Disambiguation of morphological analysis in Bantu languages

Arvi Hurskainen
Department of Asian and African Studies
Box 13
00014 University of Helsinki
Finland
Arvi.Hurskainen@helsinki.fi

Abstract

The paper describes problems in disambiguating the morphological analysis of Bantu languages by using Swahili as a test language. The main factors of ambiguity in this language group can be traced to the noun class structure on one hand and to the bi-directional word-formation on the other. In analyzing word-forms, the system applied utilizes SWATWOL, a morphological parsing program based on two-level formalism. Disambiguation is carried out with the latest version (April 1996) of the Constraint Grammar Parser (CGP). Statistics on ambiguity are provided. Solutions for resolving different types of ambiguity are presented and they are demonstrated by examples from corpus text. Finally, statistics on the performance of the disambiguator are presented.

Introduction

There are five principal factors in Bantu languages which contribute to ambiguous analysis of word-forms. First, nouns are grouped into more than ten marked noun classes. The marking of these classes extends across the noun phrase, whereby the noun governs the choice of markers in dependent constituents. Second, verbs inflect stem-initially and mark the subject, object, and relative referent by prefixes, whereby the actual form of each prefix is governed by the noun class of the noun it refers to. In addition, verb derivation also adds to the complexity of verbal morphology. Third, reduplication is a productive phenomenon. Because its accurate description in lexicon is not possible, alternative ways in handling it are discussed. Fourth, the majority of Bantu languages have a tone system, but rarely this is indicated in writing. This adds to morphological ambiguity. Fifth, various semantic functions of word-forms are also a source of ambiguity.

In this paper I shall discuss the points one and two by using Swahili as a test language.

1 Morphological analysis

The morphological analysis of Swahili is carried out by SWATWOL, which is based on the two-level formalism (Koskenniemi 1983). The application of this formalism to Swahili has been under process since 1987, and it has now, after having been tested with a corpus of one million words, reached a mature phase with a recall of 99.8% in average running text, and precision of close to 100%. The performance of SWATWOL corresponds to what is reported of ENGTWOL, the morphological parser of English (Voutilainen et al 1992; Tapanainen and Järvinen 1994), and SWETWOL, the morphological analyzer of Swedish (Karlsson 1992).

SWATWOL uses a two-level rule system for describing morphophonological variation, as well as a lexicon with 288 sub-lexicons. Unlike in languages with right-branching word formation, where word roots can be grouped together into a root lexicon, here word roots have been divided into several sub-lexicons.

Because SWATWOL has been described in detail elsewhere (Hurskainen 1992), only a sketchy description of its parts is given here.

1.1 SWATWOL rules

Two-level rules have been written mainly for handling morhophonological processes, which occur principally in morpheme boundaries. Part of such processes take place also in verbal extensions, whereby the quality of the stem vowel(s) defines the surface form of the suffix. The total number of rules is 18, part of them being combined rules. An example of a combined rule:

```
U:w <=>      k _ :Vo ;
             t _ /: a: ;
```

Change lexical 'U' to surface 'w' iff there is 'k' on the left and a surface character belonging to the set 'Vo' on the right; or there is 't' on the left and a lexical diacritic '/' on the right followed by a lexical 'a'.

1.2 SWATWOL lexicon

SWATWOL lexicon is a tree, where the morphemes of Swahili are located so that each route from the root lexicon leads to a well-formed word-form.

The most complicated part of the lexicon is the description of verb-forms, which requires a total of 125 sub-lexicons. For describing verbs, there are a number of consecutive prefix and suffix 'slots', which may or may not be filled by morphemes. The verb root is in the middle, and verbal extensions used mainly for derivation are suffixed to the root.

A noun is composed of a class prefix and root. Noun roots are located in 22 separate sub-lexicons, and access to them is permitted from the corresponding class prefix(es). Adjectives are grouped according to whether they take class prefixes or not. Also numerals are grouped according to the same principle. The lexicon has a total of about 27,000 'words'.

Here is a simplified example of a sub-lexicon:

```
LEXICON M/MI
 mU M/MIr   "mU 3/4-SG N";
 mi M/MIr   "mU 3/4-PL N";
```

This is a sub-lexicon with the name 'M/MI' containing prefixes of the noun classes 3 and 4. Each entry may have three parts, but only the middle part is compulsory. In the first entry, 'mU' is the lexical representation of a morpheme, and 'M/MIr' is the name of the sub-lexicon where the processing will continue. The third part within quotes is the output string.

In constructing the lexicon, underspecification of analysis was avoided. Although it may be used for decreasing the number of ambiguous readings (cf. Karlsson 1992), it leaves ambiguity within readings themselves in the form of underspecification, and it has to be resolved later in any case.

