This is to certify that the

thesis entitled

WORD FREQUENCY STUDY AND MORPHOLOGICAL ANALYSIS
OF TWO BANTU LANGUAGES

presented by

Chege John Githiora

has been accepted towards fulfillment
of the requirements for

Master of Arts degree in Linguistics

Department of Linguistics, Germanic, Slavic
Asian and African Languages

Major professor

Date 2 April 1993

WORD FREQUENCY STUDY AND MORPHOLOGICAL ANALYSIS OF TWO BANTU LANGUAGES

By

Chege John Githiora

A THESIS

Submitted to
Michigan State University
In partial fulfillment for the degree of

MASTER OF ARTS

Department of Linguistics, Germanic, Slavic, Asian and African Languages

1993

ABSTRACT

WORD FREQUENCY STUDY AND MORPHOLOGICAL ANALYSIS OF TWO BANTU LANGUAGES

By

Chege John Githiora

This thesis describes the theory and method of producing basic 3000-word frequency lists of the written vocabulary of two important African Bantu languages: Kiswahili and Gĩkũyũ. Toward this practical end, I discuss their morphology, with special reference to nominal and verb constructions. The first section presents the linguistic background of the two languages and related work that has been done so far on them and others in the area of lexical studies. A second section explains in detail the procedure and method used in producing the word lists, and the third part discusses some general aspects of Bantu morphology including two approaches to general morphological theory. It also covers in detail a study of word structure of the two languages with an emphasis on morpheme order and cooccurrences. Using a structural approach Bantu morphology, a template of Bantu morpheme order is elaborated. Finallly I discuss the problems and limitations of the thesis, with some concluding remarks.

ACKNOWLEDGMENTS

I would like to express my sincere gratitude to my academic advisors who steered me through two years of study here at MSU. Special thanks go to Dr. Grover Hudson for his mentoring and belief in my abilities, from the very beginning. To Dr. Dennis R. Preston for clearly perceiving my ideas and ideals, for nurturing my intuitions and providing inspiration. To Dr. David Dwyer for his great patience and open heart, willingness to listen and offer helpful advice. I would also like to thank Dr. Carolyn Harford for laying open to me her knowledge and expertise of Bantu linguistics.

I also wish to thank my friends and colleagues at MSU, for their encouragement and moral support during my studies. To the faculty and staff of Linguistics and Languages Department for their friendliness and cooperation, both which made the pains of writing this thesis a bearable experience.

TABLE OF CONTENTS

LIST OF TABLES

Page

1. **Introduction**

The aim of this thesis is to describe the theory and method of producing basic 3000-word frequency lists of the written vocabulary of two important African Bantu languages: Kiswahili and Gĩkũyũ. It also discusses their morphology, with special reference to the concept 'word'. The paper is divided into four major parts. The first one gives the linguistic background of the two languages and related work that has been done so far in the area of lexical studies. The second section explains in detail the procedure and method used in producing the word lists, and the third part discusses some general aspects of Bantu morphology including two approaches to general morphological theory. It also covers in detail a study of word structure of the two languages with an emphasis on morpheme order and cooccurences. A final part discusses the problems and limitations of the thesis, and presents some concluding remarks. The thesis describes the study of at least 40,000 'words' of each language.'Word' at this point refers to the graphic word or 'a sequence of contiguous alphanumeric characters between two spaces or punctuation' (Kucera 1967:9). But because of the rich and complex morphology of agglutinating languages such as the Bantu ones dealt with in this study, much more needs to be said before such a definition can be applied as the basis for any type of analysis.

(1) mtu anayekuandikia barua

'person who is writing you a letter'

The above sentence is made up of three graphic words. While the first and last are single lexical entries, the middle graphic word is a verb construction which can be analyzed down to a number of bound morphemes (4 prefixes, an extension and a final vowel) in the following manner:

(2) a-na-ye-ku-andik-i-a

SUBJ-TS(present)-REL-OBJ-STEM-APPLICATIVE-FV.

Note: The subject (SUBJ), relative marker (REL) and direct object (OBJ) are all in the 3rd person singular form. FV stands for final vowel.

It is necessary to analyze such a construction, to arrive at its base form, isolate the morphemes, derived forms and any inflectional variants. Each category can then be dealt with accordingly. What constitutes a 'word' and how morphemes, which play important grammatical roles, are to be treated forms a crucial part of the present paper. One fundamental theoretical goal of this research is to determine what constitutes a Unit of Lexical Analysis (ULA) for these Bantu languages. After a more complete discussion, I will provide a description of ULA as a methodological concept in section 5.3.

1.1. **Word Studies and Bantu Languages**

Studies of the lexicon have tended to be sidelined in the general study of linguistics. Work in the general field of lexicology has concentrated on dictionary writing and compilation. While the need for scientific analyses of lexical inventories and statistical studies of particular languages has long been recognized, actual work in that direction was restricted by the logistical requirements of managing large bodies of data that needed constant manipulation. Frequency or basic word studies were expensive and extremely time consuming to produce in the past. But with the spread in use of and access to computer technology, these studies have become highly feasible.

The 1960s mark the initiating period of lexico-statistical studies proper. Before then, studies of various Indoeuropean languages were done on a smaller scale--well planned and intended but with scope and results that reflected the inadequacies of their technological era. For instance, Allwood and Wilhelmsen's 1947 study of Basic Swedish word frequency involved a panel of "wisemen" who sat to decide the frequency of each word of their language. Another example is Dabb's (1966) study of Bengali done at Texas A&M University in which manual counts of words in newspaper texts were done by a team of researchers.

Even for a language like English, with its long history of scholarship, a thoroughly complete study of its lexicon was not accomplished until 1967 with the publication of the Brown Study (Kucera & Francis 1967 & 1982)

The technological advances and their impact on lexical studies are not so well reflected in African languages. Hardly any substantial work has been done and published. Even for Kiswahili, a relatively well studied language of Africa, there are no published frequency or basic word lists, much less for Gĩkũyũ. In addition, past lexicographic works (mainly dictionaries) do not integrate important underlying characteristics of Bantu word morphology. A good example is their treatment of derived verb forms as subentries of the base form despite significant meaning changes. See for instance, Johnson (1939) and Hamisi et al (1981).

2. **Objectives**

Frequency lists of the vocabulary of languages are essential bases for many types of research in both theoretical and applied linguistics. Some specific examples are given in section 2.1. While the production of such lists is greatly facilitated by technology, it is not an entirely mechanical enterprise, nor does it entail mere description of words and parts of language. A comprehensive application of linguistic theory--from morphology to semantics--is required, as is a thorough knowledge of the structure of individual languages.

This thesis attempts come up with a linguistically sound methodology of producing basic and frequency word lists of two Bantu languages and, in doing so, to elaborate on aspects of linguistic theory relevant to specific areas of the project. The work will provide a solid basis upon which the actual lists of the two languages will be produced. By the same token, it is anticipated that the credibility of this method will be such that it will inform future word studies of Bantu and other similarly structured languages. Finally, I expect that the study will produce results that may add to the body of knowledge of linguistic theory.

The fact that linguistic theory is being studied at the same time as the work of producing the actual lists is significant in that useful deductions *and* inductions can be made from the data at hand, and our theoretical questions may target specific problems encountered during the analysis. While this is a paper on theory and method, at certain stages of analysis throughout this paper I shall point out what I have achieved so far, in actual practice.

2.1. **Areas of Application of Results**

Apart from insights into the theory of language, there are several possible areas in which the results of this work are of more immediate application. The final product will be of interest to linguists, language teachers of Kiswahili and Gikũyũ, psychologists, computer scientists and others. The study will be of special relevance in the following areas.

2.1.1. Preparation of teaching materials. A teaching grammar or text-book for a language course requires knowledge of relative basicness of vocabulary, so that important words can be taught early and less important words later, and so that what is taught can be expected to be naturally reinforced in typically encountered reading materials. A list of basic words will serve to guide writers of text-books and other language learning materials.

2.1.2. Dictionaries for the languages. Basic learner's dictionaries including at least about 3000 words presuppose knowledge of what the most important (frequent) words are. A sensitive native-speaker of a language has good but imperfect intuitions of the basic vocabulary known by other native speakers, but these intuitions are unreliable for second language users of the language, whose exposure to the language is small in relation to that of a native speaker. I am currently working on a Kiswahili-Spanish-Gikũyũ dictionary whose quality and authority will be enhanced by the incorporation of the results of this study.

2.1.3. Preparation of all sorts of print materials addressed to non-native speakers of the languages. In multilingual parts of the world such as East Africa, many persons have varying but functional knowledge of several languages. Although Heine (1981:21) reports that 'the average Kenyan is proficient in at least 1.1 languages', many of those living in the central and urban areas of the country speak at least 3. A native speaker of Gîkũyũ in north central Kenya, for example, frequently has cause to read and speak Kiswahili. Print materials are readily available in Kiswahili and are much used throughout East Africa; they are becoming increasingly available in languages like Gîkũyũ. More and more, these writings must be used by persons who know the languages of the materials only as a second or third language. These important writings include such things as instructions for census reports and public-use questionnaire, voting, and instructions for use of medicines among many other uses. The effectiveness of such materials (based on corpora produced by both native and non-native speakers) can be improved if the vocabulary they use can be statistically evaluated for its likelihood to be understood by non-native speakers.

