1. Introduction

Ferdinand de Saussure's assertion that language is a principle of classification is played out in a particularly relentless way in languages with noun class systems. Among systems of linguistic categorization, noun class systems (including systems of grammatical gender, as in German or Arabic) are usually defined as follows:

(a) all nouns in the language are divided into a small and closed set of classes, signalled by inflectional morphology;
(b) the class of a noun is obligatorily co-referenced on other elements in the sentence via grammatical agreement (see e.g. Dixon 1982; Craig 1986).

Noun class systems are common, though by no means universal, among the languages of the world, and are an especially prominent areal feature in Africa (Heine 1982).

From a semantic point of view, the phenomenon of noun classification has been of interest to linguists and anthropologists because understanding the basis for grouping nouns together as members of a class hints at a system of cognitive or cultural classification underlying the system of linguistic classification. From a grammatical point of view, noun classes are interesting because they mediate between grammar and lexicon, and fall somewhere between inflection and derivation. Also, the analysis of grammatical agreement has played an important role in arguments for and against various models of syntactic theory.

The noun class system of Swahili is interesting first of all because, as a Bantu language, Swahili has the kind of complex and well-developed system that is the hallmark of that language family. But Swahili is atypical among Bantu languages in that its status as a lingua franca has led to the assimilation of an unusually large number of loanwords from genetically unrelated languages, especially Omani Arabic, Persian, and various Indian languages (and more recently English). The need to accommodate nouns of foreign origin, some of which fit the phonological forms but not the semantic content associated with the various noun classes, has had significant effects on the semantics of the noun classes (see Nurse and Hinnebusch 1993, chapter 3; henceforth N + H). Thus Swahili is an interesting case study for looking at continuity and change in noun class systems.

This paper is a report on research in progress on the semantics and syntax of noun classes in Swahili. It is organized as follows. In section 2 I describe the Swahili noun class system, discuss some earlier work on noun classes in this and related languages, and introduce the theoretical approach that is being used in the present study. In section 3 I explain how computer technology is being used in this research. Section 4 presents some results of the study: an analysis of the semantic structure of Classes 3 and 7. In Section 5 I discuss the relationship between the methodology and the analysis draw some general conclusions.
2.1. Outline of the Swahili noun class system.

Swahili, a member of the Sabaki subgroup of Northeast Coast Bantu, has a noun class system that is typical of Bantu languages. All nouns are divided into 11 classes\[^3\]. The class of a noun is signalled by

(a) a pair of prefixes attached to the nominal stem, one for singular, one for plural;
(b) a characteristic pattern of grammatical agreement, whereby possessive pronouns, demonstratives, verb subject and object prefixes, and other sentence elements coreferential with a noun are assigned a prefix that co-indexes the class of the noun, if it denotes an inanimate object. Sentence elements relating to nouns that denote animate beings are indexed by a special set of "animate concords", regardless of the prefix on the noun.

The following examples illustrate a portion of the noun class/agreement system\[^4\]:

\[(1)\]

(a. m-toto yu-le a-li-anguka  
Class 1-child Cl. 1-that Cl. 1 subj.-past-fall  
"that child fell"
  
b. wa-toto wa-le wa-li-anguka  
Class 2-child Cl. 2-that Cl. 2 subj.-past-fall  
"those children fell"

c. m-fuko u-le u-li-anguka  
Class 3-bag Cl. 3-that Cl. 3 subj.-past-fall  
"that bag fell"
  
d. mi-fuko i-le i-li-anguka  
Class 4-bag Cl. 4-that Cl. 3 subj.-past-fall  
"those bags fell"

e. ua li-le li-li-anguka  
flower (Cl. 5) Cl. 5-that Cl. 5 subj.-past-fall  
"that flower fell"
  
f. ma-ua ya-le ya-li-anguka  
Class 6-flower Cl. 6-that Cl. 6 subj.-past-fall  
"those flowers fell"
  
g. saa i-le li-li-anguka  
clock (Cl. 9) Cl. 9-that Cl. 9 subj.-past-fall  
"that clock fell"
  
h. saa zi-le zi-li-anguka  
Clock (Cl. 10) Cl. 10-that Cl. 10 subj.-past-fall  
"those clocks fell"

Comparison of (a) and (c), or (e), (g), and (h) above, shows that noun class cannot be determined solely by inspecting the form of the noun: the prefixes for Classes 1 and 3 (\(m\) in both cases) are homonymous; this is also true of Classes 9 and 10 (where the noun may have no prefix at all, so that singular- plural information is signalled only by the grammatical agreement). Many nouns in Class 5 also have no prefix, and so are indistinguishable, in the absence of grammatical agreement, from those of Classes 9/10. The agreement prefixes also show some homonymy (cf. (d) and (g)). This is why the definition of "noun class" crucially involves reference both to the prefix on the noun (if there is one) and to the pattern of grammatical agreement\[^5\].

