Technical Reports in Language Technology
Report No 20, 20015
http://www.njas.helsinki.fi/salama

Salama Dictionary

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

Abstract

This is a brief description of how Salama Dictionary was compiled and how it can be used. Also critical comments on it are included.

Key Words: *dictionary compilation, multiword expressions, disambiguation.*

1 Construction and use of Salama Dictionary

Salama Dictionary was originally compiled in June 2007 on the basis of various corpus texts. The text was first analyzed and then processed so that each word got only one analysis. Disambiguation was carried out on the basis of context in sentence. This method results in fine-grained analyses. A lexical word may get more than one interpretation, depending on the context. Therefore a noun, which sometimes inflects according to class 5/6, in other times inflects according to class 9/10. Even the noun of the same class may have more than one interpretation.

This method of producing a dictionary makes it possible to get frequency information of the use of words. If needed, the numerical frequency information can be converted into more readable symbols. Frequency information helps in selecting the most appropriate word among synonyms.

Typical dictionary entries thus produced are:

(1)
```
[fanya] Verb [fanya] {do, act, commit, make, manufacture,
manipulate} Frequency:19687
[fanyia] Verb [fanya] {do, act, commit, make, manufacture,
manipulate} APPL Frequency:1360
[fanyiwa] Verb [fanya] {do, act, commit, make, manufacture,
manipulate} APPL PASS Frequency:755
[fanyiza] Verb [fanya] {compose, mend} CAUS Frequency:145
[fanyizia] Verb [fanya] {compose, mend} CAUS APPL Frequency:62
[fanyizwa] Verb [fanya] {compose, mend} CAUS PASS Frequency:42
```

Not only base forms of verbs are produced but also the extended forms. However, one has to note that the question of whether an extended verb form should be a separate entry or whether it should be merged with its parent form is not clear. Some fully lexicalized verb extensions can be a separate entry, and some others not. I have followed the

principle that all verbal extensions, except applicative, have been given a separate entry. However, there are several exceptions to this rule.

Another interesting feature is that the dictionary contains a large number of idioms, or more generally, multi-word expressions. Examples are below:

(2)
```
[fanya_hila] Verb IDIOM-V <IDIOM {betray} Frequency:6
[fanya_hima] Verb IDIOM-V <IDIOM {hurry} Frequency:14
[fanya_hisani] Verb IDIOM-V <IDIOM {do a favour} Frequency:1
[fanya_juu_chini] Verb IDIOM-V {do whatever possible} Frequency:11
[fanya_kazi] Verb IDIOM-V <IDIOM {work} Frequency:2958
[fanya_magendo] Verb IDIOM-V <IDIOM {share unlawful trade}
Frequency:3
[fanya_njama] Verb IDIOM-V <IDIOM {plot} Frequency:9
[fanya_shauri] Verb IDIOM-V <IDIOM {decide} Frequency:74
```

These multi-word expressions that contain a verb are not restricted to base forms only. Also extended forms are recognised:

(3)
```
[fanyia_dhihaka] Verb APPL IDIOM-V <IDIOM {mock} Frequency:6
[fanyia_hila] Verb APPL IDIOM-V <IDIOM {betray} Frequency:3
[fanyia_hisani] Verb APPL IDIOM-V <IDIOM {do a favour} Frequency:8
[fanyia_kazi] Verb APPL IDIOM-V <IDIOM {work} Frequency:318
[fanyia_shaka] Verb APPL IDIOM-V <IDIOM {doubt} Frequency:4
[fanyiana_hiana] Verb APPL REC {each other} IDIOM-V <IDIOM {make
an unjust deed} Frequency:1
[fanyiana_masihara] Verb APPL REC {each other} IDIOM-V <IDIOM
{mock} Frequency:1
[fanyika_kazi] Verb STAT IDIOM-V <IDIOM {work} Frequency:5
[fanyika_njama] Verb STAT IDIOM-V <IDIOM {plot} Frequency:1
[fanyisha_kazi] Verb CAUS IDIOM-V <IDIOM {work} Frequency:10
[fanyishwa_kazi] Verb CAUS PASS IDIOM-V <IDIOM {work} Frequency:17
[fanyiwa_dhihaka] Verb APPL PASS IDIOM-V <IDIOM {mock} Frequency:2
[fanyiwa_hiana] Verb APPL PASS IDIOM-V <IDIOM {make an unjust
deed} Frequency:1
[fanyiwa_kazi] Verb APPL PASS IDIOM-V <IDIOM {work} Frequency:67
[fanywa_arusi] Verb PASS IDIOM-V <IDIOM {celebrate a wedding}
Frequency:1
[fanywa_karamu] Verb PASS IDIOM-V <IDIOM {prepare a feast}
Frequency:1
[fanywa_kazi] Verb PASS IDIOM-V <IDIOM {work} Frequency:5
```