2.1.4. Innumerable types of essential linguistics research are based on vocabulary lists. For example, psycholinguistic investigations of the connotation and prototypic uses of words provides important knowledge which may be used as a basis in translating between non-kindred languages such as Kiswahili and Spanish. In such investigations it might be of interest to compare the frequency or other statistical characteristics of words of approximately equal meaning across these languages.

One hypothesis I have proposed previously [1]is that Kiswahili has received enough influence from Arabic and some Indoeuropean languages such as English and Portuguese, to the extent that certain aspects of its structure today reflect a significant shift away from other structurally related (Bantu) languages, such as Gĩkũyũ, toward that of Indoeuropean ones. For example, there is a widespread use of prepositions where they are not required by the language.

From the perspective of componential analysis which treats word meaning as the sum of its constituent parts (e.g. 'woman' = +HUMAN; +FEMALE; +ADULT), word meanings seem able to 'decompose' into their constituent parts which subsequently acquire specialized meanings. For example ndugu at an earlier stage in the history of the language had the

[1] 'Definition and equivalence in a Bilingual Dictionary of Non-Kindered Languages: Kiswahili-Gĩkũyũ-Spanish', a paper I presented at the 23rd Annual Conference of African Linguistics (ACAL), MSU, March 1992.

generic meaning of 'fellow sibling, male or female' along with other extended meanings including 'cousin, kinsman' and 'tribesman'. To denote 'sister' or 'brother', this word would be qualified by kike (female) or *kiume* (male) respectively. In present day standard Kiswahili however, *dada,* a new lexical item is used in place of the former *ndugu wa kike* 'sister', while *ndugu* is used to refer exclusively, within the realm of kinships, to 'brother' *ndugu wa kiume.* Such a change affecting a basic kinship term may be attributable to analogical change as a result of language contact with languages that have such a marked gender distinction, possibly initiated by second language speakers of Kiswahili. Gĩkũyũ has not undergone this particular change.

There are other examples of the mutual influence (at the lexical, syntactic and other levels) among the languages in contact in East Africa, which have been the subject of a few investigations in language contact and change (e.g. Scotton 1992). The results of this frequency study will enable us to statistically compare the two languages, to see if the relative basicness of words between these two languages supports such hypotheses or not.

3. **Kiswahili and Gĩkũyũ: Speakers and Orthography**

The languages of this project are important ones of the Bantu stock. Within the Eastern Bantu subfamily, Kiswahili is one of the languages of the Northeast Coast sub-group, while Gĩkũyũ belongs to the Thagicũ one (Spear, Guthrie and others). Kiswahili is the most widely spoken African language, the national language of Tanzania and Kenya, and the lingua franca for all of East Africa and much of Central Africa in Rwanda, Burundi and Zaire. With at least 5 million native speakers found within Kenya, Gĩkũyũ is the second most widely spoken language of Kenya after Kiswahili. It is often the lingua franca along fuzzy linguistic borders in the central region, and in urban areas. Gĩkũyũ is also important to the project as a representative Bantu language. Of the many Bantu languages found in East and Central Africa, many of the more important ones, such as Gĩkũyũ, have a considerable body of print materials. All are written using, largely, the basic letters of the Roman alphabet as does English. This is important because it allows us to utilize technology that has been designed for such languages as English, without need for great modifications (e.g. scanning, storage and manipulation of data stored in electronic format). Once the project has been accomplished for Kiswahili and Gĩkũyũ, it should be readily possible to accomplish it for others of the typologically close but mutually unintelligible Bantu languages.

4. **Method and Procedure:**

A concrete example of the extent to which technology permits this type of work is found in the very first stage of the project. In order to subject a corpus of large proportions to analysis, it is necessary to create a database that can be manipulated in many ways on computer. Without the appropriate means, it is an uphill task--even with a sizable team of typists--to enter into a computer such an extensive body of language on the keyboard. The otherwise insurmountable job was resolved by relying on state-of-art technology.

To build a database for this project, I used optical scanners (or Kurzweil Scanners) available in MSU computer laboratories to convert print texts to electronic "ASCII" texts. The scanners take an image of a page of text and "read it to disk". I obtained approximately 200 words with each exposure of the scanner, scanning about 200 computer pages for a total-text length in graphic words of at least 40,000 for each language.

I then edited these texts in preparation for treatment by a text-analyzing program. For Gikũyũ, there was the additional problem of the scanner being unable to recognize--and misreading--the graphemes <ũ>, <î>. Such editing is a time consuming aspect of the project and must be done by one with sufficient linguistic sophistication and knowledge of the languages being studied.

Morphemic homophones occur frequently, making it necessary to edit manually where the simple word processing commands that I use may not be effective. Some Word Perfect 5.1 functions are very useful and sufficient where generalities can be made, but it is not possible to for it recognize the difference between the sequence **<ki>** as an adverbial affix in *kijerumani* 'German' or as a (continuous) tense marker in *akienda* 's/he going' or as a diminutive prefix marker as in *kijiti*('small stick'), or as a mere (grammatically) insignificant sequence as in *ha***kimu** 'judge'.

A crucial methodological aspect of editing these texts is the tagging of each part of speech, which has to be done before doing the actual frequency counts. It is an important task which forms a crucial component of this thesis so I discuss it in greater detail in a section below.

4.1. **Zipf's Law**

A significant part of the project whose methodology I am describing here involves statistical operations. The final goal of the project is to determine the 3000 most frequent words of the two languages. The first step required was to determine how such a list of words may be obtained using statistically credible methods.

The number of tokens (i.e untagged graphic words) to be studied, 40,000, was a bit more than that strictly needed to derive a vocabulary of 3000 different words according to a ratio which is known as Zipf's law (Paul Zipf, 1945). This formula provides the probability of a word of given frequency rank. In Miller 1981:107, it is stated that

> Although some people have larger vocabularies than others have, a few words are used so frequently that everyone knows them. The 50 most frequent words in speech make up 60% of what we say ("I" ranks first); the 50 most frequent in writing make up 45% of what we write ("the" ranks first)...if word frequencies are tabulated and the words ranked from the least frequent, a simple formula describes the relation...known as Zipf's Law
>
> If p_r is the probability of the rth most frequent word, then
>
> $$p_r = \frac{0.1}{r}$$

Zipf showed that the formula gives a good approximation of word probabilities in many languages. Zipf's law is considered a universal which is expected to have rough validity for all languages, and so in a 40,000-word Bantu language text, one expects to find approximately 3,000 different words, the least frequent several hundred of which occur only once or twice in the sample. The present study is based on this principle.

4.2. The Corpus

The corpus consists of machine readable texts and has been assembled on electronic format, creating a database of these two languages for possible use in word studies of many types. The corpus seeks to provide a representative sample of modern written language which is computer-accessible for all manners of analysis and for addition or modification. Its composition reflects these efforts as well as the limitations involved in undertaking the project. Text samples range from newspaper articles to poetry, fiction and sociolinguistic studies, among other genres. In the long run, this database should become an important resource for scholars of these two languages, once it has been edited and stored in easily retrievable electronic format.

A crucial aspect of the project is selecting the texts to assure having a valid representation of each language. Of course for Kiswahili there are texts of all sorts of print materials, as for many major world languages. For this language, I have used newspapers heavily, because they provide writing over a range of topics and styles, including the editorial page, letters to the editor, news, and sports. The sample of the language obtained may be expected to be a valid cross-section of the modern written language. It is possible that spoken, colloquial, language is somewhat different in its basic vocabulary, but for the purposes mentioned above, the written, more formal language is more appropriate. However, I have included transcriptions of Kiswahili conversations which are available from a previous discourse analysis project that I did in November of 1991. This will provide at least 3000 tokens of the spoken Kenyandialect of the language. While the inclusion of these conversations may be inconsistent with our aim of studying written language, it should better be considered a pilot study which might give useful insights for a future comparative study between spoken and written language.

Corpus gathering for Gĩkũyũ was more problematic because of the unavailability of newspapers or magazines for inclusion in the study. I had to rely on a limited number of authors who have done most of the writing in this language, notably Ngũgĩ wa Thiong'o and Gakaara wa Wanjau. The representativeness of this sample is more restricted than I would prefer, but at the same time, there are not as many authors or publications in the language (though certainly much more than I have available here now at MSU) and it may be argued that the sample is valid enough in that it does represent what limited written work exists. In any case, I hope to add more diverse printed material to this corpus in the future, as they become available.

Aside from independent sources and helpful tips from my academic advisors, selection of the corpus is modelled after three similar studies of larger proportions. Two are on English--the Brown Corpus (1967) and the LOB (London-Olsen-Bergen, 1989) Corpus. (See reference section for full citations.) The former is the most famous and comprehensive word study of the English language. The latter, more recent, one was modelled after the Brown study with a few innovations. It analyzed British English. The third study is on Mexican Spanish, a project that was carried out at El Colegio de Mexico by Dr.Fernando Lara et al, whose results were published as a book in 1989. It was a comprehensive study, very similar to the Brown study in its methodology.

It also provided the basis for a subsequent Dictionary of Basic Mexican Spanish.

All these studies fit within a paradigm of lexicology in that they are based on certain common principles and all have a closely resembling methodology of compilation and analysis. They are also applied to two of the most widely spoken world languages of the Indoeuropean stock. I have stayed within the framework of this tradition to the extent where it applies in this study of vastly different languages. Both typological and linguistic differences between the languages of the cited works and the ones of this study have called for divergent and creative approaches. While always on the lookout for possible universals that may be applicable, it has been necessary to rely on my knowledge of the structure of these languages in order to solve some methodological problems.

