Nordic Journal of African Studies 9(2): 60-84 (2000)

Tapping the Neglected Resource in Kiswahili Terminology: Automatic Compilation of the Domain-Specific Terms from Corpus

SELEMAN S. SEWANGI
University of Dar-es-Salaam, Tanzania

ABSTRACT

This paper is about corpus-based computational terminography. It highlights computational techniques and tools for compiling Kiswahili terms in the domain-particular texts written by field experts. At present, a reasonable number of domain-specific texts are available into Kiswahili. However, no efforts have been made so far to collect terms in such texts and integrate them in Kiswahili terminology. The absence of appropriate techniques and tools for carrying out the task should be the explanation for this lack of interest. But recent individual efforts in the development of computational tools for Kiswahili language, particularly by Arvi Hurskainen, at the University of Helsinki, paves the way for using computers for compiling the terms. This paper focuses on the application of Kiswahili-based computational tools developed at the University of Helsinki for describing and using corpus-based term formation patterns for the automatic compilation of Kiswahili terms. The presentation postulates the central role of the computer lexicon in carrying out such a task. The computer lexicon plays a central role in annotating a corpus for the description of the term formation patterns and for the compilation of terms from the text. The paper provides a test case to verify the applicability of the techniques and the tools in the actual compilation of the terms.

INTRODUCTION

When Tanganyika, the present Tanzania mainland, became independent in 1962, Kiswahili became the national language and hence assumed a number of official responsibilities. Accordingly, deliberate efforts were made to promote the language, including the development of its terminology. The official responsibility of developing the terminology was given to the language experts through the official organs, such as the National Swahili Council (BAKITA),[1] and through academic institutions, such as the Institute of Kiswahili Research (TUKI). At the same time, the users of the language, especially the field experts, embarked on unofficial promotion of the terminology whenever the need arose in their relevant fields. Unfortunately, the two parties have been in an uncompromising relation. The language experts considered terms which are

[1] BAKITA was established in 1967 by an Act of Parliament and TUKI, which is a Research Institute of the University of Dar-es- Salaam, was formed in 1974 by transforming the then East Africa Inter-Territorial Language Committee, which served Kenya, Uganda, Tanganyika, and Zanzibar.

developed by the field practitioners, i.e. the domain specialists, as incompatible with the traditional conventions of terminology work, particularly with the principles of terminology standardization. On the other hand, the practitioners despise most of the terms developed by the language experts as too strange and obscure for encoding concepts in their respective fields. In practical terms, the antagonism is reflected in two ways: the language experts look at the work of the practitioners not as a promotion but as a confusion of Kiswahili terminology (Massamba 1997: 89). As a result, the language experts see no reason to deal with terms that are produced by the practitioners. On the other hand, the practitioners resist most of the terms developed by the language experts and continue to coin and use their own terms as the demand arises (Mwansoko 1993: 185). Consequently, in more than thirty years of Kiswahili terminology work, no efforts have been made in the circles of the language experts to compile terms that are developed and used by the practitioners in their areas of specialisation. The official measuring of the growth rate of Kiswahili terminology is done on the basis of the terms developed by the language experts only and is so far very discouraging.[2].

This paper advocates the collection and inclusion of terms developed and used by the domain specialists in the promotion of Kiswahili terminology. It presents corpus-based computational techniques, with their implementation tools, for the compilation of such terms from the domain-specific texts in Kiswahili. It focuses especially on the techniques for compiling terms from domain-specific text corpora by the use of term formation patterns and a pattern matching a program. The coverage of the paper centres on four stages of the compilation: the collection of domain-specific texts and their conversion into a computer-readable form; the annotation of the texts and identification of terms in the text for describing term formation patterns, the representation of terms into term formation patterns and the use of the term formation patterns for the compilation of terms by the pattern matching program.

1. COLLECTION OF TEXTS IN KISWAHILI AND THEIR CONVERSION INTO A COMPUTER READABLE FORM

The domain-specific texts in Kiswahili are presently available in different sources, such as, those in the hands of the domain specialist writers, in bookstores, libraries and public offices. But it is hard to find such texts in a computer-readable form. Therefore, the compilation of the texts and their conversion into computer-readable form is a prerequisite for using the texts in computational applications. There are two computational techniques for carrying out this task: the keyboard entry technique which requires a computer, text-processing software and a typist, and the scanning technique

[2] According to Mulokozi (1991: 9) by 1988 only 15000 terms were ready, out of which 9000 were already standardized by BAKITA. However, Mulokozi does not specify the subject fields for the terms.

which relies on scanning hardware, text recognition software and text editor software, such as the Epsilon or the EMACS[3], which are used for editing errors in scanned text. The scanning method is better than the keyboard method because it is fast and not prone to human error. However this method requires texts that are orthographically good quality and it uses hardware and software which are expensive. Texts which are already in a computer readable form can be obtained by approaching the publishers, along with permission to use them. A machine-readable text selected to represent a particular variety of a language for particular research objectives is known as a text corpus.

2. USING TEXT CORPUS FOR TERM COMPILATION

The processing of a corpus text for linguistic investigation is determined by the intended purposes of the investigation and the computational tools available for performing the intended tasks in the investigation. Techniques for using a text corpus in computer-aided linguistic investigation can only be employed if there are appropriate computer tools for processing the corpus.

2.1 COMPUTER TOOLS FOR PROCESSING KISWAHILI CORPUS

The techniques that are presented in this paper are for identifying Kiswahili terms in a domain-specific corpus, representing the terms in term formation patterns and using the patterns to retrieve domain-specific terms. The implementation of such techniques is facilitated by the use of the Kiswahili-based computational tools that have been developed at the University of Helsinki. The tools include the text preprocessor, the two-level morphological analyser (SWATWOL), the Constraint Grammar morphological disambiguator (SWACGP), and the pattern matching program[4].

