Technical Reports on Language Technology
Report No 93 2023
http://www.njas.helsinki.fi/salama

Managing articles in Swahili to English machine translation[1]

Arvi Hurskainen
Department of Languages
FIN-00014 University of Helsinki, Finland
arvi.hurskainen@helsinki.fi
DOI: 10.13140/RG.2.2.25290.24002

Abstract

When the source language does not have such a concept as articles, the management of articles in a target language such as English is problematic. As the articles are missing in the source language, the articles in target language must be added using the properties of nouns as well as the context, where nouns are used. We can consider two measures for managing the articles. In the morphological lexicon, we mark all non-countable nouns, because they behave in a different way than countable nouns. We also can use Constraint Grammar rules for adding articles on the basis of context. This report describes these procedures using Swahili as source language and English as target language.

Key Words: *machine translation, assignment of articles.*

1 Introduction

In English, articles are words that define a noun as specific or unspecific. English has two types of articles: indefinite (*a, an*) and definite (*the*). The articles *a* and *an* are used for singular nouns for defining indefiniteness, while the omission of the article marks the indefiniteness in plural. The article *the* is used for defining definiteness in singular and plural.

There are two basic rules for the use of articles. (1) If the noun is singular and countable, and this is the first time it has been mentioned, then you will usually need an indefinite article. (2) If you believe your reader or listener knows exactly what you are referring to, then you will usually need the definite article in front of a noun.

The first rule indicates that if the noun is noncountable, the article will very likely be omitted. Such words do not usually have a plural form. If noncountable nouns are separately marked in the lexicon, the use of articles with such nouns can be handled.

2 How should the insertion of articles be approached?

When the source language does not have articles, there is nothing that could be converted into articles in target language. However, we are not without clues. For example, the countability of a noun can be marked. This can be done in the lexicon in two ways. In one

method, each noncountable word is marked separately, while the unmarked nouns are considered countable. In the other method, noncountable nouns of each noun class are moved into a separate sub-lexicon. The latter method is better, because using this method there is no need to mark each noun. Marking can be done through a sub-lexicon with one entry only.

Also the noun-class system itself makes the identification somewhat easier than anticipated. For example, the classes 1/2, 3/4, 5/6, and 7/8 contain countable nouns. The classes 9/10 and 11 contain both types of nouns. Therefore, the separation should be done within these classes.

When rules for adding articles are written, the uncountability tag can be used for preventing the adding of the article code.

By default, the lexicon does not have articles for any nouns. Therefore, when text is analysed, no articles or article codes will appear in the result (1).

(1)

```
"<*taa>"
      "taa" N { lamp , lantern } CAP 9/10-SG AR @SUBJ
"<inawaka>"
      "waka" V 9-SG-SP VFIN NO-SP-GLOSS PR:na z [waa] { be :alight
} SV STAT PREFR @FMAINVintr
"<.>"
      "." STOP { . } **CLB
"<*maziwa>"
      "maziwa" N { :milk } CAP 6-PLSG MASS @SUBJ
"<yanapatikana>"
      "patikana" V 6-PLSG-SP VFIN NO-SP-GLOSS PR:na z [pata] { be
available } SV PREFR REC @FMAINVintr
"<.>"
      "." STOP { . } **CLB
"<*maziwa>"
      "maziwa" N { :milk } CAP 6-PLSG MASS @SUBJ
"<ya>"
      "ya" GEN-CON 6-PL { of } @GCON
"<ng'ombe>"
      "ng'ombe" N { cow , cattle , ox , bull , bullock } 9/10-SG
AN DOM-AN @<NH
"<yameuzwa>"
      "uza" V 6-PL-SP VFIN NO-SP-GLOSS PERF:me z [uza] { sell }
SVO PASS @FMAINVtr-OBJ>
"<.>"
      "." STOP { . } **CLB
"<*yeye>"
      "yeye" CAP PRON PERS-PRON SG3 { he } @SUBJ
"<alinunua>"
      "nunua" V 1-SG3-SP VFIN NO-SP-GLOSS PAST z [nunua] { buy ,
purchase } SVO HUM-ACT @FMAINVtr+OBJ>
"<gari>"
      "gari" N { car , vehicle } 9/10-SG IND @OBJ
"<.>"
      "." STOP { . } **CLB
```

```
"<*yeye>"
      "yeye" CAP PRON PERS-PRON SG3 { he } @SUBJ
"<alinunua>"
      "nunua" V 1-SG3-SP VFIN NO-SP-GLOSS PAST z [nunua] { buy ,
purchase } SVO HUM-ACT @FMAINVtr+OBJ>
"<gari>"
      "gari" N { car , vehicle } 9/10-SG IND @OBJ
"<nzuri>"
      "zuri" ADJ { good , beautiful , pretty , gorgeous } A-INFL
9-SG @<NADJ
"<.>"
      "." STOP { . } **CLB
"<*maziwa>"
      "maziwa" N { :milk } CAP 6-PLSG MASS @SUBJ
"<mazuri>"
      "zuri" ADJ { good , beautiful , pretty , gorgeous } A-INFL
6-PL @<NADJ
"<ya>"
      "ya" GEN-CON 6-PL { of } @GCON
"<ng'ombe>"
      "ng'ombe" N { cow , cattle , ox , bull , bullock } 9/10-SG
AN DOM-AN @<NH
"<yameuzwa>"
      "uza" V 6-PL-SP VFIN NO-SP-GLOSS PERF:me z [uza] { sell }
SVO PASS @FMAINVtr-OBJ>
"<.>"
      "." STOP { . } **CLB
```