4.2.1. Selection of Texts

With few exceptions, the corpus encompasses genres that have been used in the studies cited above. The only areas that I have had trouble representing are those of science and technology. The following genres are represented in the Corpus. Full citations of the sources of text follow.

(a) Literature (novels, short stories, poetry; drama); (b) journalistic (news, opinion; editorials, politics, sports, international news); (c) political discourse (Kenyatta, Nyerere speeches); (d) religion (Bible Gospel; missionary works); (e) cultural studies (essays); (f) popular literature (detective novels, short stories); (g) transcribed conversations.

With the two exceptions noted above, these are the genres included in all the major studies I have examined. It is also worth noting that I have selected the texts, to the extent possible, without regard to the preeminence of their authors--the representativeness of the writing being the determining factor. Also, I have made attempts to control the length of each sample to avoid overrepresenting a particular author or style. I have yet to determine the exact amount of words I used from each text or author. As I stated in 3.2. above, the serious limitations I encountered included a lack of scientific material available, and a limited pool of Gĩkũyũ authors to choose from.

4.2.2. Background to Corpus Compilation

I began by assembling all publications available in personal and MSU libraries. From these books I selected a few pages of each book or pamphlet and photocopied them. Whenever possible (e.g., with short stories or news items, I selected whole passages, but in many cases continuity ended where I skipped pages of a particular book. Text coherence (i.e. a whole continuous story or article) was not considered necessary. In the final analysis our interest is in words, and the context that may eventually be required to eliminate ambiguity is not expected to go beyond the sentence-level.

After making the cleanest photocopies with best print-clarity, (newspaper texts often have to be enlarged for treatment by optical scanners, with frequently some loss of effective print clarity) these were submitted to the scanning specialist for scanning and storage in computer disk. The texts were stored in separate files according to their print type. The scanner required a "training" session for every print face it encountered; given the wide variation of sources of texts (books of different publications, newspapers, magazines, special graphemes of Gikũyũ etc), it required many training sessions which led to scanning difficulties, higher costs, and much loss of print clarity. With the scanned texts back in my hands again, I went through each file, removing undesirable marks that the scanner had misread/misrepresented, rejoining words and sentences that may have been erroneously separated and so on. However, it should be noted that in cleansing the texts in this manner, the original texts were not tampered with, modified, or rearranged.

4.2.3. Sources of Corpus

The general sources of the corpus, the difficulties and limitations involved in the corpus gathering have been discussed. The corpus at hand includes texts from the following sources.

Kiswahili

AlKUIN, P.,O.S.B, Kabla ya Kuolewa Benedictine Publications - Peramiho, Tanzania, 1982, 7ed. [Religion, social commentary, customs and culture] Pages: 5,6,7,12,13-21 & 4 pages without numbers.

MBAABU, Ireri, *Kabla ya Kuolewa*, Longman, Kenya, 1991. [academic/sociolinguistic/language history]. Pages: 7, 95, 26, 27, 42, 43, 72,73.

KEZILAHABI, Euphrase, *Rosa Mistika*, Kenya Lit. Bureau, 1971. [Hadithi/Novel] Pages 62, 63, 69,76,77,86,87

MWENEGOHA, H.A.K. *Mwenye Uhuru*, Transafrica Press. [Biography- Nyerere], pages 10,11

TAIFA LEO [Newspaper, Kenya, Julai 31, 1992 - opinion column by Walter Mbotela on the standardization of Kiswahili; second letter in same daily opinion by Kimani Ruo "Tafsiri zinavyokanganya" [The hazards of translation].

TAIFA LEO, many other texts of various sections of this paper and the one below, KENYA LEO, have been included.

KENYA LEO [Newspaper, Kenya, Agosti 3, 1992 - 6 letters to the editor]

AMEIR ISSA HAJI et al, *Misingi ya Nadharia ya Fasihi* Taasisi ya Kiswahili na lugha za kigeni, wizara ya elimu Zanzibar-Tanzania, 1981. [Literary theory] Pages: 4-11, 17-21, 30-35, 56-62.

NASSIR, Med, *Malenga wa Mvita: Diwani ya Ustadh Bhalo,* Oxford University Press, 1971. [collection of modern short Stories], Pages 6,7,11,12,13,20,21.

MUNGIA, Justin D., *Hadithi za Mfalme Sinsin* Tanzania Publishing House Ltd, Nairobi, 1971. [Tales of old] Pages: 6,7,8,9,14

ROBERT, Shaaban, *Utenzi wa Vita Vya Uhuru 1939 hata 1945,* Oxford Univ. Press, Nairobi, 1967. [epic of WWII, Mashairi]; Pages 1-15.

POETRY ASSORTED, (4, pages)

Gîkũyũ

GAKAARA, Wanjau, *Wa Nduuta Hingo ya Paawa*, Gakaara Press, Nyeri, Kenya, June 1984. [Popular Literature].Page 2-12

________, *Wa Nduuta Kũrega na Mũtî waake*, July 1982, Gakaara Press, Nyeri, Kenya, No. 32 [Popular Literature]. Pages 6-9.

________, *Thooma Gîîgîkũyũ Kîega Ibuku ria Keerî "A"* Gakaara Press, Nyeri, Kenya, Jan 1988, (Revised Edition) [Language and Culture]
Pages; 32,33

________, *Wa-Nduuta Akîhererwo nî Mũka*, No.23, Feb., Gakaara Press, Nyeri, Kenya, 1982.

_______ , *Mwandîki wa Mau Mau Ithaamîrioinî*, Heinneman Educational
Books, 1983. [Biography-Historical Novel-Dairy]
Pages: x,xii,36,38,80,81,112,113, 117,142, 143,146,147,152,153.

NJOROGE,Lizzie N., *Nîtwendeete Rũthiomi Rwiitũ*, East African Publishing House, 1979. Kenya Institute of Education, [Folk stories]. Pages 4, 8, 16, 22].

TKK Gîkũyũ, Book 3A, Longman, Kenya Ltd. 1981. [Primary school reader]. Pages 1,2,3,4, 8,10.

THIONG'O, Ngugi, *Matigari ma Njirũũngi*, Heinneman Kenya Ltd. 1986. [Novel]
Pages; 30-37,52-55,78,79,86,87,90,91

________, *Njaamba nene na Mbaathi î Mathagu*, Heinneman Educational Books, 1982. [Children's book]
Pages;4,6,8,15,16,18,20,22-25

________, *Ngaahika Ndeenda*, Heinneman, 1986, [Play-Drama]
5 unnumbered pages).

4.3.Tagging

Methodologically the principle of "tagging" is crucial in this project. A 'tag' is a string of capital letters and/or symbols indicating the grammatical category to which a graphic morpheme is assigned (Kucera, 1967). The full list of tags used in this project is laid out below in Table I. The tagging process therefore refers to the assignment of a specific grammatical designator to each morpheme, based on the taxonomy of the language's grammatical or functional categories. Spelling variants and derivatives are noted. I have designed a tagging code for all morphemes of the two languages occurring in the corpus which shall be used for the entire data base. The tags will be the chief methodological tool that I will use in the counts--especially those of grammatical morphemes--and in other kinds of analyses arising from the project.

The tagging procedure is based on a morphosyntactic study of the two languages, and here the problems arising from the fuzziness of the boundary between morphology and syntax become obvious sometimes. In particular, the issue of what role a given morpheme --say a noun class marker-- plays in the language is challenging: is it part of the noun or not? How should such markers be tagged? Are they ULAs or not? Should they be treated the same as lexical stems or not? These questions require answers which I shall attempt to provide in the following sections.

5. **Bantu Morphology.**

To better justify the tagging procedure and categorization of tokens (graphic words), it is necessary to provide a sketch of Bantu morphology of the two languages to the extent to which it concerns this project. Although the two languages are distinct, they share many commonalities which I shall draw upon in order to construct a bilingual template which might be elaborated into an algorithm for the automated analysis of Bantu grammatical constructions.

Both Kiswahili and Gikũyũ have a morphological structure typical of most Bantu languages. Two aspects are especially relevant here:

(i) nominal constructions, with special attention to the noun class system and word formation, and

(ii) verbal constructions, with particular attention to derivation and inflection.

5.1. Nominal Constructions

This category refers primarily to nominal derivation, compounding and, in general, the system of concords affecting modifiers (adjectives, numerals and interrogative nominals such as *-pi, -ngapi*). The classificatory system of nouns plays an important role in the grammar. It vaguely reflects semantic classes and it functions in similar ways (such as determining agreement in class and number) to Indoeuropean (grammatical) gender (Jensen 1990). A Spanish example may illustrate this point:

(3) a. cuant**as** persona**s** vin**ieron**? (Spanish)
b. **wa**tu **wa**ngapi **wa**likuja? (Kiswahili)
c. nî **a**ndũ **a**igana **mo**okire (Gĩkũyũ)
'how many persons came?'

Morphemes in bold are those which are attached to a stem to indicate gender (fem.) and number (plural) for Spanish in (3a); (3b) and (3c) shows class (1/2) and number (plural) for Kiswahili and Gikũyũ respectively.