The Kiswahili morphological analyser (SWATWOL) is based on the two-level formalism (Koskenniemi 1983) and the Kiswahili morphological disambiguator (SWACGP) is based on the Constraint Grammar formalism (Karlsson et al. 1995; Tapanainen 1996). The two tools perform the task of adding descriptive information to words in a text-corpus. Such information is a prerequisite for using the text corpus in different linguistic investigations, such as the description of term formation patterns and

---

[3] Epsilon is a text editor for the IBM-PC whereas the EMACS is a text editor for UNIX.

[4] The Swahili-based text preprocessor, morphological analyser and morphological disambiguator have been developed by Arvi Hurskainen at the Department of Asian and African Studies at the University of Helsinki (see Hurskainen 1992: 87-122; 1996: 568-573), and the pattern matching program has been developed by Jussi Piitulainen at the Department of General Linguistics at the University of Helsinki.

the retrieval of terms from the text corpus. The process of adding such information to the text corpus is known as corpus annotation (Leech et al. 1997).

The annotation of Kiswahili text corpus with SWATWOL and SWACGP involves three stages: text preprocessing, morphological analysis and morphological disambiguation. The text preprocessing regulates the text corpus for morphological analysis and, to some extent, for morphological disambiguation. The regulations cover the setting of sentence delimiters in the text corpus as well as merging multi-word collocations into single orthographic word tokens by underscore '_'. For instance, the preprocessor fixes adverbial phrases, such as *'mara kwa mara', 'kila mara',* and *'moja kwa moja'* as *'mara_kwa_mara', 'kila_mara*' and *'moja_ kwa_moja'*. Moreover, the program normalises upper-case letters as a sequence consisting of '*' character and the corresponding lowercase letter, as *'Kilimanjaro'* to *'*kilimanjaro*'.

The processed text is then given as an input to the SWATWOL for morphological analysis. The SWATWOL has two components: a lexicon and rules (Hurskainen 1992: 87-122). The lexicon contains a full description of Kiswahili morphological patterns and morphological constraints plus more than 45,000 base form entries for Kiswahili vocabulary. Moreover, the lexicon contains various tags for parts of speech and inflectional and derivative categories. Additionally, the lexicon permits the addition of base form entries and of any other word-based information such as etymological, terminological and semantic information depending on the demands of the research objectives. The ability of the SWATWOL lexicon to handle full word-based information facilitates the representation of domain-specific terminological information in the lexicon. The information plays a key role in the implementation of the techniques that are proposed in this presentation.

The SWATWOL analyses Kiswahili text such as this one:

> *Magonjwa ya kuambukiza yanayotokana na kinyesi ni magonjwa gani*? "Contagious diseases which originate from faeces are which ones?"

and produces it in the following form:

```
"<*magonjwa>"
   "gonjwa" 5a/6-PL N
   "ugonjwa" 11/6-PL N

"<ya>"
   "ya" 3/4-PL GEN-CON
   "ya" 9/10-SG GEN-CON
   "ya" 5/6-PL GEN-CON
   "ya" 5/6-PL
```

```
"<kuambukiza>"
    "ambukiza" INF V SV SVO

"<yanayotokana>"
    "tokana" 5/6-PL-SP VFIN PR:na 3/4-PL REL V SV SVO STAT REC
    "tokana" 5/6-PL-SP VFIN PR:na 5/6-PL REL V SV SVO STAT REC
    "tokana" 5/6-PL-SP VFIN PR:na 9/10-SG REL V SV SVO STAT REC

"<na>"
    "na" CC @CC

"<kinyesi>"
    "kinyesi" 7/8-SG N
    "kinyesi" 9/10-0-SG N
    "kinyesi" 9/10-0-PL N

"<ni>"
    "ni" ADV:ni
    "ni" SG1

"<magonjwa>"
    "gonjwa" 5a/6-PL N
    "ugonjwa" 11/6-PL

"<gani>"
    "gani" INTERROG

"<?$>"
```

The textual words in the above analyses are surrounded by angle brackets '<>'. Below each of such words are lines which begin with base-forms of the textual words which are contained within inverted commas. Such lines represent the possible morphological analyses of the words. Each line of the analyses is a reading and, in most cases, SWATWOL assigns more than one reading to the analysed words. A word with its reading or readings form a cohort. A cohort which contains more than one reading represents morphological ambiguity. In other words the cohort contains all possible analyses of the word form regardless of the textual context of the word. The process of selecting the contextually appropriate analysis among those provided by the morphological analyser is known as morphological disambiguation. This task is performed by the Kiswahili-based morphological disambiguator, the SWACGP. The SWACGP applies context-based rules and, to a lesser extent, some guessing rules in determining the contextually appropriate analysis (Hurskainen 1996: 568-573). The SWACGP takes in the morphologically ambiguous output of the SWATWOL such as

that above, and by using context-based rules it produces the morphologically disambiguated text with each cohort containing the textual word and the only contextually appropriate reading. Below, we have the SWACGP output of the above SWATWOL analysis of the text:

Magonjwa ya kuambukiza yanayotokana na kinyesi ni magonjwa gani?

```
"<*magonjwa>"
    "ugonjwa" 11/6-PL N
"<ya>"
    "ya" 5/6-PL GEN-CON
"<kuambukiza>"
    "ambukiza" INF V SV SVO
"<yanayotokana>"
    "tokana" 5/6-PL-SP VFIN PR:na 5/6-PL REL V SV SVO STAT REC
"<na>"
    "na" CC @CC
"<kinyesi>"
    "kinyesi" 7/8-SG N
"<ni>"
    "ni" ADV:ni
"<magonjwa>"
    "ugonjwa" 11/6-PL N
"<gani>"
    "gani" INTERROG
"<?$>"
```

A text corpus is morphologically annotated when it is in the above form and can be used for different research purposes. The information that is assigned to each word of the text and the tags that represent it is obtained from the lexicon as represented by the developer of the lexicon.

