Arvi Hurskainen

Constraint Grammar in Unconventional Use: Handling complex Swahili idioms and proverbs

Abstract

The paper discusses the method of handling idiomatic expressions, proverbs and other multi-word units in the Swahili language. The problem is approached from the viewpoint of machine translation, with the aim that the correct form in the target language (English) can be produced from various grammatical Swahili forms of these expressions, and that they can be kept apart from non-idiomatic uses of similar word sequences. The rules for identifying and isolating these structures were written with the Constraint Grammar Parser CG-2, and the rules for processing the result further towards the translation in the target language were implemented with Beta and Perl.

1. Introduction

The computer applications based on the theory of Constraint Grammar have been used primarily for disambiguating the results of morphological analysis, as well as for surface syntactic mapping (Karlsson & al. (eds.) 1995). In this paper I will show how the same formalism can be used for handling multiword constructions such as idiomatic expressions, sayings, proverbs. These expressions contain features that cannot be handled with conventional methods. For instance, the meaning of an idiomatic expression cannot be deduced on the basis of the semantic meaning of its constituent parts. Therefore, idioms should be isolated from the rest of text and treated as units of their own. Also proverbs contain problems in analysis, although they are quite different from those of the idioms: problems with proverbs are primarily connected with such features as the ungrammaticality of structures, the use of nonstandard words and word-forms, and the use of extinct or archaic words. Because the problems in handling these two categories of expressions and the solutions related to them are quite different, I shall discuss them separately below.

A Man of Measure
Festschrift in Honour of Fred Karlsson, pp. 397–406

2. Proverbs

The frequency of proverbs in Swahili[1] depends on the context of use. In colloquial speech, proverbs are frequent, and one finds people competing in the use of proverbs. One of the criteria for mastering the language is in fact considered to be the correct use of proverbs in speech. Not only everyday discussion is furnished with proverbs but an appropriate use of proverbs in public speeches is also a highly valued skill. A public address without the use of proverbs is considered substandard, and thus it will not get the desired attention.

The use of proverbs in written texts varies according to the type of text. In normal news texts they are rather rare. In literary texts, especially in fiction written by skilled writers, they are quite frequent. In any case, proverbs in written text are so frequent that they need to be treated separately in the analysis system.

2.1 Grammatical features of Swahili proverbs

Proverbs are often constructed according to the rules of the language. In such cases no special treatment would be needed to obtain the correct linguistic analysis. However, the number of deviant structures is so high that a separate treatment is motivated. And if a large part of proverbs require a special treatment, why not treat all of them in a similar way? This method has the additional advantage that by isolating the proverbs they can be manipulated as a distinct category, e.g. in information extraction. Especially useful is such special treatment in bilingual applications, where a corresponding assignment of the proverb in the target language is needed.

Proverbs have a fairly constant form. If their form were fully predictable, they could be isolated already in text form without analysis. However, because their surface forms have a number of deviations, it is more economical to identify proverbs on the basis of the analyzed word forms, where rule writing can be performed by making use of various degrees of generality.

It also is the case that a proverb often includes certain words as its constituent parts which do not occur anywhere else in the language, and thus are not listed in dictionaries. They can be analyzed by heuristic means,

[1] Swahili proverbs are found in current dictionaries and especially in the excellent inventory by Wamitila (2001).

which is an unreliable method and leaves the semantic contents unanalyzed. If such a word is found to be part of a proverb, it can be handled right away in this context, and it will not be encountered as a problem in the later phases of processing (Hurskainen 2004).

2.2 Procedure in handling proverbs

The isolation of proverbs was here implemented by using the Constraint Grammar Parser CG-2 (Karlsson 1995; Tapanainen 1996, 1999). The main phases of processing are described below.

The last constituent member of the proverb is marked using appropriate context constraints, and the structure of the proverb is defined by showing the number of preceding words that are part of the proverb. Also the default interpretation of the last word of the proverb is replaced with the interpretation of the entire proverb, in this case with the literal English translation (see (1)).

(1)

```
Liwike
        ”wika” V CAP SBJN 5/6-SG-SP VFIN { it } [wika] { crow , triumph } SV SVO
lisiwike
        ”wika” V SBJN 5/6-SG-SP VFIN { it } NEG [wika] { crow , triumph } SV SVO
kutakucha
        ”cha” <<PROVERB { Whether it (i.e. the cock) crows or not, the sun will rise }
        .$
```

Then each of the other constituent members of the proverb is marked by using the proverb “template” as a key. Its place in the structure is also made explicit by showing the distance to the beginning and end of the proverb. The original English glosses of each word are removed, as they have no function any longer (see (2)).

(2)

```
Liwike
        ”wika” PROVERB>>
lisiwike
        ”wika” PROVERB<>
kutakucha
        ”cha” <<PROVERB { Whether it (i.e. the cock) crows or not, the sun will rise. }
        .$
```

When the semantic meaning of the entire proverb has been attached to the last word of the proverb, the other members need no longer be considered in translation.