Noun class markers function as indicators of class; nouns may be shifted to a different class, yielding a modified meaning. For example, they may be shifted from an original class to another to form diminutives, augmentatives or collectives. This is done by changing the class prefix. Note that as in the case of verb extensions in derivation, the resulting meaning is not always predictable as exemplified below.

(4) Kiswahili

a. **m**toto (cl. 1/2) > **ka**toto (cl. 13/12)
'child' > 'little child'
b. mlima (cl. 5/6) > **ki**lima (cl. 7/8)
'mountain' 'small hill'
c. simba (cl. 9/10) > masimba (cl. 5/6)
'lion' 'pride of lions'

(5) Gĩkũyũ

a. **mw**ana (cl. 1/2) > **ka**ana (cl. 13/12)
'child' 'little child'
b. **mũ**ndũ (cl. 1/2) > **kĩ**mũndũ (cl. 7/8)
'person' 'gigantic person'

The above examples illustrate two fundamentals of Bantu morphosyntax. A change of prefix (in bold) brings about a change in nominal class and meaning.

Compounding creates new words by combining two or more others. Various types of compounds are found throughout the lexicon. *Mwana,* 'child' as a relational term 'denoting the practitioner of a profession related to N' (Mchombo & Bresnan 1993:11) is quite productive as the following Kiswahili examples show:

(6)	wa-/**mw**anasiasa	<	mwana	+	siasa
	'politician'	<	'member'	+	'politics'
	wa-/**mw**anachama	<	mwana	+	chama
	'partymember'	<	'member'	+	'party'

Other compounds are created by juxtaposing a verbal base and a direct object or bare NP, and of course the appropriate noun class prefix (in bold).

(7)	0-/kingamwili	<	kinga	+	mwili
	'antibody'	<	'guard'	+	'body'
	vi-/kionambali	<	-ona	+	mbali
	'binoculars'	<	'see'	+	'far'
	mfanyabiashara	<	-fanya	+	biashara
	'business person'	<	'do'	+	'business'

Nouns can also be derived by adding a nominal suffix to a deverbative noun, and the appropriate noun class prefix. There are several nominal suffixes which denote a variety of effects such as agentive, instrumental and state, as in (8a,b) below, for Kiswahili and Gĩkũyũ respectively. Note that these correlations are not absolute:

(8) a. (i) instrumentals: 'doer of action; result of action.

mi-/msem**o** (cl.3/4) 'saying' < -sema 'say'

ma-/nen**o** (cl.5/6) 'word' < -nena 'speak'

(ii) agentives

wa-/mwua**ji** (cl.1/2) 'killer' < -ua 'kill'

wa-/mwandishi (cl.1/2) 'writer' < -andika 'write'

(iii) of state

utuliv**u** (cl.14) 'calmness' < -tulia 'be calm'

uharibibif**u** (cl.3(14)'destruction' < -haribu 'destroy'

(8) b. **mw**ak**o** (cl.3/4) 'building' < -aka 'build'
ci-/**ki**ug**o** (cl.7/8) 'word' < -uga 'say,speak'
0-/**mũ**raga**ni** (cl.1/2)'killer' < -ũraga 'kill'
wanang**i** (cl.11/12) 'destruction' < -ananga 'destroy'

Note: In bold are singular noun class prefixes; on the left are the corresponding plural prefixes (standard notation: 0 represents zero-prefix). (8b) are analogous Gîkũyũ examples. Numerical class markers are indicated.

5.1.1. Modifiers

In addition, any phrase that modifies or is predicate of a noun phrase agrees with the head noun in class, person and number. Adjectives and determiners agree with the head noun and verbs agree with subject. This can be exemplified using the same examples of (4) as head nouns in the following sentences:

Kiswahili
m-toto **w-**ake **w-**a tatu **a-**ta-kuja kesho
child his/hers of three he/she (fut.) come tomorrow
'his/her third child will come tomorrow'
ka-toto **ka-**dogo **ka-**ta-kuja kesho
child DIM-small FUT-come tomorrow
'a small child will come tomorrow'

Gîkũyũ
mw-ana **w-**ake nî-**a-**go-oka rũciũ
child his/hers-FOC-he/she-FUT-come tomorrow
his/her child will come tomorrow
ka-ana **ka-**niini ni-**ga-**go-oka haaha rũciũ
DIM-child DIM-small FOC-he/she-FUT-come here
'a little child will come here tommorow'.

The examples show how a change in noun class prefix (in bold) results in a meaning change and, how the syntax of Kiswahili reflects these changes. The same behavior can be seen in the examples of Gĩkũyũ. They clearly show not only the morphosyntactic processes above mentioned, but also the structural closeness of the two languages. They also provide an important basis for the argumentation which I will use to justify my methodology.

5.2. **Unit of Lexical Analysis (ULA)**

> Not only are there considerable difficulties pinning down any universally applicable notion of 'word', it appears that even when we restrict ourselves to morphological criteria within a single language we find that the term itself covers a multitude of sins, which need to be carefully distinguished.- (Spencer, 1991:45)

In light of the above sketch of the nature of Bantu nominal constructions it is pertinent to pose the question: what part of the construction should be the basis of a word study, given that "word" is so elusive a concept in Bantu? In reply to this question, I propose to name that unit I am interested in for counting purposes, a Unit of Lexical Analysis (ULA). The ULA is an unbound minimal unit, essentially the stem without prefixes in the case of verbs. It includes therefore, verb roots and stems (i.e., verb roots plus extension(s)), and modifiers in their base form. It does not include noun class markers and most pre-verb stem morphemes such as tense and subject markers. These shall be tagged and counted separately, not as ULAs but as morphosyntactic units. This distinction does not diminish their importance, rather it is a methodological necessity for the purposes of this word study since it is the counting of ULAs that is of primary importance in the project, but at some point the prefixes may also be counted.

5.2.1. Independent Nominals as ULAs

To justify the above definition, let me begin by going back to the previous examples and identify the independent nominals found there:

(9) *mtoto, neno, mwuaji, mwandishi, kionambali,*
mwana, kaana, kiugo, mũragani

The nouns in (9) are independent nominals which, although they have (variable) class prefixes, may be considered lexical entries by their own right. Without those prefixes, they cannot be readily identified nor can they stand on their own. Nor can their meaning be immediately discerned without their respective affixes. That is to say that the stems and affixes are bound to each other. In any given text--whether written or spoken--they are never divorced from their prefixes. For the purposes of frequency or other counts, these are ULAs. The prefixes alone are of no significance taken out of context, or where grammar is not the object of study. Independent nouns such as *mwana, mtoto, msemo, wanangi, utulivu, kionambali, mwanachama* and so on are therefore ULA's. Plural and singular forms, those that are the result of shift of class of the same lexical item are also ULAs, independently of their status as 'variants', members of a lexical paradigm, for example:

(10) Gĩkũyũ: *mwana(sg.*'child'*),ciana (pl.),kaana*
(sg.dim),twana (pl.dim.)

Kiswahili: mtoto (sg.'child'),*watoto (pl.)*,*katoto (sg.dim)*,
tutoto (pl.dim)

Compounds and all other derived nominals shall also be treated as ULA's *not* as members of the paradigm of their original components, for example:

(11) *kionambali; mwanachama; mũragani; mũrĩraikĩhia;*

All graphic words that are independent nouns (as in 11 and 12) remain untagged. Only modifiers and prefixes bear a tag as described before. This procedure will enable me to generate the lists of verb stems and nouns in alphabetical order without the difficulties that would arise if they bore a tag (which is placed at word initial). It also reduces the number of tags on the corpus. The verb and nominal ULAs will thus be identified by default.

5.2.2. Modifiers as ULAs.

Adjectives and interrogatives, however, will receive different treatment. In the lexicon of the language, these are not inherently bound to any one or two particular prefixes. The shape of the prefix that they take is unstable in that it changes constantly depending on the noun that they modify, unlike the independent nominals whose variation is restricted to only two--singular and plural. Hence the ULA of this category of lexemes will be the base form such as:

(12) *-ake, -ngapi, -refu, -dogo* , etc.

These will then be tagged with AS, IS for adjectival stem and interrogative stem respectively, and the prefixes that have been detached will also carry a tag D, E or Q etc. to indicate their original syntactic function i.e. as adjectival, noun class, interrogative prefixes and so on (see Tagging Code in 4.3.1). This will permit the determination of the frequency of either the prefix or stem. As mentioned above, the counting of these pre-stem morphemes will be for the purposes of grammatical analysis in contrast to the counting of lexical stems (ULAs) for the purposes of word study. It may be of interest for instance, to find out the distribution and frequency of certain grammatical pre-stem markers or the statistical relations between morphemes and the stems they are attached to.

5.3. Verb Constructions

As for verb constructions, similar principles of categorization will be followed in extracting the ULA. As discussed in section 5.2., this is basically an effort to isolate stems. I will return to example (1) repeated here as (14) to clarify my postulation.

