Technical Reports in Language Technology
Report No 16, 20014
http://www.njas.helsinki.fi/salama

Weekdays as Person Names in Swahili to English Machine Translation

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

Abstract

This paper discusses the problem that the words for weekdays can appear in the meaning of indicating time, and also as person names. In both meanings the word is written with a capital-initial letter. However, in one meaning it must be translated, and in another meaning it must be left as such. The problem is elucidated with examples, and solutions for disambiguating between these meanings are presented.

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

1 Introduction

Earlier in this report series (Technical Report No 12, 2013) I discussed the problem of handling various types of proper names in machine translation (MT). The proper names can be classified into (a) proper names that need no translation and can have no other meaning in text (e.g. Paulo); (b) proper names that have a unique meaning, but must be translated (e.g. names of countries); and (c) proper names that have ordinary meanings but also roles as proper names (e.g. Rehema).

The general approach in the report was that if the word is not sentence initial, but is written with a capital-initial letter, it must be interpreted as a proper name. Sentence-initial words form a special problem, because all words are written with a capital-initial letter in this position.[1]

In this paper I discuss a special sub-group of weekdays. They are not normal nouns, because they are always written with a capital-initial letter. They would normally belong to group (b) above, but they are frequently used also as person names. In the meaning of expressing time they should be translated, but in the meaning expressing the name of a person they should not be translated. In brief, in both meanings they are proper names with a capital-initial letter, and they can appear anywhere in the sentence.

2 Weekdays as ordinary nouns

In order to decrease unnecessary ambiguity, I have analysed weekdays as ordinary nouns, although they start with a capital-initial letter (1)

[1] In addition to the three groups listed above, there is a large number of organization names, permanent and ad hoc ones, that are written with capital-initial letters. These confuse further the handling of proper names. The frequent ones at least must be isolated and handled as multiword expressions.

(1)
```
"<*jumamosi>"
      "jumamosi" N 9/10-SG { *saturday } WEEK CAP
      "jumamosi" N 9/10-PL { *saturday } WEEK CAP
"<*jumapili>"
      "jumapili" N 9/10-SG { *sunday } } WEEK CAP
      "jumapili" N 9/10-PL { *sunday } } WEEK CAP
"<*jumatatu>"
      "jumatatu" N 9/10-SG { *monday } WEEK CAP
      "jumatatu" N 9/10-PL { *monday } WEEK CAP
"<*jumanne>"
      "jumanne" N 9/10-SG { *tuesday } WEEK CAP
      "jumanne" N 9/10-PL { *tuesday } WEEK CAP
"<*jumatano>"
      "jumatano" N 9/10-SG { *wednesday } WEEK CAP
      "jumatano" N 9/10-PL { *wednesday } WEEK CAP
"<*alhamisi>"
      "alhamisi" N 9/10-SG { *thursday } WEEK CAP
      "alhamisi" N 9/10-PL { *thursday } WEEK CAP
"<*ijumaa>"
      "ijumaa" N 9/10-SG { *friday } WEEK CAP
      "ijumaa" N 9/10-PL { *friday } WEEK CAP
```

This interpretation allows the weekday names to be used in singular and plural, although singular and plural are formally identical. If the words are not used in plural, the disambiguated result is as in (2).

(2)
```
"<*jumamosi>"
      "jumamosi" N 9/10-SG { *saturday } WEEK CAP
"<*jumapili>"
      "jumapili" N 9/10-SG { *sunday } } WEEK CAP
"<*jumatatu>"
      "jumatatu" N 9/10-SG { *monday } WEEK CAP
"<*jumanne>"
      "jumanne" N 9/10-SG { *tuesday } WEEK CAP
"<*jumatano>"
      "jumatano" N 9/10-SG { *wednesday } WEEK CAP
"<*alhamisi>"
      "alhamisi" N 9/10-SG { *thursday } WEEK CAP
"<*ijumaa>"
      "ijumaa" N 9/10-SG { *friday } WEEK CAP
```

Then we add the tag PROP-CAND to the readings to indicate that these words are candidates for proper name interpretation (in this case as a person name). The result of this process is in (3).

(3)
```
"<*jumamosi>"
      "jumamosi" N 9/10-SG { *saturday } WEEK CAP
"<*jumapili>"
      "jumapili" N 9/10-SG { *sunday } } WEEK CAP PROP-CAND
"<*jumatatu>"
      "jumatatu" N 9/10-SG { *monday } WEEK CAP PROP-CAND
"<*jumanne>"
```

```
      "jumanne" N 9/10-SG { *tuesday } WEEK CAP PROP-CAND
"<*jumatano>"
      "jumatano" N 9/10-SG { *wednesday } WEEK CAP PROP-CAND
"<*alhamisi>"
      "alhamisi" N 9/10-SG { *thursday } WEEK CAP PROP-CAND
"<*ijumaa>"
      "ijumaa" N 9/10-SG { *friday } WEEK CAP PROP-CAND
```

Note that the first word in the list does not have the tag PROP-CAND. It is interpreted as a sentence-initial position, and therefore the tag is omitted.