Now what follows depends on the application. Each of the constituent parts has still the token, the lemma, and the tag showing its location in the proverb. One can, for instance, retrieve from the text proverbs and their correct English translation. If the aim is to produce only an English translation, the proverb can be neatly translated, because its translation is already there and the glosses of other members of the proverb have been removed (see (3)).

(3)
Liwike lisiwike kutakucha { Whether it (i.e. the cock) crows or not, the sun will rise }

Because the rules for identifying proverbs can be written by using a degree of abstraction, some variation in surface form can be covered by a single rule. Consider the two variants of the proverb above (see (4)).

(4)
```
Liwike
        ”wika” PROVERB>>
lisiwike
        ”wika” PROVERB<>
kutakucha
        ”cha” <<PROVERB { Whether it (i.e. the cock) crows or not, the sun will rise. }
        .$
Uwike
        ”wika” PROVERB>>
usiwike
        ”wika” PROVERB<>
kutakucha
        ”cha” <<PROVERB { Whether it (i.e. the cock) crows or not, the sun will rise. }
        .$
```

When the rule uses the lemma forms instead of the surface forms as the criteria for identifying proverbs, both variants can be described with one rule.

3. Idiomatic expressions

Many types of multiword concepts[2] can be described in the morphological lexicon. This is especially the case when the constituent parts of the expression are in strict order and do not inflect. Minimal inflection, such as separate singular and plural forms, can still be described in the lexicon. There are, however, multi-word idioms that have a verb as a constituent part which can be used in a variety of inflected forms. Such expressions cannot be described in the lexicon, especially if the correct meaning of the expression needs to be preserved, as is the case in machine translation. Consider the example in (5).

(5)
Alizunguka
    ”zunguka” V CAP 1/2-SG3-SP VFIN { he/she } PAST [zungua] { go around , surround, revolve wander about , loiter } SV SVO STAT
mbuyu
    ”mbuyu” N 3/4-SG { baobab tree } .$

Although the sentence as such is intelligible, the expression is seldom used in this meaning. It is an idiom and means ‘He/she accepted a bribe’. Here we can write a rule that assigns the correct meaning to the last element of the expression (see (6)).

(6)
Alizunguka
    ”zunguka” V CAP 1/2-SG3-SP VFIN { he/she } PAST [zungua] { go around , surround, revolve , wander about , loiter } SV SVO STAT
mbuyu
    ”mbuyu” <IDIOM { accept a bribe } .$

When the lexical meaning of the expression is attached to the last element, the glosses of the verb become unnecessary and can be removed, but the morphological tags should be retained for further processing. This is demonstrated in (7).

[2] Especially Chuwa (1995) and Wamitila (1999) have greatly contributed to the inventory of Swahili idioms.

(7)
Alizunguka
”zunguka” V CAP 1/2-SG3-SP VFIN { he/she } PAST SV SVO STAT IDIOM-V>
mbuyu
”mbuyu” <IDIOM { accept a bribe } .$

The correct translation can be processed on the basis of the remaining information (see (8)).

(8)
He/she accepted a bribe.

The same method can be used in describing more complex idioms, as exemplified in (9). The last member of the expression is marked first.

(9)
Alipaka
”paka” V CAP 1/2-SG3-SP VFIN { he/she } PAST [paka] { smear , spread , apply } SV SVO
mafuta
”mafuta” N 6-PL { oil , animal fat , lard }
kwa
”kwa” PREP { with }
mgongo
”mgongo” N 3/4-SG { a/the } DER:o { back }
wa
”wa” GEN-CON 3/4-SG { of }
chupa
”chupa” <<<<<IDIOM { flatter , praise falsely } .$

Subsequently the other members of the expression are marked and their interpretation are removed, except the verb, which loses the gloss only and retains morphological tags. The lexical form of the gloss is attached to the last constituent, and the grammatical information is stored in the verb (see (10)).

(10)

```
Alipaka
	”paka” V CAP 1/2-SG3-SP VFIN { he/she } PAST SV SVO  IDIOM-V>>>>>
mafuta
	”mafuta” IDIOM<>>>>
kwa
	”kwa” IDIOM<<>>>
mgongo
	”mgongo” IDIOM<<<>>
wa
	”wa” IDIOM<<<<>
chupa
	”chupa” <<<<<IDIOM { flatter , praise falsely } .$
```

Because the idiomatic expression, regardless of the number of its constituents, has the function of a single verb, it should be treated as a verb. Therefore, it is useful to isolate it as a single unit, so that in subsequent processing it can be treated in the same way as the other words. An example of isolation, needed in machine translation, is in (11). Note that the token and lemma of each word, as well as the unnecessary members of the idiom, have been removed and only the verb with grammatical information as well as the last member of the idiom with the gloss have been retained.

