[*Mechanical Translation and Computational Linguistics*, vol.11, nos.1 and 2, March and June 1968]

A Note on the Translation of Swahili into English

by David Woodhouse, La Trobe University, Bundoora, Victoria, Australia

Some features of the morphology of Swahili are discussed from the point of view of mechanizing a dictionary. A preliminary program is described.

1. Basic Features of the Swahili Language

To the best of my knowledge, no work has previously been carried out on the mechanical translation of any Bantu language. This note is therefore a first suggestion of a possible basis for a scheme for the mechanical translation of Swahili into English.

Swahili, in common with other Bantu languages, makes great use of prefixes. This is its most distinctive feature when compared with European languages. All agreements between adjectives, nouns, and verbs are shown by means of prefixes. There are prefixes for the subject and object of a verb and for the verb tense. Negation of a verb is also shown by means of prefixes. Suffixes are also used, but a lot of Swahili can be spoken without using them. Suffixes are used to show motion to or from a place and, apart from this, are used almost exclusively in modifying the form of verbs. The passive, causative, prepositional, reciprocal, subjunctive, plural imperative, and some singular imperative forms are all constructed by adding a suffix to the verb stem. As is usually the case, addition of a suffix often causes modification of the stem itself. For example, the passive form of a verb ending with the letter *a* is made by changing the final *a* to *wa,* as in *kuandika* ("to write") and *kuandikwa* ("to be written"). However, *kununua* ("to buy") gives rise to *kununuliwa* ("to be bought"). Prefixes, on the other hand, are added with no amendment to the verb stem, and I see this as one of the reasons why the strong reliance on prefixes will make Swahili reasonably susceptible to mechanical translation. Other advantages of the prefix structure are:

1. There is less need for context-dependent analysis. For example, if the present tense of the verb "run" is recognized in English, one still does not know the final form of the word: it could be "they run" or "he runs." In Swahili, however, no such distinction is made:

wa-na-kimbea,
a-na-kimbea.

(*Wa* means "they"; *a* means "he"; *na* denotes the present tense; *kimbea* is the verb stem, meaning "run." The hyphens are not part of the Swahili word but are inserted for clarity.)

2. While a noun or adjective takes only one prefix at a time, a verb stem may have several prefixes concatenated with it. This usually entails no amendments to the prefixes or stem. It also means that many related parts of the sentence are joined in the same word. Thus, for example, "he will buy it" becomes

a-ta-ki-nunua.

(*Ta* denotes the future tense; *ki* denotes "it"; *nunua* means "buy.") Thus, by translating one word, a large part of the sentence has been dealt with. Furthermore, the subject, object, and tense indicator of the verb have all been obtained without searching the rest of the sentence.

3. All the above may be used without parsing. When we come to parsing, it is of great assistance that adjectives, nouns, and verbs must agree.

Wa-toto wa-zuri wa-na-kimbea.
"Good children are running."

Toto is the stem of the word for "child"; *zuri,* the stem of the word for "good.")

M-toto m-zuri a-na-kimbea.
"A good child is running."

(Note that adjectives follow their nouns and that there are no articles.)

There are eight different classes of nouns. Each has its own prefixes for showing singular and plural, and corresponding prefixes to attach to adjectives and verbs. For example, the prefixes for the class to which *-toto* belongs are:

	Singular	Plural
Noun	*m*	*wa*
Adjective.....	*m*	*wa*
Verb	*a*	*wa*

Another class has the following table:

	Singular	Plural
Noun	*u*	*n*
Adjective.....	*m*	*n*
Verb	*u*	*zi*

and so on.

Unfortunately, not all the prefixes are unique in meaning. *Ku,* for example, can mean "you" in the singular as the object of a verb and can also denote the in-

finitive. These ambiguities can be resolved, without too much difficulty, by considering the combination of prefixes in which the prefix in question occurs.

Suffixes differ from prefixes in two respects: (1) As already exemplified, suffixes can cause modification of the word stem. (2) In all but one case, only one suffix is used at a time. The exceptional case is supplied by two particular suffixes (*e* and *ni*) which can occur together (as *eni).* This may be considered as giving rise to another single suffix, namely, the concatenation (*eni*) of the two individual suffixes. We may then write the translation program as if, without exception, only one suffix is used at any one time.

These differences make it more efficient to deal quite differently with prefixes and suffixes. We note in passing that a disadvantage of Swahili is the absence of articles. Some work must be done on this problem (paralleling [1]) to determine whether there are word patterns which are indicative of the need to insert an article, and of which article to insert.

2. Structure of the Translation Scheme

Three dictionaries are envisaged: a stem dictionary, a prefix dictionary, and a suffix dictionary. If one were dealing with suffixes only (rather than with suffixes and prefixes), the appropriate procedure would clearly be as follows: If no match is found in the stem dictionary for a source-language word, the last letter is elided, and a match sought for the truncated word. This elision and comparison is continued until the first few letters of the original word are found as an entry in the stem dictionary. Thus, we know that, given any input string (word) of n letters, either (1) there is some integer $m \leq n$ such that the first m letters of the input word appear as an entry in the stem dictionary, or (2) no such m exists, and the word is unrecognizable by this dictionary. Since we wish to permit recognition of prefixes, however, with these entered in a separate dictionary, we have a third possibility: (3) there are integers $r, s,$ $0 < r \leq s \leq n$ such that letters r to s inclusive of the input word appear as an entry in the stem dictionary. We no longer have a fixed base (the beginning of the word), and we have introduced much more freedom, and many more subsets of each input string to be checked.

