Nordic Journal of African Studies 12(3): 405–423 (2003)

SEMANTIC ANALYSIS OF KISWAHILI WORDS USING THE SELF ORGANIZING MAP

WANJIKU NG'ANG'A
University of Helsinki, Finland

ABSTRACT

Acquisition of semantic knowledge to support natural language processing tasks is a non-trivial task, and more so if manually undertaken. This paper presents an automatic lexical acquisition method that learns semantic properties of Kiswahili words directly from data. The method exploits Kiswahili's system of nominal and concordial agreement that is inherently rich with semantic information, to capture the morphological and syntactic contexts of words. Classification of nouns and verbs into clusters of semantically-similar words is done based on this contextual encoding. The method uses training data from the Helsinki corpus of Kiswahili while the machine-learning component is implemented using the Self-organizing Map algorithm.

The proposed method offers an efficient and consistent way of augmenting lexicons with semantic information, where electronic corpora of the language in question are available. It also provides researchers with an investigative tool that can be used to identify dependencies within linguistic data and represent them in an understandable form, for further analysis.

Keywords: lexical acquisition, corpora, machine-learning

1. INTRODUCTION

Semantic Knowledge is of critical importance to natural language processing (NLP) applications, due to the ambiguous nature of natural language. Many words in a language usually have two or more associated meanings depending on the context of use. In such cases, it is imperative for a NLP system to resolve the ambiguity and select the appropriate sense of the ambiguous word. For example, when translating an ambiguous source language word, a Machine Translation (MT) system must choose one of its possible meanings, so as to select the intended or appropriate target language translation. Other NLP applications that benefit from semantic knowledge for ambiguity resolution include Information Retrieval, Speech Recognition, Part-of-speech tagging, syntax parsing etc.

However, acquisition of semantic knowledge is a challenging problem – one that has been referred to as the "knowledge-acquisition bottleneck", in the literature (Gale et al. 1992). Researchers have had to encode semantic knowledge manually for specific tasks, a process that is labour-intensive and therefore expensive. Consequently, the linguist usually limits information acquisition to suit just the task at hand, resulting in knowledge bases that cover only a small subset of the language. Moving to a new task entails repeating the manual encoding process to either augment the existing knowledge base or build a new one from scratch, to suit the new application. Updating the knowledge bases to cope with the dynamic nature of natural language where new words and meanings are created frequently becomes a daunting and expensive task if done manually. There have been few large-scale efforts to create broad semantic knowledge bases, such as *WordNet* (Miller 1990) and *Cyc* (Lenat et al. 1986) mainly for the English language. While these efforts may be useful for some applications, they may never fully satisfy the need for dynamic semantic knowledge.

The availability of natural language data in the form of massive computer-readable corpora, as well as the corresponding research in statistical techniques and machine-learning algorithms, has created the possibility to explore 'data-driven' approaches for linguistic analysis. Such approaches (corpus linguistics) are becoming increasingly important and provide new ways of deriving linguistic information and knowledge from the data itself. Natural language acquisition by computers is an area of much potential and recent research, with work focusing on replacing hand-built language parsers with models generated automatically by training on corpora (Berwick, 1985; Brill, 1993; Zelle & Mooney, 1993; Charniak, 1993; Magerman, 1994) and part-of-speech taggers (Charniak et al. 1993; Merialdo, 1994). Recent research efforts have concentrated on building systems that can automatically acquire lexical (semantic) information from data, with a view to eliminating or minimizing the problems associated with manual acquisition of lexical data. Such systems require large corpora as well as proper computational tools that can be used to identify hidden dependencies within the data and represent them in a form that is both understandable and that allows easy use of the obtained knowledge.

In this paper, I discuss an approach to automatic acquisition of semantic properties of Kiswahili words from corpora. In section two, a brief introduction to machine-learning is given, with particular emphasis on the Self-Organizing Map algorithm that is used in this study. The third section gives a brief introduction to the Kiswahili language and discusses linguistic features specific to Kiswahili that have been chosen to encode the context of a target word. The data used in the study, experiments and results are presented in the fourth section while discussions and conclusions are given in the final section.

2. TEXTUAL DATA MINING

Data mining is a term that refers to the process of fitting models to or determining patterns from very large datasets. Data mining also comprises the aspect of knowledge discovery, where new information may be derived from data. Textual Data Mining therefore refers to the process of discovering new information or determining patterns from textual (natural language) data.