(13) mtu anayekuandikia barua
   person he/she-PRES-REL(who)-OB(you)-write-APPL-FV letter
        'the person who writes a letter to you'

Following the above criteria, the first thing in this sentence would be to isolate the initial and final independent nouns **mtu** ('person') and **barua** ('letter') as a ULAs. The *<m>* in **mtu** is a prefix which is an integral part of the whole noun and it shall thus remain attached. The middle graphic word would then be analyzed into the following morphological constituents:

(14)   a- na-ye- ku-andik-i-a
SUB-TS-REL-OB-STEM-APPL-FV

The first four morphemes are markers of important grammatical information, viz subject, (present) tense, subject relative (3rd pers. sing.) and direct object. A verb root with an extension follows. These pre-stem morphemes however are bound to the verb stem in the sense that they cannot stand on their own. It is only the fifth and final sequence *-andikia* (= 'to write to, for') which bears these characteristics and, the only one in this particular construction that will be isolated for the frequency counts. This stem has an applicative extension *-i-*, a morpheme in its own right. In the present analysis, this is considered as an integral part of one of the members of a set of lexical forms *(i.e.,-andikia)* having the same root and belonging to the same major word class (verb). A full inventory of such related forms of a verb root and its different extensions is termed a **paradigm**.

In many Bantu language dictionaries (e.g Johnson 1939, Hamisi 1989 for Kiswahili and Benson 1960 for Gîkũyũ), entries are similarly grouped, ignoring the meaning changes that result from verbal extensions. Such a procedure lacks the crucial distinction between morphological and semantic relationships and, is detrimental to the usefulness of such Bantu dictionaries by non-experts, an issue I take up later in this paper. The following examples illustrate this problem.

(15)a. Kiswahili:

-end-a	’to go’
-end-e-a	’go for, against, toward’
-end-esh-a	’cause to go; drive; have diarrhea’
-end-ele-a	’continue; ’develop’
-end-ele-z-a	’cause to develop’
-end-e-an-a	’go toward, against, for each other’

b. Gîkũyũ:

-in-a	’sing’
-in-îr-a	’sing for’
-in-ithi-a	’make sing’
-in-ain-a	’swing to and fro’
-in-a-in-ithi-a	’cause to shake’
-in-îr-îr-a	’boo’; ’pester’

On the left side of the paradigm is the verb root (*-end-* for (15a) and *-in-* for (15b), with affixed extensions which add a specific meaning to the base form which, in both languages, has an ending or final vowel (FV) **-a**. The shape of these extensions is constant (e.g. *-e-, -esh-, -ir-, -thi-*, etc.), but morphophonemic processes such as assimilation, vowel harmony and coalescence, and effects of syllable structure constraints and others may intervene depending on context. Furthermore, extensions may be built upon other extensions as in the last two examples of (15a), where a causative and reciprocal extension respectively, are added onto the already inflected applicative forms.

While the derivational process involved in creating a paradigm is regular, semantic changes may occur making it impossible, in many cases, to treat the meanings of the derivations within a paradigm as mechanical outputs as illustrated in (15a,b). However, I will not discuss the nature and effects of these semantic shifts here. This is because semantic idiosyncracies of individual forms do not affect their grammatical categories; all the forms basically remain verbs; it is only their meanings that are affected by the subsequent change in argument structure. For instance, *-andika* 'write' requires two arguments: a subject and a direct object (writer and letter), whereas *-andikia* 'write to, for' requires a subject, a direct object and an indirect object (writer, letter and recipient of letter). Derivation in these two languages does not affect functional categories and so, these changes of meaning will not be relevant to the counting of individual forms of each category.

My primary concern in this project is to obtain lexemes ULAs that are potential dictionary entries. The study and count of morphemes which are found in the grammar of the language, such as the prefixes of (15) is secondary but also important for other purposes. I shall be concerned with stripping verb constructions of their pre-stem prefixes but leaving extensions intact. I shall illustrate this point using the following Kiswahili passage as an example:

(16) *...kwa siku nyingi alitaka* ***kumwandikia*** *rafiki yake barua ya mapenzi. Lakini kila alipoanza* ***kuiandika****, alishindwa;* ***haikuandikika****.* ***Aliendelea*** *kujaribu bila kuweza hata* ***kwenda*** *kwake nyumbani. Siku moja,* ***akiendesha*** *gari lake, alimwona njiani,* ***akamwendea*** *na kumwomba* ***waendelee*** *na urafiki wao...*

'...for many days s/he wanted **to write to** his/her friend a love letter. But every time s/he started **to write it** s/he could not; it could **not be written** (it

> was impossible to write it). S/he **continued** to try, unable even **to go** to her/his house. Later, **while driving** in his/her car, s/he saw her/him in the street, **went to** (approached) her/him and begged her/him **to continue** with their friendship...'

There are many verb and nominal constructions present in the above passage. However I shall identify only those that are based on two verb-roots which I have discussed before: -**andik-** 'to write'; -end- 'to go' . Nominal forms are ignored here. The relevant constructions are highlighted both in the Kiswahili text and in the English gloss, and they are isolated in (17) below. Note that this is also what the sorting program does when asked to sort *graphic* words:

(17)
a. *kuandika, kumwandikia, haikuandikika*
b. *kwenda, aliendelea, akiendesha, akamwendea, waendelee*

Stripped of the pre-stem morphemes and maintaining inflections intact, the above lists yield the following verb stems:

(18)
a. *-andika, -andikia, -andikika*
b. -enda, -endelea, -endesha, -endea, -endelee

This is as far as the analysis goes. I consider all forms of (19a,b) to be ULAs which shall be subjected to frequency and possibly other types of counts. According to the discussion in 5.3. these are entries or ULAs in their own right. The complete list of different forms of the same verb in (a) or (b) represents a paradigm for each one of the two verb-roots *-andik-* and *-end-,* respectively. The diverse additions or changes of meaning brought about by the suffixation are not relevant for our present purposes, and hence do not affect the methodology. Their idiosyncracies will be dealt with in the dictionary where they shall be defined, giving special attention to such meaning changes. Since it is the individual members of these paradigms that shall be counted, the main task for the present is how to arrive at these forms.

5.4. The Tagging Code

Following the taxonomy I have described in the above section, the parsing exercise will be completed through tagging. Below are the tagging codes, with one tag for every pre-stem morpheme found in Kiswahili and Gĩkũyũ graphic words (in the sense established in 1.1). For each language there is a total of 27 tags (see Table I.). These Tags shall be used consistently during the project to identify every information bearing grammatical affix. What is left untagged will be lexical units (ULAs), the principal parts which will be counted. A full description of each category is provided in Table II, in section 6.3.2. Only examples are given here.

Kiswahili

S = subject markers prefixes, e.g. *u-,a-,tu ki-,i-,* etc

T1 = past tense marker present tense marker, *-na-*

T3 = completive/present perfect (*-me-; -ja-*)

T4 = conditional present *-nge-*)

T4B = conditional past (- *ngali-*

T5= future tense marker (*-ta-,-taka-)*

T6 = habitual tense (*hu-*)

T7 = consecutive (*-ka-*)

T8 = present continuous(*-ki-*)

O = Object marker e.g *--zi-ya,-i-,* etc)

P = noun class markers e.g. *ma-,wa-* etc, with phnological variants, e.g. *mw-,w-* etc.)

L = locatives (*po-, mo-, ko-, -ni*)

R = subject relative markers (e.g. *-cho-,-zo- -o-,* etc)

A = adverbial prefix (*ki-,* e.g. **Ki**afrika)

I = infinitival prefix *ku-*

N = Pre-initial negative (*ha-, si-* including *si (=negative of* ni)

Q = interrogative *-pi, wapi, nini etc.*

J = noun ending (*-ji,* e.g. mjua*ji*)

D = adjectival concords and possessive concords (similar to P)

C = conditional (*-ki-*)

PP = preposition e.g ch-*a*

CP = copula (ni);*'si'* neg under **N)**

RF = reflexive (*-ji-*)

AS = adjectival stem (e.g *-ake, -kubwa, etc.)*

7S = numeral stem (e.g. *-tatu, -tano, etc.)*

IS = Interrogative stem (*-pi, -ngapi)*

Gĩkũyũ

S = subject markers (prefixes, e.g.*ũ-,a-,tũ-,gĩ-, ma-,* etc)

T1 past tense marker (various remote, recent, etc.)

T3 = completive

T4 = conditional present (*-ngĩ*)

T4B = conditional past tense (*-ngĩa-*)

T5= future tense marker

T6 = consecutive (*-ga-/-ka-*)

T7 = present continuous (*kĩ-*)

O = object marker (various, e.g *-o-,-ci,-ya-,-ma-* etc, including reflexive **ĩ**)

P = noun class markers (see adjectival concords D, usually of same form e.g. *ma-, wa-*, etc, with phonological variants, e.g. *w-, mw-* etc.)

L = locatives (*-ho, -nĩ*)

R = relative markers

A = adverbial prefix (*kĩ-)* e.g. **Gĩ**ikamba;**Kĩ**jeremaani)

I =infinitival prefix (*kũ-)*

N = negative Pre-initial (*ha-, ti,* etc.)

Q = interrogative (*nũũ, rĩ, kũ-,including -kĩ* suffix)

D = adjectival concords and possessive concords (similar to P)

C = conditional (*-ngĩ-*)

PP = preposition (e.g *ci-a)*

CP = copula (*nĩ*). Note that'*ti*' (negative counterpart) will be categorized as **N)**

7S = numeral stem (e.g. *-thatũ, -thano, etc.)*

IS = Interrogative stem (*-ũ, -igana)*

6. Bantu word structure and Morphological Theory

It is clear that there is a specific ordering of morphemes in the structures we have looked at so far. This is not accidental. Linear ordering of pre-stem morphemes, or, prefixes in verb morphology has long been acknowledged in Bantu language studies, and has been the subject of much theoretical discussion. In many ways, Kiswahili is rather straightforward and the order of morphemes has been well studied; this is less so for Gikũyũ; I shall return to this point later. There are two major approaches to the study of Bantu morphological structure which I will briefly discuss.