2.2 TECHNIQUES FOR COMPILING TERMS FROM A TEXT CORPUS

The availability of the Kiswahili-based computational tools paves the way for devising techniques that can be implemented by the use of such tools to facilitate the compilation of terms from domain-specific texts in Kiswahili. The techniques which are proposed here are for handling the following tasks: first, the identification of terms in a domain-specific corpus for the representation of term formation patterns; second, the representation of the terms in domain-specific term formation patterns; and third, the use of the term formation patterns for the compilation of terms.

2.2.1 Identification of terms for representing term formation patterns

In the compilation of terms by the use of term formation patterns and the pattern matching program the first task is to identify terms in a domain-specific corpus for the representation of the term formation patterns. The technique for such identification divides the task into two phases: identification of single-word terms and identification of multi-word terms.

2.2.1.1 Identification of single-word terms

The presupposition behind the identification of single-word terms is that single-word terms form the core of multi-word terms. The multi-term words are extensions of single-word terms and they are formed during the process of concept specification by compounding. Moreover, the base forms of single-word terms form the nuclei for deriving other single-word terms by the process of affixation. Thus, the identification of the single-word terms forms the basis for the identification of the multi-word terms as well as of other single-word-terms that are derived from the base forms of the identified terms. This is facilitated by adding the identified single-word terms to the SWATWOL lexicon and defining their base forms as terms in a particular domain.

However, the identification of single-word terms in a text corpus is difficult because there are no structural criteria that can be used to separate term-words from non-term-words in the text. Additionally, words behave differently in different communicative settings where, with the exception of a few technical terms, a word could function as a term in one communicative setting and as an item of general vocabulary in another setting. Subsequently, a number of criteria have been proposed for the identification[5]. One of the criteria is the frequency occurrence which is applied in the statistic-based techniques for term compilation (Ahmed et al. 1994: 267-278; Yang 1986: 93-103). The frequency of occurrence is obtained by comparative method where the terminological status of a word is judged by its frequency of occurrence across a number of compared texts. However, this criterion discriminates genuine terms with low frequency of occurrence in the comparison. Another criterion that has been used is the linguistic signals. Pearson (1998: 129) argues that since terms represent generic concepts they should either be preceded by the indefinite article or not be preceded by any article at all. However this criterion is not compatible with languages which do not use articles, such as Kiswahili.

Concept representation is the traditional criterion for determining the terminological status of a word. By this criterion, a word qualifies as a term if it represents a concept in the respective subject field or communicative setting. Judging whether a word represents a concept or not is achieved by the use of knowledge of concepts and their

[5] A comprehensive deliberation on the techniques and criteria for identifying terms and non-term words in text corpora is given in Pearson (1998: 7-40).

linguistic representation in the relevant domain (Sager 1991). In other words, the identification of single-word terms should involve subject specialists and language experts. Our techniques for term compilation rely on this criterion.

When considered as names for concepts, terms belong to the part of speech category *noun*, including *verbal nominals*, simply because words for names are grammatically classified as *nouns*. The association of terms with the nominal category confined the processes of identifying single-word terms to nominal words only. Thus, the first task in the identification process should be to locate all words of the nominal category in the text-corpus. This should be done by first annotating the text corpus with the SWATWOL and the SWACGP, and then using the pattern matching program to compile all words of the nominal category in the annotated text. The pattern matching program should be given two input files: a pattern file with the ‘ N’ tag and the ‘INF’ tag patterns for matching the nouns and the infinitive verbs respectively, and an annotated text file from which the nominal words are to be compiled. The compiled nominal words should then be sorted out in order to determine the words which represent concepts in the subject field. This should be done manually on the basis of the knowledge of the subject-domain and of the language. When the sorting is done, the selected terms should be added to the SWATWOL lexicon and their base forms defined by a domain-specific tag as terms of the relevant field. For example, the ‘HC’ tag could be used for defining base form entries of health terms in the lexicon. The addition of single-word terms to the computer lexicon updates the lexicon with the corpus-based terminological information. The updated lexicon is necessary for the annotation of a text corpus to discover the multi-word terms and to retrieve terms in the text-corpus.

2.2.1.2 Identification of multi-word terms

When the text-corpus is annotated with the terminologically updated lexicon, all word forms derived from the base forms that are defined as terms in the lexicon are assigned the domain-specific tag. For example, the updated SWATWOL lexicon with health care terminological information that is represented by the ‘HC’ tag is used in the annotation of the following text:

“*Magonjwa ya kuambukiza yanayotokana na kinyesi ni magonjwa gani*”

All words in the annotated text that have been derived from the base forms which have been defined as terms in the lexicons are assigned the ‘HC’ tag as follows:

```
"<*magonjwa>"
    "ugonjwa" 11/6-PL N HC
"<ya>"
    "ya" 5/6-PL GEN-CON
"<kuambukiza>"
    "ambukiza" INF V SV SVO HC
```

```
"<yanayotokana>"
    "tokana" 5/6-PL-SP VFIN PR:na 5/6-PL REL V SV SVO STAT REC
"<na>"
    "na" CC @CC
"<kinyesi>"
    "kinyesi" 7/8-SG N HC
"<ni>"
    "ni" ADV:ni
"<magonjwa>"
    "ugonjwa" 11/6-PL N HC
"<gani>"
    "gani" INTERROG
"<?$>"
```

The words in the above text that are marked as health care terms with the 'HC' tag are:

magonjwa "diseases"
kuambukiza "infecting"
kinyesi "faeces"

The annotated text-corpus, like the one above, is the basis for the discovery of multi-word terms in the corpus. The text is also important for the representation of the terms into term formation patterns and for the retrieval of both single-word and multi-word terms.