The dictionary also contains cross-references to structures, where the lexical entry to be described is part of the structure. For example:

(4)
```
[faini] Noun 9/10 {fine} Origin:ENG Frequency:221
```

```
[faini] taz. [piga_faini] Verb ACT IDIOM-V <IDIOM {fine}
Frequency:3
[faini] taz. [pigwa_faini] Verb ACT PASS IDIOM-V <IDIOM {fine}
Frequency:14
```

Perhaps the most interesting feature in the dictionary is that for each searched word the system produces one or more use examples, according to choice. Each example is also translated to English. Because the dictionary system was compiled computationally, the result includes inevitably also less ideal examples. Long sentences were also cut for saving space, and the translation suffers from incomplete sentences. However, the translation gives a fairly good picture of the example, although the style is not perfect.

The unbiased choice of examples was a complex process. Many entries have thousands of use examples. The process of selecting examples was as follows:

(1) Sentences were shuffled document-internally, so that the code of each document was retained.
(2) All example sentences were extracted for each lexical entry from all documents.
(3) To reduce the number of examples, three first examples were taken from each shuffled document.
(4) These sentences of the documents were joined together and shuffled so that no feature in the sentence affected the sentence order. It is assumed that by these procedures a true unbiased random choice was achieved.
(4) The lexical entries and example sentences were merged together, and examples were selected starting from the beginning of the example list under each lexical entry.

The number of examples can be selected. It may turn out that the user is not satisfied with the first example. The second or third example may be better. The variability of examples meets different user tastes.

When we enter the verb 'alifanya' and select the one-example option, we get:

(5)
$ echo alifanya | find-word1

```
[fanya] Verb [fanya] {do, act, commit, make, manufacture,
manipulate} Frequency:19687
[fanya]] <MIS> kwa vile anataka afanya [fanya] na yeye mapenzi kwa
kuzingatia (for so wants he/she does and he the favourite for
considering)
ANALYSIS: [fanya] Verb 1-SG3-SP VFIN {he} PAST [fanya] {do} SVO
```

If we use the option of three examples, we get:

(6)
$ echo alifanya | find-word3

```
[fanya] Verb [fanya] {do, act, commit, make, manufacture,
manipulate} Frequency:19687
```

```
[fanya]] <MIS> kwa vile anataka afanya [fanya] na yeye mapenzi kwa
kuzingatia (for so wants he/she does and he the favourite for
considering)
[fanya]] <MIS> Kufanya [fanya] maandalizi ya uanzishaji wa (To do
the preparations of the beginning of)
[fanya]] <ALA>, jioni hii itafanya [fanya] onyesho lake kwa mara
ya (this evening will do his/her/its exhibition to the time of)
ANALYSIS: [fanya] Verb 1-SG3-SP VFIN {he} PAST [fanya] {do} SVO
```

Here we have three examples, and perhaps one of them satisfies the needs of the user.

If we enter a multi-word expression 'alifanya kazi' with one example, we get:

(7)
$ echo alifanya kazi | find-word1

```
[fanya_kazi] Verb IDIOM-V <IDIOM {work} Frequency:2958
[fanya_kazi]] <ALA> wa jeshi wanajikuta wakifanya [fanya_kazi]
kazi sehemu moja. (of the army they meet selves they doing the
work the part one.)
ANALYSIS: [fanya_kazi] Verb 1-SG3-SP VFIN {he} PAST [fanya] SVO
Verb {work}
```

When we enter a multi-word expression 'alifanya hima' with two examples, we get:

(8)
$ echo alifanya hima | find-word2

```
[fanya_hima] Verb IDIOM-V <IDIOM {hurry} Frequency:14
[fanya_hima]] <NUR> ile ya Zanzibar zimetakiwa zifanye
[fanya_hima] hima kutatua mgogoro wa Zanzibar. (That of Zanzibar
have been wished they should do halt to solve the dispute of
Zanzibar.)
[fanya_hima]] <NIP> Alitahadharisha kuwa CCM isipofanya
[fanya_hima] hima kueleza sera zake na mafanikio (He/she was wary
that when CCM does not do halt to explain their policies and the
successes)
ANALYSIS: [fanya_hima] Verb 1-SG3-SP VFIN {he} PAST [fanya] SVO
Verb {hurry}
```

Again, if we want three examples, we get:

(9)
$ echo alifanya hima | find-word3

```
[fanya_hima] Verb IDIOM-V <IDIOM {hurry} Frequency:14
[fanya_hima]] <NUR> ile ya Zanzibar zimetakiwa zifanye
[fanya_hima] hima kutatua mgogoro wa Zanzibar. (That of Zanzibar
```

```
have been wished they should do halt to solve the dispute of
Zanzibar.)
[fanya_hima]] <NIP> Alitahadharisha kuwa CCM isipofanya
[fanya_hima] hima kueleza sera zake na mafanikio (He/she was wary
that when CCM does not do halt to explain their policies and the
successes)
[fanya_hima]] <NIP> Seif Shariff Hamad, kufanya [fanya_hima] hima
bila kukawia, kuwaomba radhi (Seif Shariff Hamad, to do halt
without being late, to ask the contentments)
ANALYSIS: [fanya_hima] Verb 1-SG3-SP VFIN {he} PAST [fanya] SVO
Verb {hurry}
```

Finally a multi-word expression 'alifaya juu chini' with three members:

(10)
$ echo alifanya juu chini | find-word3

```
[fanya_juu_chini] Verb IDIOM-V {do whatever possible} Frequency:11
[fanya_juu_chini]] <NUR> Pia walifanya [fanya_juu_chini] juu chini
kuzamisha meli na kuuwa Waamerika (Also they did on below to sink
the ship and to be it Americans)
[fanya_juu_chini]] <KIO> Popote siku_hizi kanisa linafanya
[fanya_juu_chini] juu chini ili kuwaelimisha wananchi kuhusu
wajibu (Anywhere these days the church does on below in order to
teach the citizens concerning the responsibility)
[fanya_juu_chini]] <KIO> Hao wazazi hufanya [fanya_juu_chini] juu
chini ili mradi mtoto wake amekwenda (These parents do on below so
that provided that his/her child has to gone)
ANALYSIS: [fanya_juu_chini] Verb 1-SG3-SP VFIN {he} PAST [fanya]
SVO {do whatever possible}
```

Note that it is possible to enter any inflected word form, provided that it returns the desired stem, for example:

(11)
$ echo wanaotufanyizia | find-word1

```
[fanyizia] Verb [fanyiza] {compose, mend} CAUS APPL Frequency:62
[fanyizia]] <MAJ> "Leo tunakufanyizia [fanyiza] kwa umezoea
kutuandika sana na ("today compose you for you have been used to
write us much and)
ANALYSIS: [fanyizia] Verb 2-PL3-SP VFIN PR:na 2-PL-SUB-REL {who}
2-PL1-OBJ OBJ {us} [fanyiza] {compose} SVO CAUS APPL
```

2 Known bugs

There is no guarantee that the example or examples under the found lexical entry are precisely examples of the entered word form. This is due to the fact that examples were retrieved on the basis of base form only, without taking into consideration other criteria, such as part-of-speech or morphological features.

The problem was partly solved by arranging the lexical entries so that the most obvious examples would be found. It is possible to take also other criteria into consideration for improving the accuracy of examples. This method has also problems, because it may easily turn out that no example is available for that particular word form. It is better to produce examples of the base form than leave without any output.

Perhaps it is possible to develop a system, where preference is on morphology-sensitive choice of examples, and if this fails, examples are given on the basis of more general criteria, such as base form only or with the combination of base form plus part-of-speech.