In the next phase we modify the output so that each weekday word has two interpretations, one for a proper name and one for the original weekday meaning (4).

(4)

```
"<*jumamosi>"
      "jumamosi" N 9/10-SG { *saturday } WEEK CAP @SUBJ
"<*jumapili>"
      "jumapili" N 9/10-SG { *sunday } } CAP PROP-CAND
      "jumapili" N 9/10-SG { *sunday } } WEEK CAP
"<*jumatatu>"
      "jumatatu" N 9/10-SG { *monday } CAP PROP-CAND
      "jumatatu" N 9/10-SG { *monday } WEEK CAP
"<*jumanne>"
      "jumanne" N 9/10-SG { *tuesday } CAP PROP-CAND
      "jumanne" N 9/10-SG { *tuesday } WEEK CAP
"<*jumatano>"
      "jumatano" N 9/10-SG { *wednesday } CAP PROP-CAND
      "jumatano" N 9/10-SG { *wednesday } WEEK CAP
"<*alhamisi>"
      "alhamisi" N 9/10-SG { *thursday } CAP PROP-CAND
      "alhamisi" N 9/10-SG { *thursday } WEEK CAP
"<*ijumaa>"
      "ijumaa" N 9/10-SG { *friday } CAP PROP-CAND
      "ijumaa" N 9/10-SG { *friday } WEEK CAP
```

On the basis of the above format, where the weekdays are simply listed, we cannot proceed. We have to locate the words into real example sentences. Consider the examples in (5).

(5)

```
"<*profesa>"
      "*profesa" N TITLE { *professor } AN HUM @SUBJ
"<*jumanne>"
      "jumanne" N 9/10-SG { *tuesday } CAP PROP-CAND
      "jumanne" N 9/10-SG { *tuesday } WEEK CAP
"<*maghembe>"
      "*maghembe" PROPNAME SG { *maghembe } @SUBJ
"<alijibu>"
      "jibu" V 1-SG3-SP VFIN NO-SP-GLOSS PAST z [jibu] { answer }
SVO @FMAINVtr-OBJ>
"<.$>"
      ".$" { .$ } **CLB
"<*nitakuja>"
      "ja" V 1-SG1-SP VFIN { *i } FUT:ta INFMARK z [ja] { come }
SV MONOSLB CAP @FMAINVintr
```

```
"<huko>"
      "huko" ADV { in } @ADVL
"<*jumanne>"
      "jumanne" N 9/10-SG { *tuesday } CAP PROP-CAND @SUBJ
      "jumanne" N 9/10-SG { *tuesday } WEEK CAP @SUBJ
"<wiki_ijayo>"
      "wiki_ja" ADV { next week } TIME @ADVL
"<.$>"
      ".$" { .$ } **CLB
```

The word *Jumanne* appears in two contexts with different meanings. How can we select the correct interpretation in each context? We could think that if a weekday name is followed by a proper name, the weekday is likely to be in the role of a person name. In the second sentence we could think that if the weekday name is followed, immediately or within the same clause, by a word that has the tag TIME, the weekday name should be interpreted as a proper weekday. After the application of these rules we get the result as in (6).

(6)

```
"<*profesa>"
      "*profesa" N TITLE { *professor } AN HUM @SUBJ
"<*jumanne>"
      "jumanne" N 9/10-SG { *tuesday } CAP PROP-CAND
"<*maghembe>"
      "*maghembe" PROPNAME SG { *maghembe } @SUBJ
"<alijibu>"
            "jibu" V 1-SG3-SP VFIN NO-SP-GLOSS PAST z [jibu] {
answer } SVO @FMAINVtr-OBJ>
"<.$>"
      ".$" { .$ } **CLB
"<*nitakuja>"
      "ja" V 1-SG1-SP VFIN { *i } FUT:ta INFMARK z [ja] { come }
SV MONOSLB CAP @FMAINVintr
"<huko>"
      "huko" ADV { in } @ADVL
"<*jumanne>"
      "jumanne" N 9/10-SG { *tuesday } WEEK CAP @SUBJ
"<wiki_ijayo>"
      "wiki_ja" ADV { next week } TIME @ADVL
```

Now each sentence has the word *Jumanne* with appropriate tags.

For further processing we need to put the result into sentence-per-line format (7).