2.2. Earlier treatments of the noun classes.

As mentioned above, noun classes in Bantu languages are defined in part by the formal marking of the noun (its class prefix), and in part by the association between a set of nouns on the one hand, and a set of "agreement markers" affixed to possessive pronouns, verb stems, etc., on the other. Although there is a wide range of opinions about whether the noun classes in Swahili and other Bantu languages have semantic content, there is great uniformity on the treatment of grammatical agreement. Agreement is assumed to be a purely syntactic phenomenon, in which the grammatical properties of one element in the sentence (the agreeing element, or "target", in the terminology of Corbett 1991) are determined by those of another element (the "controller", in this case the noun). In other words, it is assumed that "agreement" morphology contributes no independent semantic content to the message being communicated, but is merely a mechanical copying of features of the "controller" onto the "target". Modern treatments of grammatical agreement in Bantu and other languages, though differing in detail, are essentially a formalization of this inheritance from the Graeco-Roman grammatical tradition. Even more basic, so much so that it is never explicitly stated, is the assumption that "controllers", such as noun stems, "have" fixed grammatical properties (e.g. membership in a particular noun class); without this assumption one could hardly speak of copying features...
from the noun to, say, a coreferential demonstrative pronoun.

Given this prevailing view of the notion of agreement, the question of meaningfulness of the noun classes in Bantu languages has only been raised with respect to the prefixes attached to noun stems. The question that has been addressed is, can any regular semantic principles be identified to explain the assignment of noun stems to classes? Or put in another way, do the noun class prefixes have constant semantic content? Answers to this question, perhaps predictably, range from no to yes, but the majority opinion lies somewhere in between.

Perhaps the most extreme position on the "no" side is that of Irvine Richardson (1967:378), whose assertion that "...it is impossible to prove conclusively by any reputable methodology that nominal classification in Proto-Bantu was indeed widely based on conceptual implication..." is widely quoted, especially by those who disagree with it. At the other end of the spectrum are those who have tried to define each noun class in terms of a single abstract meaning, such as Denny and Creider's (1976) analysis of Proto-Bantu, and Zawawi's (1979) analysis of noun classes in Swahili.

Denny and Creider argue that Proto-Bantu had two subsystems of categorization, with partially overlapping morphology, one for count nouns and one for mass nouns. The count noun categories are further subdivided into "kind" classes, that identify objects as animates vs. artifact, and "spatial configuration" classes, that subclassify objects according to shape. The mass noun categories make a distinction between "cohesive" (substances that stick together) and "dispersive" (substances composed of dry particles that are readily dispersed).

In the same spirit, though differing in content, Zawawi (1979) assigns a single invariant meaning to each noun class prefix in Swahili. Zawawi's analysis is very innovative: she abandons the traditional criteria for definition of the classes, pointing out correctly that they are often inconsistent, and even groups together prefixes traditionally treated as homonyms (e.g. the m of Classes 1 and 3). According to her analysis, the singular classes subclassify nouns as "substance of life" (m-, traditionally divided into Classes 1 and 3), "substance of abstractness" (u, traditionally defined as a merger of Classes 11 and 14), "comparison of size or manner" (k, Class 7), "intensification" (zero ~ ji, Class 5), "large" (ba, not a traditional Bantu class), and a residual "catch-all" category (zero ~ r, Classes 9/10).

Unfortunately, analyses of the noun classes in terms of invariant meaning have failed to convince the skeptics, because there are always some examples that conflict with the invariant meanings that are posited. For instance, Denny and Creider's definition of Proto-Bantu Class 3 as "solid, extended in one dimension" does not seem to cover terms like 'ancestral spirit', 'heart', 'moon', 'bellows', 'load', 'daylight', etc. And Zawawi's definition of m- as "substance of life, singular" does not cover terms like 'metal chain', 'ramrod', 'town', 'irrigation ditch', deverbal nouns, etc.

