to" { täytyy } %+FMAINV S-ACC MW MW V PRES PL
"<be>"
      "be" { olla V67b } %-FAUXV V-3INF-ILL O-LOC1 V INF SG
"<bailed_out>"
      "bail_out" { vapauttaa V53-C } %-FMAINV TRV MW V EN *>MW
ADV> PL
"<takuita vastaan>"
      "takuita vastaan" { +takuita vastaan } X PL
```

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

Especially note that the pronoun subject has a different inflection tag in sentences with active and passive mood.

Also note that the MWE *bail out* has the gloss *vapauttaa takuita vastaan*, which also is a MWE. This MWE is split into two sections using a double vertical bar ||, so that they can be handled as separate units.

We see the usefulness of this method in final translation, when the object *them* is inserted between the two parts of the MWE (23).

(23)
*Heidän täytyi vapauttaa **heidät** takuita vastaan.*
Heidät täytyi vapauttaa takuita vastaan.
*Heidän täytyy vapauttaa **heidät** takuita vastaan.*
Heidät täytyy vapauttaa takuita vastaan.

6. Conclusion

We have discussed and demonstrated two methods of isolating MWEs in English text. The conclusion is that all such MWEs, which are not dependent on context, or which are consecutive, should be isolated as early as possible. In practice, the best place for isolation is after tokenisation, but before analysis. The isolated MWEs are also listed into the morphological lexicon, so that they can be directly handled as single units, with no need for morphological disambiguation.

The second phase of MWE isolation comes after morphological and lexical disambiguation, but before adding the glosses of the target text. In the second phase, such MWEs are isolated, which require checking the context, or which are non-consecutive.

The tests show also that the set of MWEs are dependent on the target language. Finally, the most important criterium for deciding whether a word cluster should be interpreted as a MWE or not is the translation result. There are also a number of cases, where direct translation and translation via a MWE isolation process are equally good solutions.

On the basis of the above discussion we can also conclude that the first phase MWE isolation includes such cases, which can be considered global. That is, they are not dependent on target language.

The MWEs of the second phase include many such cases, which suit to some language types, but not to others. It is likely that for each target language a separate isolation system should be constructed and maintained.