(7)

```
( "<*profesa>" "*profesa" N TITLE { *professor } AN HUM @SUBJ ) (
"<*jumanne>" "jumanne" N 9/10-SG { *tuesday } CAP PROP-CAND ) (
"<*maghembe>" "*maghembe" PROPNAME SG { *maghembe } @SUBJ ) (
"<alijibu>" "jibu" V 1-SG3-SP VFIN NO-SP-GLOSS PAST z { answer }
SVO @FMAINVtr-OBJ> ) ( "<.$>" ".$" { .$ } **CLB )
( "<*nitakuja>" "ja" V 1-SG1-SP VFIN { *i } FUT:ta INFMARK z {
come } SV CAP @FMAINVintr ) ( "<huko>" "huko" ADV { on } @ADVL ) (
"<*jumanne>" "jumanne" N 9/10-SG { *tuesday } WEEK CAP @SUBJ ) (
```

```
"<wiki_ijayo>" "wiki_ja" ADV { next week } TIME @ADVL ) ( "<.$>"
".$" { .$ } **CLB )
```

The gloss of *Jumanne* in the first sentence will be replaced by the token. In the second sentence the original gloss is retained (8).

(8)
```
( "<*profesa>" "*profesa" N TITLE { *professor } AN HUM @SUBJ ) (
"<*jumanne>" PROPNAME { *jumanne } ) ( "<*maghembe>" "*maghembe"
PROPNAME SG { *maghembe } @SUBJ ) ( "<alijibu>" "jibu" V 1-SG3-SP
VFIN NO-SP-GLOSS PAST z { answer } SVO @FMAINVtr-OBJ> ) ( "<.$>"
".$" { .$ } **CLB )
( "<*nitakuja>" "ja" V 1-SG1-SP VFIN { *i } FUT:ta INFMARK z {
come } SV CAP @FMAINVintr ) ( "<huko>" "huko" ADV { on } @ADVL ) (
"<*jumanne>" "jumanne" N 9/10-SG { *tuesday } WEEK CAP @SUBJ ) (
"<wiki_ijayo>" "wiki_ja" ADV { next week } TIME @ADVL ) ( "<.$>"
".$" { .$ } **CLB )
```

Now the gloss of *Jumanne* in the first sentence is **jumanne*, and in the second sentence it is **tuesday*.

After further processing we get the translation as in (9)

(9)
Professor Jumanne Maghembe answered.
I will come on Tuesday next week.

3 The treatment of month names with week names

The month names are written with capital-initial letters in Swahili. Formally they behave like proper names. Because month names are hardly used in other meanings, such as person manes, they can be treated as proper names, which have to be translated. Because the target language (English in this case) does not have noun classes, we need not know what class concord for month names is used in Swahili. Therefore they can be treated as proper names without noun class assignment.

Consider example (10).

(10)
```
"<*nitakuja>"
      "ja" V 1-SG1-SP VFIN { *i } FUT:ta INFMARK z [ja] { come }
SV MONOSLB CAP @FMAINVintr
"<*jumanne>"
      "jumanne" N 9/10-SG { *tuesday } WEEK CAP @<P
"<*novemba>"
      "*novemba" N PROPNAME SG { *november } TIME @<P
"<mwaka>"
      "mwaka" N 3/4-SG { the } { year } TIME @<P
"<huu>"
      "huu" PRON DEM :hV 3-SG { this } @<NDEM
```

In (10) *Jumanne* is interpreted as a weekday and translated. The selection rule applies, because the tag TIME is found on the right two times, on *Novemba* and *mwaka*. Therefore we get the translation (11).

(11)
I will come on Tuesday November this year.

If wee use a more precise expression, such as in (12), we have the tag TIME on three words.

(12)
```
"<*nitakuja>"
      "ja" V 1-SG1-SP VFIN { *i } FUT:ta INFMARK z [ja] { come }
SV MONOSLB CAP @FMAINVintr
"<*jumanne>"
      "jumanne" N 9/10-SG { *tuesday } WEEK CAP @<P
"<tarehe>"
      "tarehe" PREP { on } TIME @ADVL
"<20>"
      "20" NUM { 20 }
"<*novemba>"
      "*novemba" N PROPNAME SG { *november } TIME @<P
"<mwaka>"
      "mwaka" N 3/4-SG { the } { year } TIME @<P
"<huu>"
      "huu" PRON DEM :hV 3-SG { this } @<NDEM
```

The full translation is in (13).

(13)
I will come on Tuesday on 20 November this year.

Let us see what happens, when we have *Jumanne* in both senses in the same sentence (14).