The middle-of-the-road position on the semantics of the noun classes is to divide the noun classes into two subsets: a "derived" set of classes, assumed to be meaningful, to which noun stems from any class can be freely assigned with predictable effects on meaning, and an "inherent" set of classes, whose membership is largely arbitrary. Formally, these sets overlap: the same morphology is used for both "inherent" and "derived" class. The "derived" classes include the diminutive classes (with ki-/vi- prefixes, homonymous with the prefixes of Classes 7/8) and the augmentative classes (with zero/ma- prefixes, homonymous with the prefixes of Classes 5/6). With respect to the "inherent" classes, it is usually argued that although some semantic generalizations can be made about the groupings of nouns into classes, there is also a great deal of arbitrariness. It is often surmised that the present, disorganized system is a breakdown of an earlier, more coherent system assumed to have existed in the ancestor language. Examples of studies employing the "inherent/derived" distinction are Givon (1972), for ChiBemba; Heine (1982), who uses the terms "free" vs. "fixed" gender; Reynolds (1989), Reynolds and Eastman (1989), and Nurse and Hinnenburg (1993) for Swahili.

Although it is useful to distinguish between productive and non-productive processes of noun formation, the "inherent/derived" distinction ignores the question whether there are any semantic regularities in "inherent" classes, and also ignores semantic relationships, if any, between "inherent" and "derived" class markers. Rarely is any attempt made to connect the various groupings of nouns in a given class with one another, to investigate systematic relationships among different classes, or to explain the exceptions to the generalizations (such as names of animals that are not in the "animal" class). Also, the claim that the modern languages represent a breakdown of an earlier, more coherent system that used to exist in the ancestor language has no evidence to support it. As can be seen from Denny and Creider's (1976) list, the Proto-Bantu noun class system also had many apparent anomalies-- if it didn't, there would be no controversy about whether the Proto-Bantu noun classes were meaningful (see also Herbert 1985). In fact, claims about a mythical, semantically transparent system assumed to have existed in an ancestor language are commonplace in discussions of noun categorization, not only in Bantu (cf. Meillet 1923 on gender in Indo-European), yet no modern noun class language is attested with such a transparent system. It seems implausible to attribute a property to an ancestral language that is not found in any language of which we have direct knowledge. But if noun class systems are so full of anomalies, why do they persist for so long essentially intact (in the case of Bantu, some 3000 years or more)?

The problem seems to lie not with the languages, but with the assumptions about the nature of linguistic categorization that are brought to bear on this question. It seems to be assumed that either noun class semantics must be defined in terms of a set of common properties shared by all nouns in a given class, or one must abandon the search for semantic coherence and settle for a heterogeneous list. As pointed out by Lakoff (1987), this assumption is based on a view of linguistic categories as equivalent to sets in Aristotelian logic, which must be defined in terms of a set of necessary and sufficient conditions for membership. This view of linguistic categorization has been widely challenged in recent years, especially from the point of view of Cognitive Grammar (cf. Lakoff 1987; Langacker 1987, 1990; Rudzka-Ostyn 1988). It has been argued that membership in a given linguistic category (for example, a noun class) may be based on multiple criteria, including "family resemblances", metaphor, and metonymy, and that linguistic categories may exhibit an internal structure in which some members of the category are more central, or prototypical, and others are more peripheral. Although work within Cognitive Grammar has tended to concentrate on the semantics of individual lexical items, there are some detailed and illuminating studies of noun categorization that make use of similar insights: the work of Zubin and Kopcke on gender in German (Zubin and Kopcke...
3. Methodology

This project is being conducted in two phases. In phase I (still in progress), a database is being compiled of all nouns listed in the *Standard Swahili-English Dictionary* (Johnson 1939, henceforth SSED), using a commercial database program, DBase IV. Phase II of the project will involve investigation of contemporary usage of the noun class system in connected discourse, with special attention to loanwords, neologisms, and other phenomena that challenge the resources of the system. The two phases of the project are described in the next two subsections.

3.1 The noun database.

As mentioned above, all nouns from the SSED are being entered into a database[6]. So far 3976 nouns have been entered, which includes relatively complete coverage for Classes 1/2, 3/4, 5/6, 7/8, and 9/10, but only part of Class 11[7]. Each noun in the database is subcategorized according to a combination of morphological and semantic criteria. The morphological criteria are:

(a) noun class affiliation, using the traditional Bantu numbering system;
(b) if the noun is derived, the source of the derivation (e.g. verb stem, adjective stem, etc.).

Most of the categories of the database are semantic, since this is the purpose of the enterprise. For each noun I have included its dictionary definition, as well as classifying its meaning according to several semantic categories, each constituting a separate "field" of the database. The major categories are HUMAN, ANIMAL, PLANT, SHAPE, SIZE, AFFECT, FORCE OF NATURE (such as wind, rain, etc.), and NUMBER[8]. Within each of these categories further, more specific information is provided. For example, within the field HUMAN the nouns are further subclassified into agentive (denoting the agent of an action), kinship term, religious (e.g. "prophet", "saint"), occupation (e.g. "tailor"), etc. (To browse through the database structure in detail, click on the word database. To see a sample of the database with nouns tagged, click on the word sample.)