The retrieval of single-word terms from the annotated text is carried out when the pattern matching program is given two input files: a pattern file with a single 'HC' tag pattern and an annotated text file. In the annotated text file, the pattern matching program matches the 'HC' tag with words in the cohorts which contain the 'HC' tag. This matching, however, overgenerates by matching non-nominal category words, especially verbs which have been derived from the base forms that are defined in the lexicon as terms. For instance, the verbs *anaambukiza* "he or she/it is infecting" and *anayeambukiza* "who infect" are wrongly picked as terms by the program because they have been derived from the base form *ambukiza*, "infect", from the term *kuambukiza* "infecting", which has been entered in the lexicon and defined as a term. This problem is solved by the use of a debugging program which removes the problematic words from the annotated text before the compilation of terms is carried out.

The discovery of multi-word terms is done manually by going through the annotated text corpus, particularly the cohorts which contain the domain-specific term tag such as the 'HC' tag in the above case, and then by checking whether words in such cohorts function as single-word terms or as part of multi-word terms. Where the word forms function as parts of multi-word terms, the collocations are picked as multi-word terms. To exemplify the process we have the annotated form of the following text:

Mengine ni magonjea ya minyoo kama safura, kichocho, tegu, minyoo, askari, ugonjwa wa ameba" "Others are worm diseases such as hookworm, bilharzia, tapeworm, ascaris, and amoeba disease"

The annotated form of the text looks as follows:

```
"<*mengine>"
    "ingine" 5/6-PL A-INFL ' other '
"<ni>"
    "ni" ADV:ni
"<magonjwa>"
    "ugonjwa" 11/6-PL N HC
"<ya>"
    "ya" 5/6-PL GEN-CON
"<minyoo>"
    "mnyoo" 3/4-PL N HC
"<kama_vile>"
    "kama_vile" ADV COLLOC @ADVL
"<safura>"
    "safura" 9/10-0-SG N AR HC
"<,>"
"<kichocho>"
    "kichocho" 9/10-0-SG N HC
"<,>"
"<tegu>"
    "tegu" 9/10-0-SG N HC
"<,>"
"<minyoo>"
    "mnyoo" 3/4-PL N HC
"<askari>"
    "askari" 9/6-0-SG N AR HC
"<,>"
"<ugonjwa>"
    "ugonjwa" 11/6-SG N HC
"<wa>"
    "wa" 11-SG GEN-CON
"<ameba>"
    "ameba" 9/10-0-SG N HC
```

The cohorts which contain the 'HC' tag in the above annotated text contain the following words:

magonjwa "diseases"
minyoo "worms"

safura	“hookworm”
kichocho	“bilharzia”
tegu	“tapeworm”
askari	“askaris”
ameba	“amoeba”

Among these words the word *ameba* forms part of the multi-word term collocation *ugonjwa wa ameba* “amoeba disease”, and hence the collocation represents a multi-word term.

After the selection of the multi-word term collocations from the annotated text-corpus, the next step is to represent the collocations as patterns of tag sequences.

2.2.3 Representing terms as domain-specific term formation patterns

In order to obtain the tags for representing the multi-word term collocations, the term collocations should be annotated with the terminologically updated SWATWOL lexicon and disambiguated by the SWACGP morphological disambiguator. The following are cases of collocations with their annotated forms.

maradhi ya ngozi	**“skin disease”**
ugonjwa wa uti wa mgongo	**“back borne disease”**
kutunga mimba nje ya mji wa mimba	**“fertilization outside the uterus”**

"<maradhi>"
"radhi" 5a/6-PL N HC
"<ya>"
"ya" 5/6-PL GEN-CON
"<ngozi>"
"ngozi" 9/10-NI-SG N HC
(maradhi ya ngozi **“skin diseases**”)
<,>
"<ugonjwa>"
"ugonjwa" 11/6-SG N HC
"<wa>"
"wa" 3/4-SG GEN-CON
"<uti>"
"uti" 11-SG N
"<wa>"
"wa" 3/4-SG GEN-CON
"<mgongo>"
"mgongo" 3/4-SG N HC
(ugonjwa wa uti wa mgongo **“back borne disease**”)
<,>

```
"<kutunga>"
    "tunga" INF V SV SVO
"<mimba>"
    "mimba" 9/10-0-SG N HC
"<nje_ya>"
    "nje_ya" PREP @ADVL
"<mji>"
    "mji" 3/4-SG N
"<wa>"
    "wa" 3/4-SG GEN-CON
"<mimba>"
    "mimba" 9/10-0-SG N HC
```

(kutunga mimba nje ya mji wa mimba "**fertilization outside the uterus**")

When the collocations are annotated, the next task is to select the appropriate tags for the representation of such collocation as tag sequences of term formation patterns.

2.2 3.1 Tags for representing term formation patterns

The tags for building up the term formation patterns are selected from cohorts of each word in the collocation. Selecting the tags is done through category representation. This is because the patterns are supposed to represent the structures of terms and not of individual term collocations. The category representation criterion requires tags for the patterns to represent the categories of the words rather than individual words. There are different kinds of category tags, such as part of speech tags, inflectional tags, terminological tags and etymological tags, depending on the types of annotations used in the text corpus. The tags for the part of speech categories, such as 'N' for the category *noun*, and 'GEN-CON' for the category *genitive connector*, are more general than tags for other categories, such as the inflectional tag *'3/4-SG'* for the category *noun class 3/4 singular*. Tags for the restricted categories, such as terminological tags, are more restrictive and are useful for representing subsets of the larger categories. For example the infinitive tag 'INF' represents the category of infinitive verbs which is a subset of the larger category 'verb'. The tags for domain-specific terms such as the "HC" tag for the health care domain have the role of delimiting term formation patterns to specific domains.

