Tips for making queries in Helsinki Corpus of Swahili

Arvi Hurskainen

The corpus use manual of Lemmie2.0 applies also to the HCS. The general Lemmie2.0 manual can be opened by clicking the tab 'Manual' in the Menu Bar of the corpus interface. However, there are differences in the structure of different language corpora, and these differences must be taken into account when making queries. Below I describe how to construct queries with Swahili material.

The major deviation in the feature structure of Swahili is that many values (or tags) do not have a unique keyword (or feature name). This applies the feature category 'msd'. Examples below will give more clarification.

NOTE: A full list of tags used in HCS is in:
`http://www.aakkl.helsinki.fi/cameel/corpus/swatags.pdf`

There are four keywords (also called features) in the query system, each with a defined set of values.

wf	word-form
bf	base form
pos	part-of-speech
msd	morpho-syntactic description

The characteristics of each of them are described below.

Word-form

`[wf='aliacha']` wf = keyword denoting the word-form, that is, the surface form found in text. The word-form can be truncated using the Kleene star '*', e.g. `[wf='*acha']`. This query matches with all words that end in 'acha'. Correspondingly, the query `[wf='*ach*']` matches with any word that includes the string 'ach', e.g. 'aliacha', 'aliachisha'.

`[bf='acha']` bf = keyword denoting the base form, or lemma form, of the word. This form is approximately the same as found normally as a keyword in dictionaries.

Examples of base form types

Nouns:
mwalimu the singular form of the noun, also in cases where the noun is in plural in text, e.g. `[bf='mwalimu']`

Adjectives:
zuri the stem of the adjective, the class prefix stripped off, e.g. `[bf='zuri']`. Non-inflecting adjectives do not have a prefix, of course.

Numerals:
tatu the stem of the numeral, the class prefix stripped off, e.g. `[bf='tatu']`. Non-inflecting numerals do not have a prefix, of course.

Adverbs:

sana	the word as such, no affixes, e.g. `[bf='sana']`

Conjunctions:

ili	the word as such, no affixes, e.g. `[bf='ili']`

Prepositions:

katika the word as such, no affixes, e.g. `[bf='katika']`. Note that there are prepositions composed of two or three words. These are boud together, e.g. *baada_ya, kati_ya, badala_ya, kwa_upande_wa, kwa_niaba_ya, kwa_sababu_ya* etc. These have to be typed with underscore, e.g. `[bf='baada_ya']`.

Pronouns:

Demonstrative

huyu the word as such, e.g. `[bf='huyu']`. Demonstrative pronouns do inflect, but because it is difficult to define the base form in the way that the user can readily guess it, the base form of a demonstrative pronoun is implemented to be identical with the surface form. E.g. `[bf='huyu']`

Personal

yeye	the word as such, e.g. `[bf='yeye']`

Possessive

ake the stem of the possessive pronoun, the class prefix stripped off, e.g. `[bf='ake']`.

Verbs:

There are two methods for defining the base form of the verb. In one method, each extended verb is reduced to its shortest base form, and this form is treated as the base form also for extended verbs. In another method, each extended verb stem is treated as a base form. In HCS, the latter method was used. That is, whatever extensions the basic verb stem has, this extended stem is treated as the base form of the verb. Examples include:

`[bf='achia']`	verb with applicative extension
`[bf='achwa']`	verb with passive marker
`[bf='achisha']`	verb with causative extension
`[bf='achishwa']`	verb with causative extension and passive marker.

Note that the inflectional features such as the subjunctive marker (e.g. *achie*) and the negative marker (e.g. *achii*) do not affect the base form. The query `[bf='achia']` finds also these cases.

Part-of-speech categories

The keyword for part-of-speech is 'pos'. This feature has the following values:

`[pos='v']`	denotes verbs
`[pos='adj']`	denotes adjectives
`[pos='n']`	denotes nouns
`[pos='propname']`	denotes proper nouns and names

`[pos='adv']`	denotes adverbs
`[pos='num']`	denotes numerals
`[pos='conj']`	denotes conjunctions
`[pos='pron']`	denotes pronouns
`[pos='prep']`	denotes prepositions

Each value must be written precisely as described above. Truncation can be used where its use is safe.