Use of a database has both advantages and disadvantages. The obvious advantage is that the database makes it possible to store and manipulate very large amounts of data, and to sort it in any number of different ways. Thus a few keystrokes can generate a list of all the nouns in Class 5, all nouns referring to animals, all nouns borrowed from Persian, all large three-dimensional hollow objects in Class 9, etc. The database can also be used for other purposes besides those for which it was originally designed. A dictionary is a kind of "culture inventory": just looking at what kinds of things are named, which semantic areas are highly differentiated and which are not, yields insight into the interests and preoccupations of the speakers of that language. Scholars and students interested in particular aspects of culture-- for example, kinship and social organization, material culture, religion and so on-- can instantly generate a list of all the nouns in the database that are relevant to that topic. Scholars interested in the nature of cultural contacts between the Swahili-speaking people and other societies can generate lists of loanwords and their languages of origin.

Disadvantages of this research method are both practical and theoretical. On the practical side, entering all the nouns from a dictionary onto a database is obviously very time consuming. But one could also regard this as an advantage: reading the dictionary does allow (or force) one to become intimately familiar with the data, a point that I will return to below. A second practical problem is how to avoid entering redundant records. A dictionary may list the same word more than once. (For example, the SSED sometimes lists derived nouns both as separate entries and as sub-entries under the source of the derivation.) This problem was avoided by writing a program for DBase IV that would automatically scan the database for homographic entries each time a new noun was entered. Since it is sometimes desirable to include multiple homographs in the database (i.e. in cases of homonymy), the program was designed to display all previous examples of the relevant form, in order to make it possible to decide whether a new entry was necessary.

The database project also raises some theoretical issues. First and most important, the problem of the semantic categories used to tag the nouns. In order to create a database, one has to anticipate which classificatory categories will be useful before entering the data, in a way guessing at the very analysis that the tool is intended to help discover. Use of a bilingual dictionary potentially adds to this problem, by introducing (or imposing) semantic categories of English that may or may not be relevant to Swahili. How can I be sure that I am not just projecting English-based categories onto Swahili? The short answer is, of course, that there is no general way to insure against this. It is the familiar problem of working from the "etic" to the "emic" (in the terminology of Pike 1967). In practice, I tried to minimize the problem by drawing on previous work on noun classification in Bantu languages, especially Spiltnik's (1987, 1989) work on ChiBemba, Zawawi's (1979) work on Swahili, and Denny and Creider's (1976) work on Proto-Bantu, as well as cross-linguistic studies of noun categorization, for example Adams and Conklin (1973), Craig (1986). My categorization by animacy, shape, and size, for example, was motivated by the very high frequency of these types of categorization among the languages of the world. On the other hand, the categories of AFFECT (positive or negative connotation) and FORCE OF NATURE were motivated by earlier studies specifically of Bantu languages.
Even so, I found it necessary to modify the database in various ways as I went along. Some tags that were used initially turned out to be too general, and more specific subdivisions were necessary. For example, the tag “artifact” (a subcategory of INANIMATE) originally included names for articles of clothing, but after discovering large numbers of these I decided to create a separate tag for clothing terms. Several more specific tags were added in this way, in response to the data. Of course, modifying the tags is very laborious, because one has to change the tags for all the words of that category that have already been entered, in order to maintain consistency. But this process also shows the importance of entering the data oneself rather than saving time by delegating the chore to a research assistant, because in the end there is no substitute for total immersion: unanticipated categories cannot be discovered any other way.