Natural language corpora are primary sources of information about language use. They represent a huge linguistic knowledge bank that can be tapped through the use of various data analysis tools to discover trends, patterns or other linguistic phenomena which may be incorporated into other language processing tasks. For example, corpora can support detailed studies of how particular words are used, by providing extensive examples of natural language sentences in context. Information about word frequency, co-occurrence, collocations etc., can be derived from corpora, and used to build statistical language models, for word sense disambiguation or speech recognition.

There are various approaches to data mining such as machine-learning algorithms, statistical algorithms, example-based, rule-based or hybrid systems. Machine-learning is an area of artificial intelligence which concentrates on the development of techniques that allow computers to, in some sense, "learn", where learning can simply be defined as the gaining of knowledge (Mitchell 1997). Following is a brief discussion of Kohonen's Self Organizing Map that has been used in this study, and which is an example of a machine-learning algorithm.

2.1 SELF-ORGANIZING MAP (SOM)

The self-organizing map (Kohonen 1995) is an unsupervised artificial neural network model that is well suited for mapping complex and high-dimensional data into a two-dimensional representation space (map). Similarities between input patterns that are present in the n-dimensional input space are mirrored within the two-dimensional output space of the self-organizing map. Thus, on the SOM output space, the inputs are organized according to their cluster structure, i.e. inputs with similar properties are grouped together. The training process is based on weight vector adaptation with respect to the input vectors. The SOM has shown to be a highly effective tool for data visualization in a broad spectrum of application domains.

3. KISWAHILI LANGUAGE

Kiswahili has a typical Bantu noun class system where noun classes are signalled by a pair of prefixes attached to the nominal stem, one for singular and the other for plural. Affixes are used to mark various grammatical relations, such as subject, object, tense, aspect, and mood in the case of verbs. There is a system of concordial agreement in which nouns and other sentence constituents must agree with the verb of the sentence in class and number. Adjectives, possessive pronouns and demonstratives also agree in class and number with the noun they modify. Syntactic and functional information can therefore be derived from the meaning-bearing affixes attached to nouns, verbs and their dependent words, and used for semantic clustering. Kiswahili has a fixed word order with the subject preceding the verb and the object. Noun modifiers come after the noun just as verbal modifiers such as adverbs, follow the verb they modify. For languages with fixed word order such as English and Kiswahili, the distribution of words in the immediate context of a target word are rich in information regarding the semantic and syntactic properties of the target word. This provides an additional source of contextual information that can be used to obtain semantic clustering of Kiswahili words.

3.1 CONTEXTUAL FEATURES

The first step in any clustering algorithm is to represent the objects to be classified (words in this case), in terms of the values of their contextual features. This means that a word is represented by its context of use. The choice of contextual features (vectors) is very important as these determine the semantic similarities that will form the basis for cluster formation. For this study, contextual features were restricted to overtly-marked linguistic features.

3.1.1 Verbs

Contextual features for encoding verbs were obtained from the verb itself and from the word to its immediate right. As explained in a previous section, concordial agreement and aspect/mood which bear important functional information are marked with affixes in the verb. For example, in the following sentence,

> *Mama a-li-m-pik-i-a mtoto chakula*
> N subj-SG3 tense-PAST obj-SG3 V APPL final-vowel N N
> "Mother cooked food for the child"

the subject prefix (*a*-) and object prefix (*m*-) refer to class 1 and 2 nouns which usually, refer to animate nouns. Therefore, the subject prefix and object marker in the verb provide information about the type of agents and patients that a verb can take. Verbal extensions that express aspect and mood e.g. causative, applicative, reciprocity etc. also carry important semantic information. In the following sentence,

> *Maria a-li-ji-peleka nyumba-ni jana*
> N subj-SG3 tense-PAST obj-REFL V N-LOC ADV
> "Maria took herself home yesterday"

the reflexive marker (-*ji*-) and the locative marker in nouns (-*ni*) provide useful contextual clues, such as subject type and indicating a movement verb, respectively. Properties of the word to the immediate right of a target verb e.g. part-of-speech, finite or infinite (in the case of a verb) etc., provide important contextual information that is significant for semantic classification. For example,

Juma a-li-enda kwa dada-ke Mombasa
N subj-SG3 tense-PAST V preposition N-PRON N
"Juma went to his sister's in Mombasa"

in the above example, the preposition after the verb provides information that the verb may be a movement verb, while in

Ali a-na-taka ku-la ma-embe
N subj-SG3 tense-PRESENT V INF-V PL-N
"Ali wants to eat mangoes"

the infinitive verb immediately following the verb *taka* may be an indicator for a verb of desire or want. The complete list of contextual features used in the verb experiments are shown in the table below:

Feature	Explanation	Position
Loc	'ni' location suffix	+1
Obj	Noun class 1/2 Object marker	0
Rfx	Reflexive marker	0
Inf	Infinitive verb	+1
Fin	Finite verb	+1
PP	Preposition	+1
Adv	Adverb	+1
Con	Conjunction	+1
Noun	Noun	+1
Name	Proper Noun	+1
Rec	Reciprocal marker	0
Stat	Stative marker	0
Caus	Causative marker	0
Appl	Applicative marker	0
Pass	Passive marker	0
Spf_1/2	Subject prefix (class 1/2)	0
Any_spf	Subject prefix (any class)	0

Table 1. Contextual Features for Verb Clustering.