2.2.3.2 Coding the patterns

The selected tags for each annotated collocation are put in a sequence, beginning with the tag for the first word up to the tag for the last word in the collocation. The tags are linked together by the '+' character which is separated from the tags by a space. Each

sequence of tags is given a number where sequences of the same tags are assigned the same number. The following example demonstrates the way tags are represented as term formation patterns

"<shinikizo>"
"shinikizo" 5a/6-SG N HC
"<la>"
"la" 5/6-SG GEN-CON
"<damu>"
"damu" 9/10-0-SG N AR HC

(shinikizo la damu **“blood pressure”)**
HC + GEN-CON + HC “1"

"<uchungu>"
"uchungu" 11/6-SG N HC
"<wa>"
"wa" 11-SG GEN-CON
"<uzazi>"
"uzazi" 11-SG N DER:zi HC

(uchungu wa uzazi **“labour pain**”)
HC + GEN-CON + HC “1"

"<pima>"
"pima" 9/10-0-SG N
"<joto>"
"joto" 9/10-0-SG N HC

(pima joto **“thermometer**”)
N + HC “2"

"<magonjwa>"
"ugonjwa" 11/6-PL N HC
"<ya>"
"ya" 5/6-PL GEN-CON
"<via>"
"kia" 7/8-PL N
"<vya>"
"vya" 7/8-PL GEN-CON
"<uzazi>"
"uzazi" 11-SG N DER:zi HC

(magonjwa ya via vya uzazi **“reproductive organs diseases**”)
HC + GEN-CON + N + GEN-CON + HC “3"

In the above examples, we have four sequences of tags as follows:

HC + GEN-CO + HC “1" (shinikizo la damu “**blood pressure**”)
HC + GEN-CO + HC “1" (uchungu wa uzazi **“labour pain**”)
N + HC “2 (pima joto **“thermometer**”)
HC + GEN-CON + N + GEN-CON + HC “3" (magonjwa ya via vya uzazi **“reproductive organs diseases**”)

The sequences assigned number (1) have the same tags. Sequences with the same tags as these are counted as a single pattern, hence the sequences above are counted as three patterns and not four.

After the description of all patterns from the annotated collocations, the patterns are put into the single pattern file ready for being used in the pattern matching program for the compilation of terms from the annotated text corpus of domain which the patterns are based.

3. COMPILATION OF TERMS BY THE USE OF TERM FORMATION PATTERNS AND THE PATTERN MATCHING PROGRAM

In this section we demonstrate the actual application of the techniques in the compilation of Kiswahili health care terms. The compilation has been carried out using Kiswahili texts in the health care domain which have been written by health care specialists.

3.1 PREPARATION OF THE TEXT CORPUS

The Kiswahili texts in books and journals were collected, scanned and edited using the EMACS text editor. The total edited text, which consisted of 92,285 words, was saved in two text files named:

pattern-describing text file of 50, 877 words
pattern-testing text file of 92, 285 words.

The pattern describing text file was intended for the description of term formation patterns in the health care domain and the pattern testing text file was reserved for testing the performances of the described term formation patterns in the compilation of terms by the pattern matching program.

3.2 UPDATING THE SWATWOL LEXICON

For updating the SWATWOL lexicon, all possible single-word terms were collected from the two text files and added to the SWATWOL lexicons. This was done according to the following procedures: first, the text files were preprocessed and annotated with the SWATWOL analyser and the SWACGP morphological disambiguator. Second, the annotated files were given as input to the pattern matching program together with the pattern file which contained two patterns: the 'N' tag pattern and 'INF' tag pattern. The pattern matching and the sort programs uniquely retrieved 1724 nouns and 900 infinitive verbs from the annotated text. Third, a total of 540 single word terms were identified amongst the retrieved nouns and verbal nominals and added to the SWATWOL lexicon. In the lexicon, the base form entries for the identified single-word terms were defined as health care terms by using the 'HC' tag.

3.3 DESCRIPTION OF TERM FORMATION PATTERNS

For the description of term formation patterns, the pattern describing text file was first analysed with the updated SWATWOL lexicon and disambiguated by the SWACGP morphological disambiguator. As a result of this analysis and disambiguation, all readings in cohorts which contained word forms derived from the base forms that have been defined in the lexicon as health care terms contained the HC tag. This is exemplified by the following annotation of one of the sentences in the text:

> *Wakati wa ujauzito wanaweke wenye lishe duni wanakabiliwa na maradhi* " During pregnancy, women with poor diet are vulnerable to diseases"

The annotation form of this sentence is as follows:

```
"<*wakati>"
    "wakati" 11-SG N AR
"<wa>"
    "wa" 11-SG GEN-CON
"<ujauzito>"
    "ujauzito" 11-SG N HC
"<wanawake>"
    "mwanamke" 1/2-PL N
"<wenye>"
    "enye" 1/2-PL POSS
"<lishe>"
    "lishe" 9/10-0-SG N HC
```

```
"<duni>"
    "duni" AR A-UNINFL
"<wanakabiliwa>"
    "kabilia" 1/2-PL3-SP VFIN PR:na V AR SV SVO PASS
"<na>"
    "na" CC @CC
"<maradhi>"
    "maradhi" 5a/6-PL N AR HC
```