Morpho-syntactic description

Whereas all the three features (or keywords) above have a single-word value, the feature 'msd' has frequently several values for each analyzed word. This affects the way how queries must be written.

Below is an example of how the analysis of the word-form *lilitokea* is in the corpus.

```
<w lemma="tokea" type="V" msd="5/6-SG-SP VFIN PAST SVO EXT: STAT
APPL :EXT" trans="emerge">lilitokea</w>[1]
```

Note that the feature names here are partly different than those used in queries. This does not matter, because they are interpreted, or translated, by the system. More important is to note that the feature name 'msd' has several values. If a tag inside the feature name 'msd' should be chosen as its value, the tag must be surrounded by the star '*', especially if it is not the first or last tag in the set. The star means that anything or nothing may be in its place.

For instance, if we want to search for the past tense forms of the verb *tokea,* the query should be as follows.

```
[bf='tokea' msd='*past*']
```

It is also possible to include more than one 'msd' value.

```
[bf='tokea' msd='5/6-SG-SP*' msd='*past*']
```

Note that in the tag '5/6-SG-SP', there is the star '*' only in the end, because this tag is the first one in the set. Putting the star also in the beginning would do no harm. And it is a good practice to do so, because the user does not always know which tag is the first or last in the set.

The more general query below finds all verbs that have the causative and applicative extensions.

```
[pos='v' msd='*caus*' msd='*appl*']
```

Sometimes tags may become mixed, for example when one tag includes the same sequence of characters as another longer one. Examples are 'inf' and 'infmark'. The query `[msd='*inf*']` maches both of these tags. If these should be matched separately, the queries should be `[msd='*inf *']` and `[msd='*infmark*']`. The tag 'inf' means the infinitive marker in

[1] The Lemmie2.0 interface does not make difference between upper and lower case letters. Therefore, in examples here only lower case letters have been used.

infinitive forms, e.g. *ku-soma*. The tag 'infmark' means the infinitive marker that is present in some finite forms, for example in mono-syllabic verbs, e.g. *ali-ku-la*.

Commenting out a tag

If we want to find verbs that have the causative but not the applicative extension, the following query does it.

```
[pos='v' msd='*caus*' msd!='*appl*']
```

More examples:
`[pos='pron' msd!='pers*']` finds pronouns excluding the personal pronouns.

`[bf='wa' pos!='v']` finds words with the base form 'wa' excluding the verb 'wa'.

`[pos='v' msd='*rel*' msd='*obj*']` finds verbs with the relative marker, but excludes those which have also the object marker.

Finding collocations

So far the examples have been on individual words. Queries on more than one word can also be constructed.

`[pos='v' bf='wa'] [msd='*cond:ki*']` finds the verb 'wa' followed by a verb with the conditional -*ki*- marker, e.g. *alikuwa akisoma*

`[pos='v' bf='wa'] [msd='*pr:na*']` finds the verb 'wa' followed by a verb with the present tense marker -*na*-, e.g. *alikuwa anasoma*

`[pos='v' bf='wa'] [msd='*perf:me*']` finds the verb 'wa' followed by a verb with the perfect tense marker -*me*-, e.g. *alikuwa amesoma*

`[pos='v' bf='wa' msd='*rel*'] [msd='*perf:me*']` finds the verb 'wa' with the relative marker followed by a verb with the perfect tense marker -*me*-, e.g. *aliyekuwa amesoma*

`[pos='v' msd='*appl*' msd='*pass*'] [pos='v' bf='wa']` finds verbs with applicative and passive extensions followed by the verb 'wa'.

A full list of tags used in HCS is in:
http://www.aakkl.helsinki.fi/cameel/corpus/swatags.pdf

Last modified: 19.10. 2009