3.1.2 Nouns

Contextual features for encoding nouns were obtained from a three-word window [-1,0,+1] and comprise overtly-marked features as well as complementary semantic information derived from *WordNet*[1]. For example, in the following sentence,

Mama a-li-uza ng'ombe wa-zuri soko-ni
N subj-SG3 tense-PAST V N Class-marker-ADJ N-LOC
"Mother sold nice cows in the market "

[1] *WordNet* is an enumerative handcrafted knowledge base for English.

Mama belongs to noun class 9/10 but takes a class 1/2 (animate) subject prefix in the main verb and in the adjectival modifier. Therefore, the subject prefix from the head verb is chosen as a contextual feature, as opposed to the noun class prefix. The locative marker (*-ni*) is included as a feature that indicates a location or space. Plural and singular markers as well as numbers (both ordinal and cardinal) are chosen as additional features which provide clues as to whether a noun is a mass or count noun.

Various researchers have shown that words can be classified according to the predicate-argument structures they exhibit in a corpus (Hindle 1990). This idea is based on the premise that there exists certain restrictions in natural language, as to what words can be used together in the same construction. In particular, there are restrictions on what nouns can be arguments of what predicates i.e. there is a restricted set of verbs for which a given noun can be subject or object of. With this in mind, a set of contextual features that encode these restrictions were included as features for this study. The semantic type[2] of the predicates was obtained from *WordNet*, via the English translation of the Kiswahili verb, giving rise to features such as *is-a-subject-of-a-cognition-verb* or *is-an-object-of-a-contact verb*. For example, in the following sentence,

Mtoto a-li-imba- wimbo
N subj-SG3 tense-PAST V N
“The child sang a song”

the noun *mtoto* would be encoded by the feature *is-a-subject-of-a-creation-verb* while the noun *wimbo* would be encoded by the feature *is-an-obect-of-a-creation-verb*, where the English translation of the verb *imba* (sing) is coded as a verb of creation in *WordNet*[3]. The complete list of features used in the noun clustering experiments is shown in table 2 below:

[2] *WordNet* classifies all verbs into 15 categories (types) such as communication, creation, contact etc

[3] Where the verb has more than one semantic category associated with it, the first reading is chosen. *WordNet* lists senses for ambiguous words in order of frequency, where the most frequent reading is listed first.

Feature	Explanation	Position
Spf	Concordial (subject)prefixes	+1
Loc	Location suffix 'ni'	0
Num	Cardinal number	+1
Plu_Sg	Plural and Singular Marker	0
O_stat	Object of a stative verb	-1
O_body	Object of a body verb	-1
O_cons	Object of a consumption verb	-1
O_comm	Object of a communication verb	-1
O_cont	Object of a contact verb	-1
S_cogn	Subject of a cognition verb	+1
S_comm	Subject of a communication verb	+1
S_emot	Subject of an emotion verb	+1
S_body	Subject of a body verb	+1
S_poss	Subject of a possession verb	+1
S_social	Subject of a social verb	+1
S_comp	Subject of a competition verb	+1
S_cont	Subject of a contact verb	+1
S_motn	Subject of a motion verb	+1
S_stat	Subject of a stative verb	+1

Table 2. Contextual Features for Noun Clustering.

4. EXPERIMENTS

4.1 DATA

The data used for this experiment has been obtained from the Helsinki corpus of Kiswahili, located at the University of Helsinki's Language Corpus Server. The corpus consists of varied genres of texts such as newspaper articles, religious texts (Bible and Quran), parliamentary proceedings and standard book texts.

A total of thirty verbs (table 3) and sixty-five nouns (table 4) were selected for this study, based on their occurrence frequency as well as semantic diversity. Occurrences of these words were extracted from the corpus and pre-processed using the Swahili Language Manager (SALAMA).[4] These tools were used to perform the initial pre-processing of the raw texts, morphological analysis as well as morphological disambiguation. The output from this phase lists each word occurring in the input text, but with additional morphological and basic syntactic tags, as shown in figure 1. Any morphological or semantic ambiguities in the texts were left unresolved.