2 Extent of morphological ambiguity

For the purposes of writing and testing disambiguation rules, a corpus of about 10,000 words of prose text was compiled (Corpus 1). The text was analyzed with SWATWOL, and the results in regard to ambiguity are given in Table 1.

Table 1: Number of readings of word-forms in Swahili test corpus (Corpus 1). N(r) = number of readings, N(t) = number of word-form tokens, % = percent of the total, cum-% = cumulative percentage

N(r)	N(t)	%	cum-%
1	4653	48.74	48.74
2	2061	21.59	70.33
3	871	9.12	79.55
4	1047	10.97	90.52
5	542	5.68	96.20
6	162	1.70	97.90
7	49	0.51	98.41
8	22	0.23	98.64
9	34	0.36	99.00
10	33	0.35	99.35
11 or more	72	0.75	100.00

As can be seen in Table 1, about half of word-form tokens in Swahili are at least two-ways ambiguous. About one fifth of tokens are precisely two-ways ambiguous, and the share of three-ways and four-ways ambiguous tokens is almost equal, about 10%. The share of five-ways ambiguous tokens is 5.68%, but the number of still more ambiguous tokens decreases drastically. There are word-forms with more than 20 readings, the largest number in the corpus being 60 readings.

If we compare these numbers with those in Table 2 we note significant differences and similarities. Table 2 was constructed exactly in the same manner as Table 1, only the source text being different. Whereas in Table 1 a corpus of running text (Corpus 1) was used, in Table 2 the source text was a list of unique word-forms (Corpus 2).

The number of word-forms with more than one reading is almost equal in both corpora, slightly over 50%. The percentages in Table 2 decrease rather systematically the more readings a word-form has. While there were more four-ways ambiguous word-forms (10.97%) than three-ways ones (9.12%) in Table 1, in Table 2 the numbers are as expected. The only unexpected result is the share of six-ways ambiguous words (3.44%), which is higher than the share of the five-ways ambiguous ones (2.94%). In Corpus 2, the high percentage of four-ways ambiguous readings found in Corpus 1 does not exist.

The ambiguity rate in Swahili is somewhat lower than in Swedish (60%, Berg 1978). It seems to correspond to that of English (Voutilainen et al 1992:5), although DeRose (1988) gives somewhat

Table 2: Number of readings of word-forms in Swahili list of unique word-forms (Corpus 2). N(r) = number of readings, N(t) = number of word-form tokens, % = percent of the total, cum-% = cumulative percentage

N(r)	N(t)	%	cum-%
1	4960	48.13	48.13
2	2294	23.99	72.12
3	1031	10.78	82.90
4	568	5.94	88.84
5	281	2.94	91.78
6	329	3.44	95.22
7	102	1.07	96.29
8	88	0.92	97.21
9	85	0.89	98.10
10	34	0.36	98.46
11 or more	148	1.54	100.00

lower figures, 11% for word-form types and 40% for word-form tokens. In Finnish the corresponding figures are still lower, 3.3% for word-form types and 11.2% for word-form tokens (Niemikorpi 1979).

While the reported ambiguity counted from word-form tokens is generally much higher than that counted from word-form types, in Swahili the difference is small. This is due to the fact that in addition to ambiguity found in several of the most common words, verb-forms are typically ambiguous, as are almost half of the nouns.

Karlsson (1994:23) suggests an inverse correlation between the number of unique word-forms and rate of ambiguity. Therefore, heavily inflecting languages would tend to produce unambiguous word-forms. Swahili does not seem to fully support this hypothesis, although the numbers in Table 1 and 2 are not directly comparable with results of other studies. In Swahili lexicon, underspecification was avoided which adds to ambiguity.

3 Disambiguation with Constraint Grammar Parser

Morphological disambiguation as well as syntactic mapping is carried out with Constraint Grammar Parser (CGP). Descriptions of its development phases are found in several publications (e.g. Karlsson 1990; Karlsson 1994a, 1994b; Karlsson et al 1994; Voutilainen et al 1992; Voutilainen and Tapanainen 1993; Tapanainen 1996). It sets off from the idea that rather than trying to write rules by pointing out the conditions necessary for the acceptance of a reading in an ambiguous case, it allows the writing of such rules that discard a certain reading as illegitimate. The rule system is typically a combination of deletion and selection rules.

The morphological analyzer SWATWOL was so designed that it would be ideal for further processing with CGP. The output of SWATWOL contains such information as part-of-speech features, features for adjectives, verbs, adverbs, nouns, numerals, and pronouns, as well as information on noun class marking (also zero marking) wherever it occurs, etc. In the present application also syntactic tags are included into the morphological lexicon as far as the marking can be done unambiguously. The syntactic mapping of context-sensitive word-forms is left to the CGP.