Furthermore, we must guard against faulty recognitions. If "anti" were an entry in the prefix dictionary, we should try to remove this prefix from the beginning of a word whenever possible—but must not "recognize" it in the word "antique," for example. My suggestion for Swahili translation deals with this difficulty, as follows.

A word is taken from the incoming source text, and attempts are made to recognize prefixes and suffixes. All prefixes have one or two letters, and the two-letter ones are recognized first, in an attempt to prevent spurious recognitions. If the first two letters are the same as an entry in the prefix dictionary, a note is made of the prefix, these two letters are dropped from the word, and the third and fourth letters are compared. When no more two-letter prefixes are found, a search is made for one-letter ones. If one is found, it is noted, the letter is dropped from the word, and a search is made for two-letter prefixes again. When no more prefixes can be found, we have some recognized prefixes, the remainder of the word being regarded as the stem. The stem dictionary is now searched for this stem. If it is found, the associated meaning, and the meanings of the recognized prefixes, are printed out, and the program moves to the next word of the source text. If it is not found, however, we should not immediately assume that the word is unknown to the dictionary (see the above comment on "antique").

We now *replace* the prefixes, one by one, in all possible (order-preserving) combinations. Thus, we replace the last prefix and try to recognize the resulting stem. If we are unsuccessful, we replace the next prefix, and so on. If all the prefixes are replaced with no recognition taking place, we move to consideration of suffixes.

One suffix may be considered as a complete addition to the word it modifies, namely, *ni. Nyumba* means "house"; *nyumbani* means "to the house" (or "at the house," or "from the house," depending on context). Most other suffixes are applied to verbs.

Most verbs end with the letter *a.* (Some verbs, of Arabic origin, end in *i, u,* or *e.* We have not dealt with these, but the necessary extension is not difficult.) In a Swahili-English dictionary, the verb "to buy" is entered as *nunua* (or *kununua)* and the noun "child" as *mtoto.* In our stem dictionary, however, we enter the stem *toto* and the "normal form" *nunua,* rather than the stem *nunu.* This is because the singular and plural forms *mtoto* and *watoto* appear with comparable frequency. It is therefore more efficient always to search for the stem *toto,* and then check the prefix for number. In the case of verb forms, however, the active voice, in unmodified form, occurs far more frequently than any of the other forms, such as passive, imperative, reciprocal, and so on. It is therefore more efficient to search first for the basic form. If no recognition takes place, we may then check for suffixes. This takes place as follows. If a final *e* is found, we may suppose the word to be a verb in imperative or subjunctive mood, replace the *e* by *a,* and check the resulting word to see if it is a verb in unmodified form. If the word does not end in *e,* we look for other verb endings (such as *ana* [reciprocal], *liwa* [passive]) and, whenever one is recognized, replace it by *a* and check the resulting word. This manner of dealing with verb suffixes clearly differs from the manner of dealing with prefixes.

3. The Program

The scheme as described above has so far been implemented in FORTRAN on ICL 1900 series computers. To use a scientific language for this purpose seems ludi-

crous, but there is a good practical reason. If a program to translate Swahili into English is to be useful (rather than purely academic research), it must be usable in Tanzania. Until recently, the only computers available in Tanzania were smaller processors from ICL's 1900 series, on which no list-processing language has been implemented. In order to develop this project, it had to be made to fit the local situation.

So far, only the basic idea, described above, has been implemented as a word-for-word dictionary lookup. No parsing of the input string or restructuring of the output string takes place. Only simple sentences (not involving subordinate clauses) have been translated.

The program accepts input in a form which may easily be prepared by a typist.

4. Results

Working with 28, 12, and 230 entries in the prefix, suffix, and stem dictionaries, respectively, the results obtained have been encouraging, although not faultless. For example,

a-li-amkwa

means

"he was awoken."

(*A* means "he" or "she"; *li* denotes the past tense; *amkwa* is the verb stem meaning "be awoken.") The program translated this as

He/She Past He/She Sing To/By/With/For.

Clearly, besides the correct recognition of prefixes *a* and *li,* prefixes *a* and *m* (denoting a reference to a personal noun in the singular) have been spuriously recognized in *amkwa,* because the preposition *kwa* is entered in the stem dictionary. However, all such erroneous translations encountered so far could be avoided by simple checks on allowable sequences of prefixes.

Much, however, still remains to be done if the English reader is not to have to use great mental agility to construe the computer output. The next major step must be to implement some automatic parsing of the Swahili input.

Received January 28, 1970

References

1. Martins, G. P. "Preliminary Report on the Insertion of English Articles in Russian-English MT Output." *Mechanical Translation,* vol. 8, no. 1 (August 1964).