6.1. The Template Approach

The first one, known as the "Slot-and-filler" or "Template" approach (Lyons 1968, Gleason 1961), stems from the structural school of linguistics. Word formation is a process of stringing together of morphemes "like beads on a string" (Lyons 1968:56.). According to this approach, Bantu "word" structures consist of slots which may be filled by one of a finite list of morphemes. The graphic words we have seen then are flat structures or templates composed of constituent morphemes that are subject to certain collocational restrictions and hierarchy. The template approach has been applied to many languages, especially agglutinative ones. Given the set of coocurrence restrictions that apply, the (linear) ordering of these morphemes can be predicted.

Fundamentally, the "flat structure" referred to in this approach is equivalent to the "Dokean word" where "word" refers to a string of morphemes whose boundaries are determined by stress placement (Myers 1987). Word boundaries in a language like Kiswahili (and many other Bantu languages) which assign stress to the penultimate syllable, correspond to the limits of the graphic word I have discussed, which in the two languages under study, can be defined following similar phonological criteria.

Rubanza (1988) in an MSU Ph.D dissertation arrived at similar conclusions about the phonological word after doing an exhaustive study of Haya verb morphology. He cited phonological, syntactic, semantic and pragmatic factors that determine both the ordering and predictability of morphemes. Haya is a Bantu language and in his work, Rubanza drew many parallels to both Gikũyũ and Kiswahili morphology. The claim of predictability is motivated by the fact that "it is hard in Bantu languages (impossible in Haya) to have all (23 for Haya) morphemes in one verbal construction". He then elaborated a Morphemic Formula for the Haya Verb (MHV), a complex template which summarized all the restrictions that apply within the Dokean word.

6.2. The Generativist Approach

The second approach is termed "configurational" (Myers 1987) Ph.D dissertation which adopts the principles of the generativist school of grammar. It is configurational because it involves a binary branching structure ordered in two levels: word and stem. Briefly put, Myers utilizes a version of Chomsky's 1970 X-bar theory of natural language constituent structure that is head driven and binary branching. His position rejects the template approach discussed in the previous section. The Dokean word upon which the template approach is based is also rejected on the basis that it fails to capture certain generalizations about Bantu word structure; in this study for example, it would be quite misleading to use the Dokean word par excellence as a unit of lexical analysis because morphological boundaries between lexemes and other constituents are not easily defined phonologically. Myers's central thesis is that "Bantu morphology is configurational and binary branching, that the rules governing it are context-free and strictly local...the Bantu "word" (i.e "Dokean word") is not in fact a morphological or syntactic constituent at all, but rather a derived phonological domain i.e. a phonological word" (Myers 1987:12). All evidence there is in Shona for the Dokean word-- such as being the domain for stress, epenthesis and Meeussen's rule on tone lowering--is all phonological.

6.3. An Eclectic Approach

The two approaches I have outlined above are divergent, reflecting fundamental philosophical and practical differences. No doubt the configurational stance is of much theoretical interest for generativist grammars of Bantu languages. It seeks more to explain the deeper structure of Bantu word formation. The main critique of the template approach is, predictably, the descriptivist nature and relatively limited explanatory powers of the latter. Both schools of thought have points that are useful, though in fact the template is more attractive to me as far as the aims of this project are concerned.

The so called Dokean word clearly corresponds to the graphic word I have described, although this in itself is not of major significance. For present purposes, it is more convenient to regard structures as linearly ordered, and, without regard to the internal constituency of the verb structure (binary or not), there is no questioning the ordered character of the pre-stem morphemes (morpheme sequencing). On the surface, they are readily identifiable and isolable, a crucial first step toward my goal. In other words, description rather than explanation is the immediate goal here. Furthermore, the idea of a morphemic formula based on a study of the cooccurence restrictions of the affixes seems to hold great potential. Ideally, a template for each of the languages might be designed--perhaps a single one that captures enough generalizations for application in both languages (and other Bantu ones). From such a template, a mathematical algorithm might be worked out allowing us to analyze every graphic word automatically.

The so called Dokean word clearly corresponds to the graphic word I have described, although this in itself is not of major significance. For present purposes, it is more convenient to regard structures as linearly ordered, and, without regard to the internal constituency of the verb structure (binary or not), there is no questioning the ordered character of the pre-stem morphemes (morpheme sequencing). On the surface, they are readily identifiable and isolable, a crucial first step toward my goal. In other words, description rather than explanation is the immediate goal here. Furthermore, the idea of a morphemic formula based on a study of the cooccurence restrictions of the affixes seems to hold great potential. Ideally, a template for each of the languages might be designed--perhaps a single one that captures enough generalizations for application in both languages (and other Bantu ones). From such a template, a mathematical algorithm might be worked out allowing us to analyze every graphic word automatically. For such an algorithm to succeed completely, it would have to take very many factors into consideration: phonological (including tonology, phonetic variations, vocalic transformations etc.), morphological, syntactic, semantic, pragmatic, and so on. Upon closer examination, the feasibility of successfully doing so is limited by the complexity that such a template entails. But in spite of the complexity of Rubanza's MHV, and the number of variables that come into play especially for Gĩkũyũ, there are prospects for success of such a solution if all Also, the potential use of such a template remains far too attractive. Basing myself on Bantu morphological characteristics and actual data found in the project database, I attempt in the following section, to construct such a template.

6.4. A General Template of Bantu Verb Structures

With only a change of tense, the constituent morphemes of the verb structure in (2) can be assigned slots in a template such as follows.

(19)

FOC 1	SUB2	TS 3	REL 4	OBJ5	RT 6	EXT7	FV 8	PF 9
(0)	a	ta	ye	ku	andik	i	a	
nî	**a**	**ga**	**(0)**	**ku**	**andik**	**ir**	**a**	

The above is a highly generalized template showing the order of markers occurring in a given Bantu verb construction. Two concrete examples are provided in the template, first row for Kiswahili and in the second an analogous Gĩkũyũ one are provided. Slots with (0) (e.g. Gky.(4) and Ksw.(1)) are unfilled because they cannot occur within the exemplified constructions. Markers are coded as FOC=focus, SUB= subject, TS=Tense, REL=subject relative, OBJ=direct object, RT=verb root, EXT=extension, FV=final vowel and PF=postfinal morpheme. As a second step, I will describe each category, its individual members and, the cooccurrence restrictions that apply between them. Two morphemes *cooccur* if they can occupy their respective slot within the same construction. While there are significant differences between the order and appearance of morphemes in each language, many of the common features observed might facilitate the construction of a single, bilingual template.

Two basic assumptions apply: no two markers of similar or different categories can occupy the same slot at the same time. Second, a slot may remain unfilled. In view of the goals of this study that I delimited earlier, I shall limit the analysis to a certain level. I shall not delve into morphophonemics for instance. For the sake of consistency, I shall proceed by dealing separately with each one of two broad categories: pre-stem and post-stem morphemes. Due to lack of space, it is not possible to include all the relevant information and examples, and present it graphically in linear order as in (17). I have used numbers to encode the morpheme categories and show the sequencing.

6.4.1. Inventory and Description of Morphemes.

Table I

(1) FOCUS (prefix, Gky. only) **nĩ-** (2) Subject Markers (SUB) (prefixes) **Kiswahili:** **a-,-u-,a-,wa-, tu-,i-,li-ya-,ki-,vi-,i-,zi-,ku-,pa-ku-,mu-** **Gĩkũyũ:** **mũ-, a-, mĩ-, kĩ-, i-, ci-,n-, rũ-, ma-, ka-, tũ-, kũ-, ha-, kũ-**	(3) Tense Markers (TNS) (infixes) general present: **-na-** progressive **-ki-** past **-li-** future **-ta-** perfective **-me-,-mesha** consecutive **-ka-** habitual **hu-** conditional **-nge-,-ngali-** **Gĩkũyũ: cell (3b).**
(3b) **Gĩkũyũ** Tense markers (TNS): Past: -remote **a+...+ire/aga** -near **ra+...+ire/aga** -immediate **kũ+...+ĩte/aga** Present **a+...+a/ĩte** Future: -immediate **kũ/gũ+...+a** -near **rĩ+...+a/aga** -remote **ka+...+a/aga** Consecutive:**-ka-,-ra-,-a-,gĩ-** **Ro**+...(**+aga**) tense **Na...a** tense (habitual) progressive **-kĩ-** conditional -future **-ngĩ-** -past **-ngĩa** Habitual **-aga** Intentional **-aga** Repetitive **-aga**	(5) Object Markers (OBJ) (infixes) **Kiswahili:** **-ni-, -ku-,-m-,-wa-,-tu-,-i- -li-,-ya-,-ki-,-vi-,-i-,-zi-** **Gĩkũyũ:** **-n-, ndĩ-, -kũ-,-i-, -mũ-,tũ,-ma-,-ũ-,-kĩ,-ci-,rũ-,-mĩ-,-ka--rĩ-,-ha-, -kũ-**

Table I (cont'd)