A second problem, or set of problems, comes from the use of a dictionary as data source. A dictionary is an analysis, not just a description. The compilers make choices about which words to include, how many entries to make for a given form (the familiar problems of polysemy and homonymy), and how to deal with geographic and social variation in pronunciation, grammar, meaning, and usage. Without doing extensive archival research, the user of the dictionary has no way of knowing exactly whose language is represented in it. Moreover, a dictionary is a sociolinguistic act (cf. Hymes 1974 and even more to the point, Fabian 1986). It is produced by people with a certain socio-cultural background, for a certain intended audience, and with certain goals in mind. The SSED, for example, was compiled by a British colonial committee constituted in 1930, composed in part of Christian missionaries and intended for an English-speaking colonial audience (for detail, see Whiteley 1969). Without intending criticism of the compilers, who supply a wealth of cultural information about medicinal properties of plants, beliefs about various kinds of spirits, proverbs and expressions illustrating usage, and the like, it would be impossible for such a dictionary to be devoid of ethnocentric bias. It is not hard to find obvious examples, e.g. several varieties of fish defined only as "fish, not considered good eating by Europeans". Although examples like this are not typical of the dictionary as a whole, it is still inevitable that in defining a Swahili term for an English-speaking audience of European cultural background, points regarded as worthy of comment or elaboration would be those where there is a perceived contrast between Swahili language/culture and that of English-speaking Europeans. An evident example of this is the relative paucity of loanwords from English in this dictionary-- in contrast, for example, with the more recent Kamusi ya Kiswahili Sanifu (Taasisi 1981). Some English loanwords are included in the dictionary, to be sure, but given that English had been the language of the colonial authorities since the end of World War I, one would expect a larger number of English words to be used by Swahili speakers than is reflected in the entries. Presumably the compilers of the SSED assumed that their readers would already be familiar with English, so by and large would not need definitions of words borrowed from English. In any case, the only way of compensating for ethnocentrism in the dictionary is to learn as much as possible about the language and culture from other sources as well.

A second important problem with the SSED as data source derives from the compilers' goal of standardizing the language. Because dictionaries usually have a prescriptive function as well as a descriptive one, it is hard to determine how much variation in the language is being concealed in order to encourage uniformity of usage. For example, from my knowledge of the contemporary language, I expected to find a fair amount of variation in noun class assignment, especially between Classes 5 and 9, both of which contain large numbers of loanwords, and both of which have zero as the most frequent allomorph of the noun class prefix. However, only about 3% of the nouns entered so far were listed in the dictionary as variable in noun class membership. Has the situation changed, or were the dictionary makers trying to impose conformity on the data? There is no way to tell. Also, even for those nouns that are listed as variable, no information is given about the nature of the variation: does it stem from dialect variation? variation among individual speakers? variation based on discourse context? Again, there is no way to tell.

For the reasons just outlined, it is desirable to supplement the material obtained from the 1939 dictionary with a wider range of data, especially data from contemporary discourse. Discourse data is important because that is the place to look for the areas of uncertainty and variability in meaning and usage that are represented only sporadically in the SSED. Also, this is the place to find neologisms, loanwords, slang, and other innovative usages that may or may not make it into a dictionary. Looking at how these "unclassified" words interact with the noun class/agreement system should shed light on the semantic reality of the noun classes themselves, and on the semantic functions of the agreement system. This is the plan for Phase II of the project. [Contents]

3.2 Investigation of noun classes in discourse.

For Phase II of the project, I plan to use the electronic corpus of Swahili texts that is currently being compiled by the Department of Asian and African Studies of the University of Helsinki, Finland together with the Institute for Kiswahili Research of the University of Dar es Salaam, Tanzania. The corpus, housed at the University of Helsinki, contains prose texts in Standard Swahili, from books and newspapers, and transcriptions of folkloristic material. At present the corpus consists only of the texts, preserving as much as possible of their original format. They have not yet been coded for morphological and syntactic information, but Arvi Hurskainen of the University of Helsinki is working on this. He has already developed a morphological parser for Swahili, called SWATWOL (see Hurskainen 1992 for details), but no facilities for automatic syntactic or semantic analysis of Swahili text are available as yet. However, some text retrieving programs are available, which produce concordances with context ranging from a line to a sentence in length.[9]

I already have a list of all wordforms in the Helsinki corpus that are not included in the dictionary compiled by the Institute for Kiswahili Research (Taasisi 1981), with their sentential contexts, for which I profusely thank Professor Hurskainen. The purpose of generating this list was to collect examples of neologisms and loanwords that have not (yet) been "codified", to look at their syntactic behavior. A preliminary (eyeball) scan of the data has already uncovered some interesting examples of agreement with conjoined noun phrases, treatment of acronyms, and nouns with variable agreement patterns.

Looking through large quantities of electronic text is very laborious, of course. But pending development of a parsing program that will automatically co-index nouns with "agreeing" elements, other expedients have to be used. One possibility that I have been exploring is to use the output from a morphological parser as the basis for syntactic tagging. I already had access to AINI, a morphological parsing program for Swahili developed by Thilo Schadeberg and P.S.E. Elias of the University of Leiden, the Netherlands (Schadeberg and
As mentioned in Section 2, the Swahili noun classes have differing degrees of internal coherence in their semantic structure. At one extreme are Classes 1/2, often called the "human" classes, whose membership consists almost entirely of nouns denoting human beings, especially agents of actions. At the other extreme are Classes 5/6 and 9/10, which have absorbed the majority of foreign loanwords, and may already have been fairly homogeneous even before the major influx of loans from Omani Arabic, dating only from the 17th century (N + H 1993:320). In this section I will propose an analysis of two classes that fall somewhere in the middle, Classes 3 and 7. The reason for starting here rather than with the more heterogeneous classes is that it is easier to work from greater coherence toward lesser coherence rather than the other way around; in fact the principles of organization discovered in the more coherent classes may help to elucidate the internal structure of the less transparent ones.