(11)

```
( V CAP 1/2-SG3-SP VFIN { he/she } PAST SV SVO IDIOM-V>>>>>  <<<<<IDIOM
{ flatter , praise falsely } ) ( .$ )
```

On the basis of this information the correct translation can be processed (12).

(12)
He/she flattered.

There are also idioms where the verb is not the only constituent with varying inflected forms. Consider examples (13) and (14), where the form of the possessive pronoun depends on the form of the verb.

(13)

```
Nimepita
        ”pita” V CAP 1/2-SG1-SP VFIN { *i } PERF:me SV  IDIOM-V>>
na
        ”na” IDIOM<>
hamsini
        ”hamsini” <<IDIOM { mind only } N { business }
zangu
        ”angu” PRON POSS 9/10-PL SG1 { my , mine } .$
```

(14)

```
Tumepita
        ”pita” V CAP 1/2-PL1-SP VFIN { we } PERF:me SV  IDIOM-V>>
na
        ”na” IDIOM<>
hamsini
        ”hamsini” <<IDIOM { mind only } N { business }
zetu
        ”etu” PRON POSS 9/10-PL PL1 { our , ours } .$
```

In such cases the possessive pronoun should be left outside the idiom structure, whereby it retains its original interpretation. The normal rules for reordering the constituents of phrases will then take care of the correct word order in the target language. The final translations are in (15).

(15)

```
I have minded only my business.
We have minded only our business.
```

The method described above works for various verb constructions, as shown in (16, 17 and 18).

(16)

```
Sitapita
        ”pita” V CAP NEG SG1-SP VFIN { *i } FUT:ta SV  IDIOM-V>>
na
        ”na” IDIOM<>
hamsini
        ”hamsini” <<IDIOM { mind only } N { business }
zangu
        ”angu” PRON POSS 9/10-PL SG1 { my , mine } .$
I will not mind only my business.
```

(17)
Nisingepita
”pita” V CAP 1/2-SG1-SP VFIN { *i } COND-NEG:singe SV IDIOM-V>>
na
”na” IDIOM<>
hamsini
”hamsini” <<IDIOM { mind only } N { business }
zangu
”angu” PRON POSS 9/10-PL SG1 { my , mine } .$
I would not mind only my business.

(18)
Nisingalipita
”pita” V CAP 1/2-SG1-SP VFIN { *i } COND-NEG:singali SV IDIOM-V>>
na
”na” IDIOM<>
hamsini
”hamsini” <<IDIOM { mind only } N { business }
zangu
”angu” PRON POSS 9/10-PL SG1 { my , mine } .$
I would not have minded only my business.

4. Conclusion

Multi-word expressions which cannot be safely described in the morphological parser have to be isolated and handled as separate units in machine translation. We have shown how it is possible to design the necessary rule system with the Constraint Grammar Parser CG-2. In fact, about 2,000 proverbs and 2,200 *Swahili idioms* have so far been described with this method. The rules for describing multi-word expressions are fully integrated with the more conventional rules, such as the rules for disambiguation and surface syntactic mapping. The advantages of the method include the possibility of writing rules with various degrees of abstraction. By using abstraction, several variant cases can be covered with one rule, provided that reliability does not suffer by doing so. With more concrete rules the danger of misapplication of rules can be avoided.

References

Chuwa, Albina (1995) *Phraseological Units and Dictionary: The case of Swahili language*. Ph.D. Dissertation. University of Warsaw.

Hurskainen, Arvi (2004) Optimizing disambiguation in Swahili. In *Proceedings of the 20th International Conference on Computational Linguistics (COLING '04), Geneva, Switzerland, 23-27 August 2004*, pp. 254–260. Geneva: [International Conference on Computational Linguistics].

Karlsson, Fred (1995) Designing a parser for unrestricted text. In Karlsson & al. (eds.), pp. 1–40.

Karlsson, Fred & Atro Voutilainen & Juha Heikkilä & Arto Anttila (eds.) (1995) *Constraint Grammar: A language-independent system for parsing unrestricted text*. Natural Language Processing 4. Berlin & New York, NY: Mouton de Gruyter.

Tapanainen, Pasi (1996) *The Constraint Grammar Parser CG-2*. Publications of the Department of General Linguistics, University of Helsinki 27. Helsinki: University of Helsinki.

—— (1999) *Parsing in Two Frameworks: Finite-state and functional dependency grammar*. Ph.D. Dissertation, Department of General Linguistics, University of Helsinki.

Wamitila, Kyallo Wadi (1999) *Kamusi ya Misemo na Nahau*. Nairobi: Longhorn Publishers.

—— (2001) *Kamusi ya Methali*. Nairobi: Longhorn Publishers.

Contact information:

Arvi Hurskainen
Institute for Asian and African Studies
P.O. Box 59
FI-00014 University of Helsinki
Arvi(dot)Hurskainen(at)helsinki(dot)fi
http://www.aakkl.helsinki.fi