(4) Relative Markers (REL) (infixes) **Kiswahili:** **-ye, -o-,-yo-,-lo-,-cho-,-vyo-,-zo-,-ko-,-po-,-cho-** **Gĩkũyũ:** An independent relative pronoun **-rĩa** is used, rather than an infix within the verb construction as in Kiswahili. A marker of the same class and form as in (3) is prefixed.	(6) **VERB ROOT (RT).** e.g. Ksw: -andik- Gky: -andĩk-
(7) Extensions (EXT) (Capital letters represent archiphonemic vowels) **Kisw.** **Gky.** stative (**-Ik-**) **-Uk-,-ara** causative (**-Ish/z**) **-Ia/Ithia** passive (**-Iw-**) **-Iw-** applicative (**-I-**) -**Ira** reversive (**-U-**) **-Ura/Uka** reciprocal (**-an-**) **-ana** positional (**-am-**) repetitive (**-ag-**) **-anga** potential -- **-Ika**	(8) Final Vowel (FV) (suffixes) indicative **-a** subjunctive: **-e** imperative: **-e** **Gĩkũyũ:** indicative **-a, -aga.** subjunctive **-e, -age** imperative **-a** (The *e's* of the subjunctive are treated as FV's)
(9) Post Final (PF) Ksw.: locatives *-po, -mo--ko-* *-ni* Gky.: *-ho, -kuo* *-rĩ, -ĩ*	

This table summarizes the most ubiquitous components of a Kiswahili verb construction (Ksw.) and Gĩkũyũ (Gky.). While I have striven to be thorough, it must be noted that the layout does not describe all specific, internal characteristics of the verb. Many morphophonemic details have been left out because they are not of direct relevance to the problem at hand, even though they may affect the surface forms. There are several other language and morpheme specific characteristics that have been left out. Tone plays an important role in the morphosyntax of Gĩkũyũ, but a truly comprehensive study is beyond the scope of this paper. What is of importance is that the outline includes enough information to enable us to proceed to delimit the main cooccurrence and other restrictions that apply, the next step in our search for an adequate template.

Other peculiarities of Gĩkũyũ relate to tenses in cell (2b), which are more complex, marked in some cases by two discontiguous morphemes, e.g. **a...ire** for the remote past tense. Other tenses lack equivalents in English hence they are not given a standard grammatical category, e.g., '**na...a** tense'. Dahl's law, which voices (prefix) initial stops when the stem consonant is unvoiced, is a highly productive rule of Gikũyũ which should be assumed to apply in all appropriate contexts.

6.4.2. Order of pre-stem morphemes

In order to maintain coherence I shall analyze two types of complex verb constructions individually: a negative verb construction and an affirmative one, then draw generalizations. As before, I shall deal with pre-stem morphemes first.

(20) *nîtũgakîmũandikîraî*
'so **then** we shall then write to him/her'

nî	tũ	ga	kî	mũ	andik	îr	a	î
1	2	3	CON	5	6	7	8	9

Cell 4 is unrepresented because, as mentioned earlier, Gîkũyũ does not use an internal relative marker (see (19b)); cell 9 is a discourse marker (variant: *rî)* affixed to the verb construction after the final vowel. Its function is similar to 'the way a comma is used in English' (Barlow, 1951:13). In cell 5 is a connective particle, *kî,* of Gîkũyũ. Now let us look at the negative counterpart of (20).

(21) *tũtigakîmũandîkîraî*
'so, we will not write to him/her'

tũ	ti	ga	kî	mũ	andik	îr	a	î
2	NEG	3	CON	5	6	7	8	9

Observation: FOC(1) and NEG do not occur within the same verb construction. There is a allomorphic NEG marker *-ta-* which is used in relative clauses.

(22) *ũrîa tũ-ta-ga-kî-andîk-îr-a-î*

REL (2)-NEG-(3)-CON-(6)-(7)-(8)-(9)
'the one then, whom we shall not write to'

tũ	ta	ga	kĩ	andik	ĩr	a	ĩ
2	NEG	3	CON	6	7	8	9

The same restriction applies. *-ta-* is simply a substitute of *-ti-* of the main clause NEG. Also note that when an independent relative *ũrĩa* is used, the object marker of cell 5 cannot occur.

We can include in example (22) another marker, the reflexive **ĩ** which simply occupies cell 5, without affecting the template any further. Examples (21) and (22) represent the maximum number of pre-stem markers possible in a single verb construction of Gĩkũyũ.

Observations: The CON occupies the same slot as (3)TNS, immediately preceding (5)OBJ unless the optional (4)REL is present in which case CON fills that slot. The reflexive marker replaces (5)OBJ, so the two cannot cooccur. Consider Kiswahili (22).

(23) *tu-taka-ye-mu-andik-i-a*
2- 3- 4- 5- 6- 7-8
'he/she who we will write to'

ha-tu-ta-mu-andik-i-a
NEG- 2- 3- 5- 6- 7-8

Observations: NEG does not cooccur with 4 REL; the independent relative *amba-* is used in such cases. Note that *-taka-* is a variant future tense form (c.f.*-ta-)* which is required by relative constructions.

The examples and observations of (18)-(23) provide sufficient information for me to make a few generalizations about pre-stem morphemes of both languages.

In Gîkũyũ, a maximum of 5 pre-stem markers can occur in a single affirmative or negative constructions, 4 in Kiswahili. As Barlow (1951) does for Gîkũyũ, I consider there to be a total of 7 pre-stem morphemes in the grammar of Gikũyũ, 6 in Kiswahili. The discrepancy between the maximum possible and those that may actually occur in a single surface construction is accounted for by the cooccurrence restrictions which I attempt to summarize in the following template. Their order in an actual surface form is as they appear in the template.

6.5. A Bilingual Morpheme Template

(1)	2	(NEG)	3	(4) **(CON)**	(5) (RF)	6	(7)	8	(9)

Following the numbering code established in Table II, pre-stem morphemes in non-relative, affirmative verb clauses appear as in Table II, underlyingly, but the following restrictions apply. The parentheses around a marker indicates that it may or may not appear on the surface structure. In slots ((4a), (5)) where there are two possible markers, both may not occur at the same time.

1. Morphemes in bold occur only in Gîkũyũ.

2. Cell 4 never occurs in Gîkũyũ; its slot may or may not be

filled by CON

3. 1 and NEG do not cooccur

4. It is assumed, following Schadeberg 1990, that the Kiswahili preinitial negative marker *ha-* has fused with the subject concords. Forms such as *hu- (<ha+u), hatu- (<ha+tu), hai- (<ha+i) and si- (cl.1)* etc are units that fit in the 3rd (NEG) slot of the template. Same applies for Gîkũyũ's *ndi-,nda-,ma-,mũ-* (neg. personal pronouns.).

To this point, an analysis of pre-stem morphemes has yielded sufficient generalizations to allow for a workable template. If the template correctly describes the morphological structure of the two languages, it is possible to build the information of Table II into a computer program which would then be able to recognize slots (1) through (5), detach them from the main construction to isolate verb stems with their various suffixes. Such an operation would yield what we defined as ULAs.

7. Expected Results

Having established objectives and a methodology, I shall now pause to take stock of what I eventually hope to achieve at the completion of this project. The principal expectation of this thesis is the production of a solid methodology for Bantu lexical studies, including a definition of Bantu ULA and a template, for immediate application and future adaptation for use in other related languages. From the project itself will result:

1. A complete list containing verb stems and their derived forms. These will be alphabetically ordered.
2. A list of nouns (together with their class prefixes) and dependent nouns with their frequency rankings. These may be subject to further analysis such as concordance lists when contextual information might be required.
3. A frequency count of pre-stem morphemes (markers).
4. Descending frequency rankings of (1), (2) and (3).
5. Descending frequency rankings of (1) and (2) combined which should give a list of most frequent words (ULAs); the top 3000 ULAs will be considered the basic words. This is the most fundamental objective.
6. Frequency counts across genres for relative frequencies.
7. Comparison of results (1)-(4) with similar ones on English and Spanish.
8. A data base of Bantu lexicon and tagged items for future use in other kinds of studies.

8. Discussion and Conclusions

There are some problems with this methodology that I have not yet been able to overcome. The first has to do with the compilation of the corpus. As established earlier, there is a notable lack of representation of certain genres of writing in both languages, specifically scientific and technical material. For Gikũyũ the deficiency is greater since I have had to rely on a restricted number of authors and publications. The variety of newspapers and magazines used for both languages is also rather small.

Up to this point the method I am using to analyze structures remains unperfected. The most desirable means of doing so, which would also represent a significant advance in Bantu studies, would be as said earlier, to make a workable algorithm which could produce regular and reliable results faster, in an automated manner. I continue to strive to perfect the template to eventually develop such an algorithm. As a first study of its kind, this, and most of the problems I have mentioned are logistical rather than methodological and, can be rectified with time.

Some confusion may have arisen during the presentation of this paper in regard to what stage of which part of the project I am in. It should be clear to the reader that the present paper is an effort to bring together the accumulated knowledge, to reach the true depth of nature of the lexical study I am carrying out. It is a correct procedure, I believe, in every linguistic study, to apply theory to a practical problem. This methodology is a guide to the making of Bantu word lists whose applications were described in section 2.

I acknowledge the oftentimes arbitrary choice of one procedure or even linguistic analysis over another where one is convenient for my objectives. This is acceptable if the choices appear to be equal.