The following diagram is a schematic representation of the semantic structure of Class 3. The top-to-bottom organization moves from the more general to the more specific, but the diagram is not intended to be a "taxonomy" in the technical sense (cf. Casson 1981:75-77). I have borrowed the conventions used by Langacker (1988) for the representation of a linguistic category. Langacker defines two basic types of semantic relationship among the elements in a category:

(a) relations of "schematicity", in which one element is an "elaboration" or "instantiation" of another, more abstract element (represented by solid lines in the diagram);
(b) relations of "extension", in which some feature specifications are suspended or modified, while other features are retained (represented by dotted lines in the diagram).

Examples of nouns in each category can be viewed by clicking on that part of the diagram. [Note for text-only readers: for a text version of this diagram, use the right arrow key.]

This analysis incorporates many of the insights and observations of earlier studies of Swahili, such as Ashton (1944:23), Polome (1967:97), Hinnebusch (1979:230), Zawawi (1979:chapter 5). What I have tried to do here is make explicit the connections among the various semantic categories that had been identified by others, and add some categories that have not been mentioned before.

The topmost category in the chart, "entities with vitality", is what Langacker (1988) calls a "superschema", a maximally abstract category that holds together the various subcategories. However, since some of the subcategories are connected to the rest of the group by relationships of semantic extension, it cannot be claimed that the "superschema" includes everything below it in the chart, in the strict sense of class inclusion. "Vitality" is meant to capture various attributes of living beings, including growing and reproducing (true of plants, and metaphorically of human collectivities), but also ability to move (active body parts), to act on or affect other entities in the world or to occur independently of human volition (supernatural and natural phenomena).

The categories "exceptional animals" and "human collectivities" require some additional comments. These categories appear to be in this class not only because of properties that they share with other members of the class, but also because of the opposition between this class and other classes in the Swahili noun class system. First, the animals. In Swahili, as in other Bantu languages, most terms denoting animals are in Classes 9/10. This can be thought of as the "default" class for animals, especially mammals, virtually all of which are in Class 9/10 in Swahili. All animal terms that are not in 9/10 are therefore exceptional in some way. If one looks at the distribution of animal terms outside 9/10, the situation is as follows:

(a) those in Classes 1/2 (the "human" classes) are either generic terms for whole groups of animal species, agentive nouns, or terms derived from Class 1 nouns (already pointed out in note 10);
(b) those in Classes 7/8 denote small animals (see chart of Class 7, below);
(c) those in Classes 5/6 denote either animals that are large for their type (such as kunguru 'carion crow', panzi 'grasshopper', moma 'puff adder'), or non-mammals (e.g. kaya 'kind of shellfish', tekene 'jigger, burrowing flea', zamburu 'fish found in mangrove swamps'.

The animal terms that are in Class 3 are unusual either because of their appearance (swordfish), their behavior (kingfisher, Golden Weaver finch, termite, cuttlefish), or the fact that they do not fit easily into established categories (the eel is like a snake, but also like a fish; leeches and intestinal worms have an unusually intimate relationship with the human body). These animals are exceptional not just because they are not prototypical: each is remarkable in some way. Perhaps it is not a coincidence that they are placed in the class whose prefix is homonymous with that of the "human" class, m-.

The case of human collectivities is somewhat similar. These are entities that include human beings, but are not themselves human, so
Among all the subcategories of Class 3, that of plants/trees can be regarded as the most central, for several reasons. First, this subcategory contains the largest number of terms (almost half of all the nouns in the class). Second, it is productive in two ways: loanwords denoting trees and plants are almost invariably assigned to this class regardless whether the original term had an initial m-; trees and plants form the model for the majority of metaphorical and metonymic extensions in Class 3.

The subcategories below the second level on the diagram are interrelated in several ways, and in fact the degree of differentiation suggested by the diagram may be excessive. Many terms in the subcategories "extended things", "active things", and "extended parts of things" are also made of wood, so they could also have been placed in the subcategory "objects made of plants". My reason for separating these categories is to suggest plausible avenues for semantic extensions, not to claim that these are the only possible avenues, or that they are mutually exclusive. To the extent that a term fits into more than one category, it can also be regarded as well-entrenched within the semantic network of Class 3.