The words in the above cohorts with the readings which contain the HC tag are

ujauzito "**pregnancy**"
lishe "**nutrition**"
maradhi **"diseases"**

After the text file had been produced in the annotated form, all the collocations for the multi-word terms were selected in the text and put in a single text file then annotated with the updated SWATWOL lexicon and SWACGP morphological disambiguator. The following are three cases illustrating annotated collocations from the text:

uchungu wa uzazi **"labour pain"**
kujikaza na kuachia kwa misuli ya mji wa mimba **"uterus muscles contraction and relaxation** "
mfuko wa chakula **"stomach**"

The annotated forms of the collocations are:

```
"<uchungu>"
    "uchungu" 11/6-SG N HC
"<wa>"
    "wa" 11-SG GEN-CON
"<uzazi>"
    "uzazi" 11-SG N DER:zi HC
( uchungu wa uzazi)
"<kujikaza>"
    "kaza" INF REFL-SG OBJ V SV SVO HC
"<na>"
    "na" CC @CC
"<kuachia>"
    "achia" INF V SV SVO SVOO APPL
"<kwa>"
    "kwa" 15-SG GEN-CON
"<misuli>"
    "msuli" 3/4-PL N HC
```

```
"<ya>"
    "ya" 3/4-PL GEN-CON
"<mji>"
    "mji" 3/4-SG N ' toen, city '
"<wa>"
    "wa" 3/4-SG GEN-CON
"<mimba>"
    "mimba" 9/10-0-SG N HC
 ( kujikaza na kuachia kwa misuli ya mji wa mimba)

"<mfuko>"
    "mfuko" 3/4-SG N DER:o
"<wa>"
    "wa" 3/4-SG GEN-CON
"<chakula>"
    "chakula" 7/8-SG N HC
(mfuko wa chakula)
```

Then, the words which formed the collocations were represented with the appropriate category tags which were joined together with the '+' character to form the sequences of tags of term formation patterns. The total of 66 sequences of tags was described on the basis of the annotated collocations and placed in a single pattern file as follows:

```
HC "0"
INF + HC "1"
HC + HC "2"
HC + N "3"
HC + GEN-CON + HC "4"
HC + HC + HC "5"
N + GEN-CON + HC "6"
HC + A-INFL "7"
HC + A-UNINFL "7b"
HC + GEN-CON + ADV "8"
HC + GEN-CON + N "9"
INF + HC + LOC "10"
INF + GEN-CON + HC "11"
INF + HC + PREP + N "12"
INF + HC + PREP + HC "13"
HC + GEN-CON + INF + HC "14"
INF + N + GEN-CON + HC "15"
N + CARD + GEN-CON + HC "16"
N + PREP + HC + CC + HC "17"
HC + GEN-CON + N + CARD "18"
```

INF + GEN-CON + INF + HC "19"
HC + PREP + HC "20"
HC + GEN-CON + INF + N "21"
INF + HC + GEN-CON + ADV "22"
HC + GEN-CON + N + CC + N "23"
HC + HC + PREP + HC "24"
HC + A-INFL + PREP + HC "25"
HC + GEN-CON + N + HC "26"
HC + N + GEN-CON + HC "27"
HC + GEN-CON + N + A-INFL "28"
N + GEN-CON + HC + A-INFL "29"
N + A-INFL + GEN-CON + HC "30"
N + A-UNINFL + GEN-CON + HC "30b"
HC + GEN-CON + HC + A-INFL "31"
HC + POSS + HC + NUM-INFL "32"
HC + GEN-CON + HC + HC "33"
N + POSS + HC + GEN-CON + HC "34"
INF + HC + GEN-CON + INF "35"
HC + ADV + HC + GEN-CON + HC "36"
N + GEN-CON + HC + GEN-CON + N "37"
INF + HC + PREP + N + A-INFL "38"
HC + GEN-CON + N + GEN-CON + HC "39"
HC + GEN-CON + N + GEN-CON + N "39b"
HC + GEN-CON + HC + GEN-CON + HC "40"
HC + PREP + N + GEN-CON + HC "41"
N + GEN-CON + ADV + GEN-CON + HC "42"
INF + GEN-CON + N + GEN-CON + HC "43"
HC + A-INFL + GEN-CON + N "44"
HC + A-UNINFL + GEN-CON + N "44b"
N + GEN-CON + HC + GEN-CON + HC "45"
HC + PREP + HC + GEN-CON + HC "46"
HC + GEN-CON + HC + INF + ADV "47"
HC + GEN-CON + HC + GEN-CON + N "48"
N + GEN-CON + N + GEN-CON + HC "49"
INF + HC + PREP + N + GEN-CON + HC "50"
HC + HC + PREP + N + GEN-CON + HC "51"
N + GEN-CON + HC + INF + CC + HC "52"
N + CARD + GEN-CON + N + GEN-CON + HC "53"
HC + A-INFL + GEN-CON + HC + GEN-CON + HC "54"
N + GEN-CON + N + GEN-CON + HC + GEN-CON + N "55"
N + GEN-CON + N + GEN-CON + N + GEN-CON + HC "56"
INF + ADV + PREP + HC + GEN-CON + HC "57"
HC + GEN-CON + INF + N + GEN-CON + HC "58"

HC + A-INFL + GEN-CON + N + GEN-CON + HC "59"
HC + A-UNINFL + GEN-CON + N + GEN-CON + HC "59b"
HC + GEN-CON + HC + GEN-CON + N + GEN-CON + HC "60"
N + GEN-CON + N + GEN-CON + HC + GEN-CON + HC "61"
HC + GEN-CON + N + GEN-CON + N + GEN-CON + HC "62"
N + GEN-CON + HC + GEN-CON + HC + A-INFL + GEN-CON + HC "63"
HC + CC + INF + GEN-CON + HC + GEN-CON + N + GEN-CON + HC "64"
HC + GEN-CON + HC + GEN-CON + HC + GEN-CON + HC + A-INFL +
GEN-CON + HC "65"
HC + GEN-CON + N + GEN-CON + HC + GEN-CON + HC + A-INFL +
GEN-CON + HC "66"