Articles are added to the analysed text using Constraint Grammar (CG) rules. Articles are not added directly as such. They are first described as two codes, DEFART and INDEFART (2).

(2)

```
"<*taa>"
      "taa" N { lamp , lantern } CAP 9/10-SG AR %SUBJ DEFART
"<inawaka>"
      "waka" V 9-SG-SP VFIN NO-SP-GLOSS PR:na z [waa] { be :alight
} SV STAT PREFR %FMAINVintr
"<.>"
      "." STOP { . } **CLB
"<*maziwa>"
      "maziwa" N { :milk } CAP 6-PLSG MASS %SUBJ
"<yanapatikana>"
      "patikana" V 6-PLSG-SP VFIN NO-SP-GLOSS PR:na z [pata] { be
available } SV PREFR REC %FMAINVintr
"<.>"
      "." STOP { . } **CLB
"<*maziwa>"
      "maziwa" N { :milk } CAP 6-PLSG MASS %SUBJ DEFART
"<ya>"
      "ya" GEN-CON 6-PL { of } %GCON
```

```
"<ng'ombe>"
      "ng'ombe" N { cow , cattle , ox , bull , bullock } 9/10-SG
AN DOM-AN %<NH
"<yameuzwa>"
      "uza" V 6-PL-SP VFIN NO-SP-GLOSS PERF:me z [uza] { sell }
SVO PASS %FMAINVtr-OBJ>
"<.>"
      "." STOP { . } **CLB
"<*yeye>"
      "yeye" CAP PRON PERS-PRON SG3 { he } %SUBJ
"<alinunua>"
      "nunua" V 1-SG3-SP VFIN NO-SP-GLOSS PAST z [nunua] { buy ,
purchase } SVO HUM-ACT %FMAINVtr+OBJ>
"<gari>"
      "gari" N { car , vehicle } 9/10-SG IND %OBJ INDEFART
"<.>"
      "." STOP { . } **CLB
"<*yeye>"
      "yeye" CAP PRON PERS-PRON SG3 { he } %SUBJ
"<alinunua>"
      "nunua" V 1-SG3-SP VFIN NO-SP-GLOSS PAST z [nunua] { buy ,
purchase } SVO HUM-ACT %FMAINVtr+OBJ>
"<gari>"
      "gari" N { car , vehicle } 9/10-SG IND %OBJ
"<nzuri>"
      "zuri" ADJ { good , beautiful , pretty , gorgeous } A-INFL
9-SG %<NADJ INDEFART
"<.>"
      "." STOP { . } **CLB
"<*maziwa>"
      "maziwa" N { :milk } CAP 6-PLSG MASS %SUBJ
"<mazuri>"
      "zuri" ADJ { good , beautiful , pretty , gorgeous } A-INFL
6-PL %<NADJ DEFART
"<ya>"
      "ya" GEN-CON 6-PL { of } %GCON
"<ng'ombe>"
      "ng'ombe" N { cow , cattle , ox , bull , bullock } 9/10-SG
AN DOM-AN %<NH
"<yameuzwa>"
      "uza" V 6-PL-SP VFIN NO-SP-GLOSS PERF:me z [uza] { sell }
SVO PASS %FMAINVtr-OBJ>
"<.>"
      "." STOP { . } **CLB
```

Note that if the noun has an adjective modifier, such as *gari nzuri* or *maziwa mazuri*, the article tag is added to the adjective instead of the noun. By so doing we anticipate the place of the article in the target language.