In order to simplify disambiguation, fixed phrases, idioms, multi-word prepositions and non-ambiguous collocations are joined together already in the preprocessing phase of the text (e.g. `mbele ya > mbele_ya` 'in front of'), and the same constructions are written into the lexicon with corresponding analysis.

3.1 Constraint Grammar rule formalism

The subsequent discussion of the Constraint Grammar Parser is based on the formalism of Tapanainen (1996). A detailed description of an earlier version of CGP is in Karlsson (1994b). The CGP rule file has the following sections (optional ones in parentheses):

```
DELIMITERS
(PREFERRED-TARGET)
(SETS)
(MAPPINGS)
CONSTRAINTS
...
END
```

In DELIMITERS, those tags are listed which mark the boundary of context conditions. If the rule system tries to remove all readings of a cohort, the target listed in the section PREFERRED-TARGET is the one which survives. SETS is a section where groups of tags are defined. Syntactic parsing is carried out with rules located under the heading MAPPINGS. CONSTRAINTS contains constraint rules with the following schema:

```
[WORDFORM] OPERATION (target)
  [(context condition(s))]
```

WORDFORM can be any surface word-form, for which a rule will be written. OPERATION may have two forms: REMOVE and SELECT. These are self-explanatory. In TARGET is defined the concrete morphological tag (or sequence of tags), to which the operation is applied. A target may be also a set, which is defined in the SETS

section. If the target is left without parentheses it is interpreted as a set. CONTEXT CONDITIONS is an optional part, but in most cases necessary. In it, conditions for the application of the rule are defined in detail. Context conditions are defined in relation to the target reading, which has the default position 0. Positive integers refer to the number of words to the right, and the negative ones to the left. In context conditions, reference can be made to any of the features or tags found in the unambiguous reading, e.g. (1C ADJ), or in the whole cohort, e.g. (1 ADJ). These references can be made either directly to a tag or indirectly through sets, which are defined in a special section (SETS) of the rule formalism.

Any context may also be negated by placing the key-word NOT to the beginning of the context clause. It is also possible to refer to more than one context in the same position.

If there is a need to define further conditions for a reading found by scanning (by using position markers *-1 or *1), the linking mechanism may be used. This can be done by adding the key-word LINK to the context, whereafter the new context follows. For example, the context condition (*-1 N LINK 1 PRON LINK 1 ADJ) reads: 'there is a noun (N) on the left followed by pronoun (PRON) followed by and adjective (ADJ)'.

3.2 Order of rules

The algorithm allows a sequential rule order. This can be done by grouping the rules into separate sections. The sequential order of rules within a section does not guarantee that the rules are applied in the order where they appear. The rules of the first section are applied first. Any number of consecutive sections can be used. There are presently four sections of constraint rules in the rule file. Certain types of rules should be applied first without giving a possibility to other, less clearly stated, rules to interfere. Typical of such first-level rules are those where disambiguation is done within a phrase structure. In intermediate sections there are rules which use larger structures for disambiguation. By first disambiguating noun phrases and genitive constructions, the use of otherwise too permissive rules becomes possible, when clear cases are already disambiguated. The disambiguation of verb-forms belongs to these middle levels. The risk of wrong interpretations decreases substantially by first disambiguating noun phrases and other smaller units.

The CGP of Swahili has presently a total of 656 rules in four different sections for disambiguation and 50 rules for syntactic mapping. So far about 600 hours have been used for writing and testing rules.

4 Disambiguation of a sample sentence

Below is a Swahili sample sentence after morphological analysis and after CG disambiguation. The sentence is:

Washiriki wa semina zote walitoka katika nchi za Afrika. (Participants of all seminars came from African countries.