3.4 PERFORMANCE OF THE PATTERNS IN THE PATTERN-MATCHING A PROGRAM

The effectiveness of the term formation patterns in the compilation of health care terms was tested by compiling terms from the pattern-testing text file. Before the compilation, the text file was analysed with the updated SWATWOL lexicon and disambiguated by the SWACGP morphological disambiguator. The compilation was carried out using the pattern matching program. The program was given two input files: a patten file and a text file. The pattern file contained all of the 66 term formation patterns and the text file contained the annotated pattern testing text. The program compiled the patterns in the input pattern file and produced two function files: a tokeniser and a pattern matcher. With the two files, the program matched the patterns with actual words in the annotated text file and retrieved the matched words and collocations as possible health care terms. The retrieved items were then sorted and counted by the sort program and copied into a text file. The counting of the retrieved items revealed that the program had retrieved a total of 6980 items as possible terms. The items included single words and collocations. Below we see a few cases of the retrieved items.

163 "0"	mimba	pregnancy
124 "0"	damu	blood
123 "0"	afya	health
117 "0"	ugonjwa	disease
117 "0"	maji	water
115 "0"	kujifungua	giving birth
113 "0"	watoto	children
100 "6"	wakati wa ujauzito	expecting period
96 "6"	mji wa mimba	uterus
24 "49"	shingo ya mji wa mimba	uterus neck
16 "2"	kuhara damu	“blood diarrhea”

14 "11"	kuharibika kwa mimba	aborting
7 "7b"	lishe duni	poor nutrition
7 "56"	mlango wa shingo ya mji wa mimba	uterus neck opening
7 "49"	ukuta wa mji wa mimba	uterus wall
7 "4"	kifafa cha mimba	eclampsia
7 "4"	kichwa cha uume	penis head
7 "4"	homa ya ini	hepatic fever
6 "6"	magonjwa ya uasherati	sexually transmitted diseases (STDs)
6 "27"	vidonda sehemu za siri	private part ulcers
6 "11"	*kuharibika kwa mimba	
6 "11"	kuharibika kwa mimba	aborting
6 "10"	kutokwa damu ukeni	vaginal bleeding
6 "0"	wanaozaliwa	
6 "0"	kujifungulia	giving birth
6 "0"	itasaidia	
6 "0"	inasaidia	
5 "4"	virusi vya *ukimwi	HIV virus
5 "4"	vimelea vya ukimwi	HIV virus
5 "3"	hospitali haraka	
5 "2"	utamtibu mgonjwa	
5 "10"	*kutokwa damu ukeni	
4 "3"	uume wake	
4 "3"	malale aina	
4 "3"	kliniki mara	
1 "9"	mbegu ya mwanaume	sperm
1 "9"	mbegu ya mwanamume	sperm
1 "9"	mbegu ya mwanamme	sperm

The items above consist of three elements: a number for the frequency of occurrence, a number which is surrounded by inverted comma for the term formation pattern, and a word form or a collocation.

In going through the list of the retrieved items, a number of problems in the catching were revealed. These were:

-the presence of non term words, especially verbs such as *anayejifungua* “who is delivering” and *itavimba* “it will swell” derived from the base forms that were defined in the lexicon as health care terms

-the presence of non term collocations, such as *malale aina* “type sleeping sickness” and *kliniki mara* “?” which were wrongly matched owing to unresolved ambiguities.

-the presence of additionally matched items owing to minor orthographic variations, such as *mbegu ya mwanaume,* “sperm” *mbegu ya manamume,* “sperm” and *mbegu ya*

mwanamme “sperm” and owing to the use of small and capital letters, as in *kuharibu mimba* “abortion” and *Kuharibu mimba* “abortion”

Two strategies were used in resolving the above problems: automatic debugging and manual sorting. First, a debugging program was used to remove the cohorts which contained problematic word forms in the annotated text before the compilation of the terms. For example, the program removed the unwanted verbs by removing the cohorts which contained tags for verb-based information such as tense, objects and subject agreements. In addition, the program removed from the annotated text file certain function words that were difficult to disambiguate, such as “*aina”* and “*mara*”. After the debugging, the compilation was repeated and the pattern-matching program retrieved 5,158 items. The manual sorting was then applied to remove the extra matched items from the retrieved list. After the sorting the final list of 3,098 corpus-based health care terms was obtained. Below are a few cases of such terms

117 "0"	ugonjwa	disease
115 "0"	kujifungua	giving birth
96 "6"	mji wa mimba	uterus
29 "8"	kondo la nyuma	placentae
29 "0"	kaswende	syphilis
27 "1"	kupata mimba	becoming pregnant
26 "0"	ukimwi	AIDS
26 "0"	mkojo	urine
26 "0"	mbegu	sperm/egg
25 "0"	kisonono	gonorrhea
24 "49"	shingo ya mji wa mimba	uterus neck
24 "0"	klamidia	chlamydia
22 "6"	via vya uzazi	reproductive organs
22 "0"	viluwiluwi	maggot
21 "4"	maziwa ya mama	mother milk
19 "0"	virusi	virus
19 "0"	vichocheo	enzymes
15 "7"	kifua kikuu	tuberculose
15 "6"	huduma za afya	health service
12 "9"	vimelea vya magonjwa	disease germs
12 "2"	mama mjamzito	pregnant woman
12 "1"	kupata hedhi	reaching menstruation
10 "4"	mfumo wa afya	health system
10 "4"	maumivu ya tumbo	abdomen pain
"4"	virusi vya ukimwi	HIV virus
8 "4"	ugonjwa wa upofu	blindness disease
8 "4"	mzunguko wa hedhi	menstruation circle