(14)
```
"<*profesa>"
      "*profesa" N TITLE { *professor } AN HUM @SUBJ
"<*jumanne>"
      "jumanne" N 9/10-SG { *tuesday } CAP PROP-CAND
"<*mwagembe>"
      "*mwagembe" PROPNAME SG { *mwagembe } @SUBJ
"<alisema>"
      "sema" V 1-SG3-SP VFIN NO-SP-GLOSS PAST z [sema] { say } SV
@FMAINVintr
"<kwamba>"
      "kwamba" CONJ { that } **CLB @CS
"<atakuja>"
      "ja" V 1-SG3-SP VFIN { he } FUT:ta INFMARK z [ja] { come }
SV MONOSLB @FMAINVintr
"<*jumanne>"
      "jumanne" N 9/10-SG { *tuesday } WEEK CAP @<P
"<tarehe>"
      "tarehe" PREP { on } TIME @ADVL
```

```
"<20>"
      "20" NUM { 20 }
"<*novemba>"
      "*novemba" N PROPNAME SG { *november } TIME @<P
"<mwaka>"
      "mwaka" N 3/4-SG { the } { year } TIME @<P
"<huu>"
      "huu" PRON DEM :hV 3-SG { this } @<NDEM
```

We see that in the disambiguated sentence (14) the word *Jumanne* appears first with the tag PROP-CAND and then later with the tag WEEK. When we process this further we get the translation (21).

(21)
Professor Jumanne Mwagembe said that he will come on Tuesday on 20 November this year.

Why was the first occurrence of *Jumanne* not translated as a week day, although it was followed by three words, each with the tag TIME? The reason is that according to the rule, scanning to the right is not allowed beyond the finite verb or clause boundary. Both of these conditions, *alisema* and *kwamba*, are found before the first occurrence of the tag TIME is encountered. Therefore the first occurrence of *Jumanne* is interpreted as a person name.

To make this even clearer, let us use a very short sentence (22)
.
(22)

```
"<*profesa>"
      "*profesa" N TITLE { *professor } AN HUM @SUBJ
"<*jumanne>"
      "jumanne" N 9/10-SG { *tuesday } CAP PROP-CAND
"<atakuja>"
      "ja" V 1-SG3-SP VFIN NO-SP-GLOSS FUT:ta INFMARK z [ja] {
come } SV MONOSLB @FMAINVintr
"<*jumanne>"
      "jumanne" N 9/10-SG { *tuesday } WEEK CAP @<P
"<wiki>"
      "wiki" N 9/10-SG { the } { week } TIME @<P
"<hii>"
      "hii" PRON DEM :hV 9-SG { this } @<NDEM
```

And the translation (23).

(23)
Professor Jumanne will come on Tuesday this week.

If we still shorten the sentence, we get the correct translation, but the criteria for disambiguation become more idiosyncratic. Consider the example (24).

(24)

```
"<*profesa>"
      "*profesa" N TITLE { *professor } AN HUM @SUBJ
"<*jumanne>"
```

```
      "jumanne" N 9/10-SG { *tuesday } CAP PROP-CAND
"<atakuja>"
      "ja" V 1-SG3-SP VFIN NO-SP-GLOSS FUT:ta INFMARK z [ja] {
come } SV MONOSLB @FMAINVintr
"<*jumanne>"
      "jumanne" N 9/10-SG { *tuesday } WEEK CAP @<P
```

And the translation (25).

(25)
Professor Jumanne will come on Tuesday.

In this case the disambiguation rule for the last occurrence of *Jumanne* relies on the type of verb (motion) on the left.

4 Discussion

In disambiguation it is a good practice to have a default interpretation for each word. Tis means that if no rule applies to the word, the default interpretation will be selected. Also in the cases we have discussed here, we could consider either interpretation as a default. It is likely that the weekday interpretation is more frequent than the person name interpretation. Therefore, the weekday interpretation could be considered as a default. If this is the case, we should find appropriate criteria for deciding when the word occurs as a person name. One could think that if the word is followed by a proper name (e.g. *Jumanne Maghambe*), it should be interpreted as a person name. Another criterion could be that if the word is preceded by a title, it is a person name (e.g. *Profesa Jumanne*). However, the person name can appear without any of these criteria, whereby the disambiguation becomes difficult.

One could also consider the following finite verb, such as *Jumanne alisema hivi*. This sentence is, however, fully ambiguous, because *Jumanne* can mean the weekday or person (*Jumanne said in this way* or *On Tuesday he said in this way*).

Perhaps it is better to use the likelihood criterion rather than an explicit rule in cases such as this. But which interpretation is more likely than the other? It is very hard to know without frequency data.

5 Conclusion

Proper names constitute a difficult problem in machine translation, because a number of word forms written with a capital-initial letter have also other interpretations. In this paper we have discussed weekday names, which may also be used as person names. We have discussed the criteria for writing disambiguation rules, and demonstrated part of them. Yet we have ended up with the conclusion that a small number of cases will be left without safe disambiguation rules. It is not known, however, how often such cases occur in practice.