The most salient aspect of trees and plants, from the point of view of the subcategories associated with them, is shape, i.e. extendedness in one dimension[12]. This physical attribute is the basis for the inclusion in this class of some inanimate objects not made of plants, such as "nail", "ramrod", "metal chain", and of extended body parts of humans and animals ("bone", "blood vessels", "sinew", "porcupine spine", etc.). The inclusion of objects made of plants ("wooden platter", "straw mat", etc.), on the other hand, is based on a different sort of semantic extension.

The inclusion of long body parts in Class 3 motivates a further extension, by metonymy, to coverings that are wrapped around the body. This seems to be a recent extension within Swahili: none of the nouns in this group are reconstructed for Proto-Sabaki or Proto-Swahili by N + H, and some of them are loanwords from Arabic (but significantly, these did not originally begin with m-). The distinction between "powerful things" and "active things" may not be intuitively obvious. What I mean by "powerful things" is inanimate objects that have effects on human beings, such as substances with curative properties or religious objects. Some of these are made of plants, so they are connected both to this and to the category "supernatural phenomena", since they derive their power from some agency other than human. "Active things" are things, especially tools, that have movement as a salient characteristic: 'arrow', 'pestle', 'chopper', 'loom pedal', etc. In contrast to the "powerful things", they must be set in motion by a human agent. In this way they are similar to the "active body parts", which move but do not have independent volition.

Perhaps the most abstract distillation of the "entities with vitality" category is the use of Class 3 to derive deverbal nouns referring to the verbal process itself, such as mparuro `a scratching' (from paruru `to scratch'), mfuo `a hammering' (from fua `to hammer'), mlio `a sound' (from -lia `to make a sound'), etc. Such nouns describe a process as a thing, and so fit well with the other liminal entities in this class, that fall somewhere between animate and inanimate.

Let us move on to Class 7. [Note for text-only readers: for a text version of this diagram, use the right arrow key.]

Denny and Creider (1986 [1976]:223) state that the "primary meaning of Proto-Bantu Class 7 is instrumental artifact". If one interprets "primary meaning" as "prototypical meaning", I think this is right, and that it is still true of Swahili. I have added to this the specification "small enough to hold in the hand", because this applies to the majority of terms for instrumental artifacts in Class 7, and it provides a motivation for the major semantic extension within this class, to "small entities in general", not all of which are instrumental artifacts.

Among "small entities" there are several subgroups, most of which are self-explanatory. I will comment on the ones that are less obvious. First, the category "pieces/parts of things". Parts or subdivisions of things are smaller than the whole, so this category includes both reference to size and an implicit comparison between the part and the whole (recall Zawawi's 1979:115 definition of this class as "comparison of size or manner"). This part-whole comparison is carried over into a further extension, to "shortened things", that is things that have been truncated through being worn down or cut, and to terms for objects with physical defects, conceived as not-whole. These latter terms generally have derogatory connotations in Swahili which, not coincidentally, have a single adjective for "whole", "healthy", and "adult/mature", i.e. zima. As pointed out by Denny and Creider (ibid.), it is a fairly natural extension from 'used object' to 'despised object'. I argue here that the metaphor of size plays a role in this extension in Swahili.

The salient characteristic of terms in the subcategory "pointed things" is that a point or angle occupies a small amount of space. Even if the whole is large, such as a mountain, the pointed part ('peak') is relatively small. The same is true of the chin in relation to the face, and of the joints of the arm and leg. In the case of an arrowhead, the whole is both pointed and small. In the case of pointed parts too, there is an implicit comparison to the whole, but in this case the part is not cut off from the whole.

The category "part of substance" is more abstract than the ones just discussed, but its connection to part-whole relations is nonetheless apparent. This category includes terms denoting subdivisions of time and space. Height, depth, and units of measurement divide and delimit potentially extended spaces or spans of time into measurable parts. Here, too, there is an implicit comparison between the measured entity and the undelimited remainder.

Swahili grammars often point out that the prefix ki- is used to derive "adverbs of manner" (cf. Ashton 1944:165; Polome 1967:100), but they do not explain why this particular prefix should be used for this function. Zawawi's (1979:115) suggestive definition for the Class 7 prefix ki, "comparison of size or manner", is the first to make an explicit connection between these ideas, but Zawawi does not go on to elucidate the nature of the connection. I believe the link lies in a metaphorical extension of the part-whole relation to qualities or attributes, in a way reminiscent of the English expression "a chip off the old block". The relationship of similarity can be thought of as a kind of overlapping of substance between the entities that are regarded as similar. The overlap is partial, not total, because otherwise
Finally, the terms in the category "ailments associated with body parts" are all connected to their respective body parts by metonymic extension. Although some of them are actually associated with small body parts ('inflammation of the eyelid'), one could easily imagine this category being extended to other body parts via metonymy.