7 "56"	mlango wa shingo ya mji wa mimba	uterus neck opening
7 "49"	ukuta wa mji wa mimba	uterus wall
7 "4"	kifafa cha mimba	eclampsia
7 "4"	kichwa cha uume	penis head
5 "64"	kukaza na kuachia kwa misuli ya mji wa mimba	uterus muscle contraction and relaxation
5 "28"	uchungu wa muda mrefu	prolonged labour pain
4 "45"	tarehe ya matarajio ya kujifungua	expected date for delivery
4 "39b"	vimelea vya kundi la bakteria	bacteria germs
3 "62"	kansa ya shingo ya mji wa mimba	uterus neck cancer
3 "14"	upasuaji wa kutoa mtoto	delivery operation
2 "66"	ugonjwa wa kasoro ya maumbile ya chembechembe nyekundu za damu	circle cell
2 "62"	kansa ya mlango wa mji wa mimba	uterus opening cancer
2 "54"	maumivu makali ya kichwa cha uume	penis head severe pain
1 "9"	vimelea vya uambukizo	infectious germs
1 "9"	malaria ya msimu	seasonal malaria
1 "9"	malale ya rhodesia	Rhodesian trypanosomiasis
1 "9"	malale ya gambia	Gambian trypanosomiasis
1 "62"	dalili za uambukizo wa via vya uzazi	infection symptoms in reproductive organs
1 "59b"	kliniki maalum za magonjwa ya kujamiiana	STDs special clinics
1 "59"	maumivu makali ya eneo la tumbo	severe abdomen pain
1 "7"	utapiamlo mkali	severe malnutrition
1 "7"	upasuaji mdogo	minor surgery

It may be true that certain terms that have been compiled in this exercise are in the official collections of Kiswahili terminology. However, it could also be true that a good number of terms in this collection are not in the official lists, which suggests that the compilation, if thoroughly carried out in all domains, could greatly contribute to the growth of Kiswahili terminology.

4. CONCLUSION

This paper has presented the techniques for compiling Kiswahili terms from domain-specific corpora by the use of domain-specific term formation patterns and a pattern-matching program. The techniques have been formulated on the basis of computational tools that have been developed at the University of Helsinki. The paper has systematically presented the techniques and evaluated their performance in the actual compilation of Kiswahili health care terms.

The techniques that have been presented in this paper and their subsequent performance in the compilation of the Kiswahili health care terms suggest that there are possibilities of expanding Kiswahili terminology by the automatic compilation of terms that have been developed and used by subject specialists in their writings. However, such possibilities can only be realised if there is a cooperation between Kiswahili terminographers and developers of Kiswahili-based computational tools, such as the SWATWOL morphological analyser and the SWACGP morphological disambiguator.

REFERENCES

Ahmed, K., A. Davies, H. Fulford and M. Rogers 1994.
What is a term? The semi-automatic extraction of terms from text. In: *Translation Studies: an Interdiscipline,* Snell-Hornby, F. Pöchhacker and K. Kaindl (eds.), pp. 267-277. Amsterdam: John Benjamins Publishing Company.

Blommaert, J. (ed.) 1991.
Swahili Studies: Essays in Honour of Marcel Van Spaandonck. Ghent: Academic Press.

Hurskainen, A. 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. Helsinki: University of Helsinki Press.

1996 *Disambiguation of morphological analysis in Bantu languages.* **COLINGS -96: The 16th International Conference on Computational Linguistics. Proceedings vol 1,** pp. 568-573**.** Copenhagen: Center for Sprogteknologi.

Karlsson, F., A. Voutilainen, J. Heikkilä and A. Anttila (eds.) 1995.
Constraint Grammar: A language-Independent System for Parsing Unrestricted Text. Berlin & New York: Mounton de Gruyter.

Koskenniemi, K. 1983.
Two-level Morphology: A General Computational Model for Word-Form Recognition and Analysis. Publications of the Department of General Linguistics, University of Helsinki, No 11. Helsinki: University of Helsinki, Department of General Linguistics.

Leech, G., R. Garside and T. McEnry (eds.) 1997.
Corpus Annotation. London & New York. Longman.

Massamba, D. P. B. 1997.
Problems in terminology development: The case of Tanzania. **Kiswahili. Journal of the Institute of Kiswahili Research,** pp. 86-98. Dar es Salaam: University of Dar-es Salaam.

Mulokozi, M. M. 1991.
English versus Kiswahili in Tanzania secondary education. In: *Swahili Studies: Essays in Honour of Marcel Van Spaandonck*, Jan Blommaert (ed.), pp. 7-16. Ghent: Academic Press.

Mwansoko, H. J. M. 1993.
Acceptability of terms: The case of the Swahili linguistics and literature terms in Tanzania. **Multilingua** 12-2: 177-188. Berlin: Walter de Gruyter.

Pearson, J. 1998.
Terms in Context. Amsterdam/Philadelphia: Benjamins Publishing Company.

Sager, J. C. 1990.
Practical Course in Terminology Processing. Amsterdam/Philadelphia: Benjamins Publishing Company.

Snell-Hornby, F. Pöchhacker and K. Kaindl (eds.) 1994.
Translation Studies: an Interdiscipline. Amsterdam: John Benjamins Publishing Company.

Tapanainen, P. 1996.
The Constraint Grammar Parser CG-2. Publications of the Department of General Linguistics, University of Helsinki, No.27. Helsinki: University of Helsinki, Department of General Linguistics.

Yang, H. Z. 1986.
A new technique for identifying scientific and technical terms and describing science texts. **Literary and Linguistic Computing,** 1(2): 93-103. Oxford: Oxford University Press.