The codes for articles are converted into surface form, and they are moved to the left, so that they are ready in the correct place when the process continues (3).

(3)

```
"<*taa>"
      "taa" N { the } { lamp , lantern } CAP 9/10-SG AR %SUBJ
"<inawaka>"
      "waka" V 9-SG-SP VFIN NO-SP-GLOSS PR:na z [waa] { be :alight
} SV STAT PREFR %FMAINVintr
"<.>"
      "." STOP { . } **CLB
"<*maziwa>"
      "maziwa" N { :milk } CAP 6-PLSG MASS %SUBJ
"<yanapatikana>"
      "patikana" V 6-PLSG-SP VFIN NO-SP-GLOSS PR:na z [pata] { be
available } SV PREFR REC %FMAINVintr
"<.>"
      "." STOP { . } **CLB
"<*maziwa>"
      "maziwa" N { the } { :milk } CAP 6-PLSG MASS %SUBJ
"<ya>"
      "ya" GEN-CON 6-PL { of } %GCON
"<ng'ombe>"
      "ng'ombe" N { cow , cattle , ox , bull , bullock } 9/10-SG
AN DOM-AN %<NH
"<yameuzwa>"
      "uza" V 6-PL-SP VFIN NO-SP-GLOSS PERF:me z [uza] { sell }
SVO PASS %FMAINVtr-OBJ>
"<.>"
      "." STOP { . } **CLB
"<*yeye>"
      "yeye" CAP PRON PERS-PRON SG3 { he } %SUBJ
"<alinunua>"
      "nunua" V 1-SG3-SP VFIN NO-SP-GLOSS PAST z [nunua] { buy ,
purchase } SVO HUM-ACT %FMAINVtr+OBJ>
"<gari>"
      "gari" N { a } { car , vehicle } 9/10-SG IND %OBJ
"<.>"
      "." STOP { . } **CLB
"<*yeye>"
      "yeye" CAP PRON PERS-PRON SG3 { he } %SUBJ
"<alinunua>"
      "nunua" V 1-SG3-SP VFIN NO-SP-GLOSS PAST z [nunua] { buy ,
purchase } SVO HUM-ACT %FMAINVtr+OBJ>
"<gari>"
      "gari" N { car , vehicle } 9/10-SG IND %OBJ
"<nzuri>"
      "zuri" ADJ { a } { good , beautiful , pretty , gorgeous } A-
INFL 9-SG %<NADJ
"<.>"
      "." STOP { . } **CLB
"<*maziwa>"
      "maziwa" N { :milk } CAP 6-PLSG MASS %SUBJ
```

```
"<mazuri>"
      "zuri" ADJ { the } { good , beautiful , pretty , gorgeous }
A-INFL 6-PL %<NADJ
"<ya>"
      "ya" GEN-CON 6-PL { of } %GCON
"<ng'ombe>"
      "ng'ombe" N { cow , cattle , ox , bull , bullock } 9/10-SG
AN DOM-AN %<NH
"<yameuzwa>"
      "uza" V 6-PL-SP VFIN NO-SP-GLOSS PERF:me z [uza] { sell }
SVO PASS %FMAINVtr-OBJ>
"<.>"
      "." STOP { . } **CLB
```

The glosses are disambiguated, that is, if the word has more than one gloss candidate, the most appropriate one is chosen, and other glosses are removed (4).

(4)

```
( N { *the } { lamp } CAP 9/10-SG %SUBJ )
( V 9-SG-SP VFIN NO-SP-GLOSS PR:na z { be :alight } SV STAT PREFR
%FMAINVintr )
( STOP { . } **CLB )
( N { :milk } CAP 6-PLSG MASS %SUBJ )
( V 6-PLSG-SP VFIN NO-SP-GLOSS PR:na z { be available } SV PREFR
REC %FMAINVintr )
( STOP { . } **CLB )
( N { *the } { :milk } CAP 6-PLSG MASS %SUBJ )
( GEN-CON 6-PL { of } %GCON )
( N { cow } 9/10-SG AN DOM-AN %<NH )
( V 6-PL-SP VFIN NO-SP-GLOSS PERF:me z { sell } SVO PASS
%FMAINVtr-OBJ> )
( STOP { . } **CLB )
( PRON CAP PERS-PRON SG3 { *he } %SUBJ )
( V 1-SG3-SP VFIN NO-SP-GLOSS PAST z { buy } SVO HUM-ACT
%FMAINVtr+OBJ> )
( N { a } { car } 9/10-SG %OBJ )
( STOP { . } **CLB )
( PRON CAP PERS-PRON SG3 { *he } %SUBJ )
( V 1-SG3-SP VFIN NO-SP-GLOSS PAST z { buy } SVO HUM-ACT
%FMAINVtr+OBJ> )
:( ADJ { a } { good } A-INFL 9-SG %<NADJ )
:( N { car } 9/10-SG %OBJ )
( STOP { . } **CLB )
:( ADJ { *the } { good } A-INFL 6-PL %<NADJ )
:( N { :milk } CAP 6-PLSG MASS %SUBJ )
( GEN-CON 6-PL { of } %GCON )
( N { cow } 9/10-SG AN DOM-AN %<NH )
( V 6-PL-SP VFIN NO-SP-GLOSS PERF:me z { sell } SVO PASS
%FMAINVtr-OBJ> )
( STOP { . } **CLB )
```