Sample sentence 1 Sample sentence after morphological analysis with SWATWOL before disambiguation:

```
"<*washiriki>"
  "*shiriki" SBJN VFIN 1/2-PL2 OBJ V
  "*shiriki" SBJN VFIN 1/2-PL3 OBJ V
  "*shiriki" SBJN VFIN 1/2-PL3-SP V
  "*shiriki" 1/2-SG2-SP VFIN PR:a V
  "*shiriki" 3/4-SG-SP VFIN PR:a V
  "*shiriki" 11-SG-SP VFIN PR:a V
  "*shiriki" 1/2-PL3-SP VFIN PR:a V
  "*mshiriki" 1/2-PL N
"<wa>"
  "wa" SELFSTANDING SP
  "wa" 3/4-SG GEN-CON
  "wa" 11-SG GEN-CON
  "wa" 1/2-SG GEN-CON
  "wa" 1/2-PL GEN-CON
"<semina>"
  "semina" 9/10-0-SG N
  "semina" 9/10-0-PL N
"<zote>"
  "ote" 9/10-PL-SP PRON:ote
"<walitoka>"
  "toka" 1/2-SG2-SP VFIN PR:a 5/6-SG
   OBJ V SVO
  "toka" 3/4-SG-SP VFIN PR:a 5/6-SG
   OBJ V SVO
  "toka" 11-SG-SP VFIN PR:a 5/6-SG
   OBJ V SVO
  "toka" 1/2-PL3-SP VFIN PAST V SVO
  "toka" 1/2-PL3-SP VFIN PR:a 5/6-SG
   OBJ V SVO
"<katika>"
  "katika" IMP V SVO  STAT
  "tika" NARR-COLLOQ:ka-a VFIN V SVO STAT
  "tika" NARR-COLLOQ:ka-a VFIN V STAT
  "katika" PREPOS
"<nchi>"
  "nchi" 9/10-NI-SG N
  "nchi" 9/10-NI-PL N
"<za>"
  "za" 9/10-PL GEN-CON
"<*afrika>"
```

```
  "afrika" PROPN SG
"<$.>"
```

Sample sentence 1 after disambiguation with CGP

```
"<*washiriki>" S:816, 1099
  "*mshiriki" 1/2-PL N
"<wa>" S:412
  "wa" 1/2-PL GEN-CON
"<semina>" S:1433
  "semina" 9/10-0-PL N
"<zote>"
  "ote" 9/10-PL-SP PRON:ote
"<walitoka>" S:534
  "toka" 1/2-PL3-SP VFIN PAST V SVO
"<katika>" S:244
  "katika" PREP @ADVL
"<nchi>" S:1155
  "nchi" 9/10-NI-PL N
"<za>"
  "za" 9/10-PL GEN-CON
"<*afrika>"
  "afrika" PROPN SG
"<.$>"
```

The CG rules reduce the number of multiple readings so that optimally only one reading survives. Rule S:816 removes an object reading of the word-form. After that, a selection rule S:1099 is applied.

```
SELECT (1/2-PL N)
     (1 NCL-2 + GEN-CON) ;
```

Select noun reading of Ncl 1/2-PL if followed immediately by genitive connector belonging to the set NCL-2. This description is equal to the grammatical rule. Also other rules follow the same principle. E.g. the reading 1/2-PL GEN-CON is chosen for the analysis of *wa* on the basis of the Ncl of the preceding noun. The rule states:

```
"<wa>" SELECT (1/2-PL)
     (-1 NCL-2) ;
```

Select Ncl 1/2-PL of the word 'wa' if in the preceding cohort there is a feature belonging to the set NCL-2.

Although both *washiriki* and *wa* are initially ambiguous, and in rules the context reference does not extend beyond this pair of words, we get the correct result. This is because in both of the cohorts there is only one such reading which refers to the same noun class.

The word *semina* is both SG and PL, and the following pronoun *zote*, which has the PL reading, solves the problem. The word *nchi* is disambiguated with a rule relying on the Ncl of the following genitive connector (GEN-CON).

The word *katika* has four readings. The grammatically correct way of disambiguating it is by referring to the following word.

```
"<katika>" SELECT (PREPOS)
     (1 N OR INF OR PRON) ;
```

Select the reading PREPOS of "katika" if there is a noun or infinitive of a verb or pronoun in the following cohort.

5 Success rate and remaining problems of disambiguation

The CGP of Swahili was tested with two text corpora, which had not been used as test material in writing rules: E. Kezilahabi's novel *Mzingile* (22,984 word-form tokens), and a collection of newspaper texts from the weekly paper *Mzalendo*, 1994 (49,969 word-form tokens). Test results are in Table 3.

Table 3: Ambiguity after processing with the Swahili CGP. N(t) = number of word-form tokens, N(w) = number of unique word-forms, amb-(t) = ambiguity in tokens, amb-(w) = ambiguity in unique word-forms.

Ambiguity	Mzingile	Mzalendo
N(t)	22,984	49,968
N(w)	5,914	9,359
amb-(t)	1,837	2,463
%	7.99	4.93
amb-(w)	721	831
%	12.19	8.88

The parser performed best with newspaper texts, leaving ambiguity to 4.9% of tokens. Yet the overall result has to be considered promising, given that the parser is still under development and that the rules are almost solely grammar-based.