A large part of this thesis is dedicated to parsing, an exercise that may not find much favor in current morphological work. However, it is inherent in the nature of the objectives of the paper; one cannot do text counts without parsing. Not much of this type of work has been done for the two languages under study so far, and it is hoped that this study will add to the body of linguistic literature on them. It also represents a step foward in Bantu language studies since I have no record of a similar study involving the same languages. As mentioned earlier, the results of this work will be made accessible to all kinds of scholars; these are reasonable contributions.

I have introduced the concept of ULA, one not only useful but also necessary in a lexical study with special relevance to Bantu and other related languages. As a paper on lexicological method and theory, it is an important headstart in future Bantu lexicology, but, its applications are not limited to word studies.

There are significant theoretical issues which arose from the discussion. The category "word" can only be applied in very specific sense and often unhelpful ways in the present study. The conclusions about Bantu ULA may shed further insights into this fuzzy category.

I have been able to work out a procedure and means of identifying nominal forms, compounds and derived forms as ULAs. The template, if adequate, is an exciting finding of the thesis. Apart from its lacking precedents, it contains promise of even greater advances in the search and study of typological universals. It may be imperfect but the extent to which it works will be the measure of its success.

As a theoretical base for a practical project, the writing of this thesis clearly reveals the true, significant level of linguistic theory that is demanded in lexicography and studies of lexicon. Real work may now continue.

List of References

Allwood, Martin Samuel,(1947), *Basic Swedish Word List, with English Equivalents, frequency grading and a statistical analysis,* Agustina Book Concern.

Armstrong, Lilias,(1967), *The Phonetic and Tonal Structure of Kikuyu,* Pall Mall, London.

Ashton,(1969), *Swahili Grammar including intonation,* 2nd ed., Longman, London.

Barlow, Ruffell A.,(1951), *Kikuyu Grammar and Idiom,* Blackwood and Sons, Edinburgh.

Benson, T.G.,(1964), *Kikuyu-English Dictionary,* Oxford.

Chang, Rodriguez y Alphonse J.,(1964), *Frequency Dictionary of Spanish Words,* Mouton & Co., New York.

Clements, George N,(1984), "Principles of Tone Assignment in Kikuyu" in *Autosegmental Studies in Bantu Tone,* Clements and J. Goldsmith (eds), Foris Publications, U.S.A.

Comrie, Bernard,(1989), *Language Universals and Linguistic Typology,* University of Chicago Press, 2nd ed..

Dabbs, Jack Autrey,(1966), *Word Frequency in Newspaper Bengali,* Dept. of Modern Languages, A. & M College of Texas.

Francis, W.N.,and H.Kucera,(1982), Frequency Analysis of English Usage:Lexicon and Grammar, Houghton Mifflin, Boston.

Gecaga, Mareka & Kirkaldy-Willis,(1960), *A Short Kikuyu Grammar,* Macmillan and Sons, London.

Gleason, H,(1961), *An Introduction to Descriptive Linguistics,* New York: Holt, Rinehart and Winston.

Guthrie Malcolm,(1948), *Comparative Bantu,* vols.I-IV, Farnborough: Greg Press.

Hamisi, Akida ...et al,(1981), *Kamusi ya Kiswahili Sanifu,* Taasisi ya Uchunguzi wa Kiswahili, University of Dares-Salaam, Oxford University Press.

Heine, Bernd, and William Möhlig (eds),(1980),(1981),(1982), *Language and Dialect Atlas of Kenya,* Dietrich Reimer Verlag, Berlin, Vols. I,II & III.

Hinnebusch, Thomas J. et al,(1979), *Kiswahili: Msingi wa Kusema, Kusoma na Kuandika,* University Press of America.

Jensen, John,(1990), *Morphology: Word Structure in Generative Grammar,* John Benjamins Publishing Company.

Johansson, Stig and Knut Hofland,(1989), *Frequency Analysis of English Grammar and Vocabulary* (Based on the LOB Corpus) Vol.I:Tag Frequencies and Word Frequencies), Clarendon Press, Oxford.

Johnson, Frederick,(1935), *Kamusi ya Kiswahili- yaani Kitabu cha Maneno* (Kiswahili-Kiswahili Dictionary), London, The Sheldon Press.

Johnson, Frederick,(1939), *A Swahili-English, English-Swahili Dictionary,* Oxford University Press.

Kiswahili: Jarida la Taasisi ya Uchunguzi wa Kiswahili (Journal of Kiswahili Studies, March 1974 and various others).

Kraph Ludwig,(1882), *A Dictionary of the Swahili Language,* Gregg Publications.

Lara, Fernando Luis, Chande & García Hidalgo,(1979), *Investigaones lingüísticas en lexicografía,* El Colegio de México.

Leech, John,(1971), *Semantics,* Penguin Classics.

Lyons, John,(1968), *Introduction to Theoretical Linguistics,* Camb.Univ. Press.

McGregor, Wallace,(1905), *A Grammar of the Kikuyu Language,* London.

Mchombo Sam & Bresnan Joan,(Jan.1993), UC Berkeley 'The Lexical Integrity Principle: Evidence from Bantu'(m/s).

Myers, Scott,(1987), 'Tone and the Structure of Words in Shona', Ph.D. Dissertation, University of Massachusetts.

Nurse Derek & T. Spear,(1985), *The Swahili: Reconstructing the History and Language of an African People, 500-1500*, University of Philadelphia.

Diccionario Básico del Español de México, El Colegio de Mexico, 1989,

Perrot, D.V.,(1965), *Swahili Dictionary*, Teach Yourself Series, London.

Polomé, Edgar,(1967), *Swahili Language Handbook* Center for applied linguistics.

Rubanza, Yunus Ismail,(1988), 'Linear Order in Haya Verbal Morphology: Theoretical Implications', Phd. Dissertation, MSU.

Schadeburg, Thilo,(1984), *A Sketch of Swahili Morphology*, Foris publications, 2nd ed..

Scotton-Myers, Carol, 'The origin of Ma'a: Codeswitching and Language change', University of S. Carolina, Columbia.[paper and handout read at the 23rd ACAL, MSU, March 1992.]

Spencer, Andrew,(1991), *Morphological Theory*, Basil Blackwell.

Vitale, Anthony,(1981), *Swahili Syntax*, Foris Publications.

Zgusta, Ladislav (ed),(1980), *Theory and Method in Lexicography*, Hornbeam press inc., Urbana.

Zgusta, L.,(1971), *Manual of Lexicography*, Hornbeam press.

APPENDIX A:
Sample page of untagged text.

HADITHI ZA MFALME SINSIN

mpaka akafikisha wanawake kumi. Kwa bahati njema huyu mke wa kumi akapata mimba.Wakati wote ule mfalme alipokuwa anatafuta mtoto kumbe aliweka nadhiri: "Kwamba pindi mke akishika mimba ya kwamza, ufikapo mwezi wa saba, nitaondoka na jeshi langu kwenda nchi nyingi- ne kukaa kwa muda mpaka mke wangu atakapojifungua. Halafu nitarudi ili nikute mtoto mchanga amekwisha zaliwa. Nitaingia kwa shangwe katika nchi yangu, kwenda kumbusu mrithi wa ufalme wa milki hii, mtoto wangu mpendwa." Alipoambiwa na wakunga kwa- mba sasa mkewe ana mimba ya miezi sita akakumbuka ahadi aliyo- iweka atimize. Akaandaa safari pamoja na jeshi la kumsindikiza, akaweka kila kitu tayari ili kuondoka mwezi wa saba wa mimba ya mkewe.

Alipokwisha andaa hivyo, wale wanawake tisa wasiojaliwa kupata mimba kila mmoja alikataa kufuatana na mfalme katika safari yake. Kumbe walikuwa na siri. Walimteta mke mwenzao na walimwonea wivu kupata mimba na wao wasipate mimba ijapokuwa waliolewa wadogo. Wakafikiri kwamba mke mdogo

APPENDIX B:

Sample Page of Tagged Text

mpaka aS kaT7 fikisha wanawake kumi. Kwa bahati njD ASema huyu mke wa kumi aS kaT7 pata mimba.Wakati wD ASote ule mfalme aS liT1 poR kuwa aS naT2 tafuta mtoto kumbe aS liT1 weka nadhiri: "Kwamba pindi mke aS kiT8 shika mimba ya kwanza, uS fika poR mwezi wa saba, niS taT5 ondoka na jeshi lE ASangu kwenda nchi nyD ASingine kuI kaa kwa muda mpaka mke wE ASangu aS takaT5 poR jiRF fungua. Halafu niS taT5 rudi ili niS kute mtoto mchanga aS mekwishaT3 zaliwa. niS taT5 ingia kwa shangwe katika nchi yangu, kwenda kuI mO busu mS rithi wa ufalme wa milki hii, mtoto wE ASangu mpendwa." aS liT1 poR ambiwa na wakunga kwamba sasa mkewe aS naT2 mimba ya miezi sita aS kaT7 kumbuka ahadi aS liT1 yoR iO weka aS timize. aS kaT7 andaa safari pamoja na jeshi la kuI mO sindikiza, aS kaT7 weka kila kitu tayari ili kuI ondoka mwezi wa saba wa mimba ya mkewe.

aS liT1 poR kwisha andaa hivyo, waS le wanawake tisa waS siN jaliwa kuI pata mimba kila mS ASmoja aS liT1 kataa kuI