Using the linguistic tags, the glosses are converted into surface form (5).

(5)
```
( N { *the } { lamp } CAP 9/10-SG %SUBJ )
( V 9-SG-SP VFIN NO-SP-GLOSS PR:na :z { :is :alight } SV STAT
PREFR %FMAINVintr )
( STOP { . } **CLB )
( "<<s>>" { <s> } )
( N { :milk } CAP 6-PLSG MASS %SUBJ )
( V 6-PLSG-SP VFIN NO-SP-GLOSS PR:na :z { :is available } SV PREFR
REC %FMAINVintr )
( STOP { . } **CLB )
( "<<s>>" { <s> } )
( N { *the } { :milk } CAP 6-PLSG MASS %SUBJ )
( GEN-CON 6-PL { of } %GCON )
( N { cow } 9/10-SG AN DOM-AN %<NH )
( V 6-PL-SP VFIN NO-SP-GLOSS PERF:me :z { :has } { :been } { :sold
} SVO PASS %FMAINVtr-OBJ> )
( STOP { . } **CLB )
( "<<s>>" { <s> } )
( PRON CAP PERS-PRON SG3 { *he } %SUBJ )
( V 1-SG3-SP VFIN NO-SP-GLOSS PAST :z { :bought } SVO HUM-ACT
%FMAINVtr+OBJ> )
( N { a } { car } 9/10-SG %OBJ )
( STOP { . } **CLB )
( "<<s>>" { <s> } )
( PRON CAP PERS-PRON SG3 { *he } %SUBJ )
( V 1-SG3-SP VFIN NO-SP-GLOSS PAST :z { :bought } SVO HUM-ACT
%FMAINVtr+OBJ> )
:( ADJ { a } { good } A-INFL 9-SG %<NADJ )
:( N { car } 9/10-SG %OBJ )
( STOP { . } **CLB )
( "<<s>>" { <s> } )
:( ADJ { *the } { good } A-INFL 6-PL %<NADJ )
:( N { :milk } CAP 6-PLSG MASS %SUBJ )
( GEN-CON 6-PL { of } %GCON )
( N { cow } 9/10-SG AN DOM-AN %<NH )
( V 6-PL-SP VFIN NO-SP-GLOSS PERF:me :z { :has } { :been } { :sold
} SVO PASS %FMAINVtr-OBJ> )
( STOP { . } **CLB )
```

Note that when the word stem has been converted to surface form, a colon ':' is added in front of it, so that another rule will not affect it. The translation can now be produced (6).

(6)
(a) *The lamp is alight.*
(b) *Milk is available.*
(c) *The milk of cow has been sold.*
(d) *He bought a car.*

(e) *He bought a good car*.
(f) *The good milk of cow has been sold.*

We take a look at each translation.

In (a), the noun *lamp* has a definite article, because it is countable and it is considered known, because it is in the beginning of the sentence.

In (b), although the noun *milk* is in the beginning of the sentence, it is noncountable, and therefore the article is omitted.

In (c), the noun *milk* has a definite article, because it is modified by the noun *cow*.

In (d), the noun *car* has an indefinite article, because it is in the end of the sentence, which indicates that it is not yet known.

In (e), the noun *car*, preceded by an adjective modifier, has an indefinite article, because it is assumed that the car is not yet known. In case there would be a choice between a good and bad car, the article should be definite.

In (f), the noun *milk*, preceded by an adjective modifier, has a definite article, because it is also modified by the noun *cow*. In such a structure, the noun has a definite article, although it is noncountable.

3 Rules for controlling the articles in English

Above I have shown the method of inserting articles to target language in the case when the source language does not use articles. The insertion of articles was implemented using CG rules. Below we will see in more detail the CG rules.
The rules are in (7).