Of the 490 Class 7 nouns entered into the database, 39, or about 8%, do not obviously fit into the above categories. In some cases, not enough information is provided in the dictionary gloss to decide whether they fit or not (for example, kilendo 'kind of fish'; kilua 'kind of sweet smelling flower'-- are they small?). Some may be deverbal nouns or "similarity" nouns whose source is now obsolete (e.g. kisutu 'food received after a task has been completed'; kifabakazi 'Nandi flame tree'). Some are loanwords that had ki- as initial syllable (e.g. kisirani 'mishap, misfortune'; kisasi 'vengeance, retaliation', both from Arabic).

There is one set of apparent anomalies that deserves additional comment, however. This is a group of terms referring to large, dangerous animals or birds: kifaru 'rhinoceros', kiboko 'hippopotamus', kingugwa 'large spotted hyena', kipungu 'eagle', and kipanga 'Dickinson's falcon'. These terms are strikingly anomalous: why should large, predatory animals be placed in a class whose most prominent characteristic is small size, often with a connotation of insignificance? Interestingly, three of these terms ('hippopotamus', 'rhinoceros', and 'eagle') are replacements for terms that were originally in Class 9 (the "animal" class) in Proto-Sabaki (N + H 296). In other words, these three animals were moved from the "animal" class to Class 7 (kipanga 'Dickinson's falcon' was already in Class 7 in PSA, and N + H do not give a reconstructed form for 'spotted hyena'). One possible explanation is that these terms started out as euphemisms. It is a common phenomenon for people to replace terms for threatening or taboo things with more harmless terms. Putting names of large, dangerous animals in the class of small, manipulable things could be a way of figuratively neutralizing or diminishing their power. [Contents]

Conclusion

In conclusion, I would like to make two main points about the analysis and methodology described above.

First, the principles of cognitive grammar are a useful tool for investigating the semantic structure of noun classes. This approach explicitly recognizes the fact that human beings use linguistic categories to make sense out of the world, and it provides a cognitively motivated framework for describing associative relations among the members of a category. The principles by which different nouns are grouped together into a class are similar to those that govern the connections among the various "senses" of an individual lexical item. These same principles, including metaphor and metonymy, also play an important role in lexical and grammatical change (see also Heine and Claudi 1986).

At the same time, analyses along these lines do not attempt to predict the content of a given category or the direction of meaning change. Although some developments are more likely than others, the open-ended nature of the messages that people may want to communicate militates against the possibility of absolute prediction. What this type of analysis does show is that the groupings that are found are semantically motivated rather than arbitrary. In this respect it is an advance over the point of view that linguistic categories must either be definable in terms of Aristotle's necessary and sufficient conditions for membership, or dismissed as incoherent.

The second conclusion to be drawn from this study concerns the use of the database. It must be emphasized that a database is not a discovery procedure for semantic structure. In fact, comparison of the categories in the diagrams in Section 4 with the tags used in the database shows that the tags are only indirectly reflected in the diagrams. Some tags, such as "body part", "animal", turn out to require greater differentiation, in ways not originally anticipated when the database structure was conceived. Others, such as "human collectivities" or "part of substance", were discovered as a result of inspecting larger groups of nouns, the semantic network as a whole, or the intersections between the noun classes and wider semantic domains. However, this is not meant to suggest that one could select at random a dictionary of a language that one did not know, create a database, and discover the semantic structure of its noun classes. The database tags were not randomly related to the eventual analysis; they depended on a prior familiarity with Swahili language and culture, and on the relationships of noun classification in Swahili, Bantu, and other language families. The database is an extremely useful tool but like other tools, its limits are the limits of its users. [Contents]

References


In Guthrie's classification Swahili belongs to Group G. Overview. Swahili is one of the most widely spoken of the autochthonous African languages and the largest of the Bantu group. It originated in the coast and islands of East Africa but later spread inland being adopted as an international language for communication between the Indian Ocean commercial network and the Bantu interior. Swahili's grammar is typically Bantu, including a noun-class system and a complex agglutinating verbal morphology, but unusually for a Bantu language it lacks tones. Distribution. The east coast of Africa, from Somalia to North Mozambique including Kenya, Tanzania and many offshore islands.