The most common types of ambiguity still remaining are: noun vs. adverb, adjective vs. adverb, noun vs. conjunction, verb (imperative) vs. noun, and verb (infinitive) vs. noun. Those are typically in such positions in a sentence that writing of reliable rules is difficult. A fairly large part of remaining ambiguity concerns genitive connectors *ya* and *wa*, and possessive pronouns. They are generally in positions where the governing noun is beyond the current clause or sentence boundary on the left. For such cases, the rule syntax should

allow the use of more distantly located information.

The vast majority of constraints are selection rules for resolving ambiguity based on homographic noun class agreement markers. It is possible to resolve most of this ambiguity by using contextual information.

Conclusion

The morphological analysis of Swahili tends to produce a comparatively large number of ambiguous readings. The noun class structure coupled with class agreement marking in dependent constituents, contributes significantly to ambiguity. The phenomenon is particularly evident in verb structures, where different sets of noun class markers add to the ambiguity of the same verb-form. It is assumed that the solutions suggested here apply also to other Bantu languages.

The ambiguity resolution is based on the Constraint Grammar formalism, which allows the use of grammatically motivated rules. The maximal context in the present application is a sentence, but there is a need for extending it over sentence boundaries. Constraint rules are grouped into sections, so that the most obvious cases are disambiguated first. A parser with only grammar-based rules disambiguates about 95% or Swahili word-forms from running text, which initially has about 50% of the tokens ambiguous. The remaining ambiguity is hard to resolve fully safely, but probabilistic and heuristic techniques are likely to still improve the performance.

References

Berg, Sture. 1978. *Olika lika ord. Svenskt homograflexikon.* [Different similar words. Dictionary of Swedish homographs.] Stockholm: Almqvist and Wiksell International.

DeRose, Sture. 1988. Grammatical Category Disambiguation by Statistical Optimization. *Computational Linguistics,* 14:31-39.

Hurskainen, Arvi. 1992. A Two-Level Computer Formalism for the Analysis of Bantu Morphology: An Application to Swahili. *Nordic Journal of African Studies* 1(1):87-122.

Karlsson, Fred. 1990. *Constraint Grammar as a framework for parsing running text.* In Hans Karlgren (ed.), *COLING-90. Papers presented to the 13th International Conference on Computational Linguistics.* volume 3, pp. 168-173, Helsinki, 1990.

Karlsson, Fred. 1992. SWETWOL: A comprehensive morphological analyzer for Swedish. *Nordic Journal of Linguistics,* 15:1-45.

Karlsson, Fred. 1994a. *Designing a parser for unrestricted text.* In Karlsson et al (ed.) *Constraint Grammar: A Language-Independent System for Parsing Unrestricted Text.* Mouton de Gruyter, Berlin, 1994. pp. 1-40.

Karlsson, Fred. 1994b. *The formalism and environment of Constraint Grammar Parsing.* In Karlsson et al (ed.) *Constraint Grammar: A Language-Independent System for Parsing Unrestricted Text.* Mouton de Gruyter, Berlin, 1994. pp. 41-88.

Karlsson, F., A. Voutilainen, J. Heikkilä, and A. Anttila (eds.). 1994. *Constraint Grammar: A Language-Independent System for Parsing Unrestricted Text.* Mouton de Gruyter, Berlin, 1994.

Koskenniemi, Kimmo. 1983. *Two-level morphology: A general computational model for word-form recognition and production. Publications No. 11.* Department of General Linguistics, University of Helsinki, 1983.

Niemikorpi, Antero. 1979. Automatic Data Processing in the Compilation of Word Lists. In Kaisa Häkkinen and Fred Karlsson (eds.) *Suomen kielitieteellisen yhdistyksen julkaisuja* [Publications of the Linguistic Association of Finland,] 2:117-126.

Tapanainen, Pasi. 1996. *The Constraint Grammar Parser CG-2.* Publications No. 27. Department of General Linguistics, University of Helsinki, (ISBN-951-45-7331-5).

Tapanainen, P. and Järvinen T. 1994. Syntactic analysis of natural language using linguistic rules and corpus-based patterns. In *COLING-94. Papers presented to the 15th International Conference on Computational Linguistics.* Vol. 1, pp. 629-634. Kyoto.

Voutilainen, A., J. Heikkilä, and A. Anttila. 1992. *Constraint Grammar of English - A Performance-Oriented Introduction.* Publications No. 21. Department of General Linguistics, University of Helsinki.

Voutilainen, A. and Tapanainen, P. 1993. Ambiguity resolution in a reductionistic parser. In *Proceedings of the Sixth Conference of the European Chapter of the Association for Computational Linguistics. EACL-93.* pp. 394-403, Utrecht, Netherlands, 1993.