(7)

(a) `MAP (DEFART) TARGET N (*1 GEN-CON BARRIER CLB)(NOT 1 ADJ) (NOT 0 (PROPNAME) OR TITLE OR NENT OR NOART OR (15-SG)) (NOT 1 NUM OR DEM OR POSS) (NOT -1 (have));`

(b) `MAP (DEFART) TARGET N (-1 SNTB) (NOT 1 ADJ OR DEM OR POSS OR NUM) (NOT 0 MASS);`

(c) `MAP (INDEFART) TARGET N (NOT 1 GEN-CON OR NUM OR DEM OR POSS) (NOT 0 MASS OR NOART OR NENT) (NOT -1 SNTB);`

(d) `MAP (INDEFART) TARGET N (*-1 V BARRIER CLB) (NOT 1 ADJ OR NUM OR DEM OR POSS OR GEN-CON) (NOT 0 N-PL OR (PROPNAME) OR NENT OR NOART OR (15-SG) OR LOC) (NOT -1 GEN-CON);`

(e) `MAP (INDEFART) TARGET ADJ (-1 N) (*-1 V BARRIER CLB) (NOT -1 N-PL);`

(f) `MAP (DEFART) TARGET ADJ (-1 N) (NOT *-1 V BARRIER CLB) (NOT 0 ("chache"));`

Explanation of rule (a): Add (MAP) the definite article code (DEFART) to the noun (N). On the right there is a genitive connector (GEN-CON), but do not scan beyond the clause boundary (CLB). The next word should not be an adjective (ADJ). The target noun should not be a proper name (PROPNAME), or title (TITLE), or named entity (NENT), or a noun that never gets an article (NOART), or of class 15-SG. The next word should not be a

number (NUM), or a demonstrative pronoun (DEM), or a possessive pronoun (POSS). The first word to the left should not be the verb *have*.

Explanation of rule (b): Add the definite article code (DEFART) to the noun (N), if immediately on the left there is a sentence boundary (SNTB). The next word on the right should not be an adjective (ADJ), or demonstrative pronoun (DEM), or possessive pronoun (POSS), or number (NUM). The noun should not be a mass noun (MASS).

Explanation of rule (c): Add the indefinite article code (INDEFART) to the noun (N). The first word on the right should not be a genitive connector (GEN-CON), or number (NUM), or demonstrative pronoun (DEM), or a possessive pronoun (POSS). The target noun should not be a mass noun (MASS), or a noun that never takes an article (NOART), or a named entity (NENT). The target word should not be the first word of the sentence (SNTB).

Explanation of rule (d): Add the indefinite article code (INDEFART) to the noun (N). Somewhere on the left, there should be a verb, but do not scan beyond the clause boundary. The next word should not be an adjective (ADJ), or number (NUM), or demonstrative pronoun (DEM), or possessive pronoun (POSS), or genitive connector (GEN-CON). The target noun should not have a plural form (N-PL), it should not be a proper name (PROPNAME), or a noun that never gets an article (NOART), or a noun of the class 15 (15-SG), or a locative form (LOC). The first word on the left should not be a genitive connector (GEN-CON).

Explanation of rule (e): Add the indefinite article code (INDEFART) to the adjective (ADJ). The first word on the left should be a noun (N). On the left there should be a verb (V), but do not scan beyond the clause boundary (CLB). The first word on the left should not be a noun in plural form (N-PL).

Explanation of rule (f): Add the definite article code (DEFART) to the adjective (ADJ). The first word to the left should be a noun (N). No word to the left should be a verb (V), but do not scan beyond the clause boundary (CLB). The target word should not be *chache*.

Note that many tags used in rules are set names, each containing a number of tags. For example, the set name N-PL contains all noun class tags of plural forms.

The CG rules were developed and tested using a text from BBC News. It is likely that more constraints are needed for the rules. However, the rules cover the most common instances of the use of articles.

4 Performance of the system

The performance of the system was tested using a news text of 1173 words. This gives a rough picture of the performance of the system. We can see in Table 1 that in most cases the articles were correctly inserted. The cases with incorrect insertion were due to insufficient constraints in CG rules. They can be easily corrected.

Table 1. Performance of the system to insert articles.

Type of article	correct	incorrect
sg indef	34	4
sg def	48	4
pl def	8	2
sg noart	33	3
pl noart	13	0

5 Conclusion

In the report I have shown how articles can be inserted in a translation system, where the source language has no articles. The discussion was carried out using the Swahili to English translation system. The same approach applies also to the language pairs, where the source language has no articles. The Finnish to English translation system is an example of such cases.