[4] SALAMA is a suite of computational tools that enables the efficient processing and analysis of Kiswahili texts. It comprises a text preprocessor and SWATWOL, a morphological analyzer and disambiguator, based on two-level morphology, developed at the University of Helsinki (see Hurskainen 1992: 87–122; 1996: 568–573).

```
"<*mtoto>" "mtoto"  N 1/2-SG  { child , (Kiskoti) bairn , young person , wa jicho cataract ,
inset , juvenile } @SUBJ &
"<aliimba>" "imba"  V 1/2-SG3-SP VFIN PAST z { sing , chant , twitter } SVO
@FMAINVtr+OBJ> &
"<wimbo>" "wimbo"  N 11/10-SG  DER:o { song } @OBJ &
"<mzuri>" "zuri"  ADJ A-INFL 11-SG  { beautiful , pretty , gorgeous , good } @<NADJ &
"<sana>" "sana"  AD-ADJ AR { much , very , a lot } @<A-ADJ  &
"<.$>"
```

Figure 1. Morphological analysis of the sentence **"***Mtoto aliimba wimbo mzuri sana***"**.

4.1.1 DATA MATRIX

The data matrix, **D**, can be described formally as follows:

> Let **W** be the set of words, $\mathbf{w_i} \in \mathbf{W}$, i = 1..n;
> Let **Q** be the set of morphological features, $\mathbf{q_j} \in \mathbf{Q}$, j = 1..m;
> if **D** represents the data vector, then the value of $\mathbf{d_{i,j}}$ represents the frequency of morphological feature $\mathbf{q_j}$ within the context of word $\mathbf{w_i}$, i.e. the value of $\mathbf{d_{i,j}}$ is a measure of how typical (frequent) the $\mathbf{j}^{th}$ morphological feature is, within the context of a the particular word, $\mathbf{w_i}$.

All the data vectors for both nouns and verbs were normalized by the total number of corpus occurrences for each word.

4.1.2 VERBS

Table 3 lists the 30 verbs used in the study. The frequency column indicates the total number of occurrences in the training corpus for each word. The verbs have been grouped into three major classes using Levin's classification for English verbs (Levin 1993). These classes will be used to compare and evaluate the clusters obtained using the SOM algorithm. The English translations have been obtained from TUKI's[5] Kiswahili-English dictionary.

VERB	**FREQUENCY**	
		VERBS OF COMMUNICATION (1)
sema	55947	speak, say; scold, speak against; advise, counsel; backbite, badmouth
eleza	8067	elucidate, describe, explain, brief; feel; be clear, be intelligible
ongea	6845	talk, chat; increase
zungumza	6146	talk, speak, discourse.
ambia	6019	tell sb sth; bid
jibu	6007	answer, respond, reply, react to
ita	4108	call, summon, beckon, phone; invite
andika	4058	write, pen, chronicle, inscribe; lay table for a meal; satirize.

[5] Taasisi ya Uchunguzi wa Kiswahili (Institute of Kiswahili Research), University of Dar es Salaam, Tanzania.

uliza	3913	ask, question, interrogate inquire
onyesha	1626	show, exhibit, demonstrate.
soma	3484	read; study, receive teaching; attend school; be educated; observe sb.
		VERBS OF PSYCHOLOGICAL STATE;PERCEPTION;DESIRE (2)
jua	9650	know
penda	4690	love, like; will
tazama	2636	look at, watch; help, aid; be wary/cautious
sikia	3653	hear; obey, pay attention; feel
hisi	732	feel, perceive, sense, envisage
taka	24918	wish, want, need; be on the verge of, be inclined
weza	20893	be capable, be able; have strength, have means, be in control
ona	18530	see; feel
		VERBS OF MOTION (3)
ja	11203	come, turn up
fika	10884	arrive (at), reach, find, get (to); be complete
pita	10155	pass; be out of date; be temporary; surpass, excel
kwenda	7733	go
ingia	7424	enter, get in; go into sth; incur; pierce; matriculate; join a group/association/party
fuata	6905	follow, come after, pursue; be with sb, be attached to; observe, comply, conform, abide
rudi	2684	return, reverse; repay; discipline, punish; shrink
tembea	2118	walk, move around, travel; fornicate.
panda	2086	climb, ascend, mount; rise; get upon, ride upon; plant, sow; bid
kimbia	1719	run; run away, escape
ruka	549	fly; jump, hop; deny, denounce, disown; shrink.

Table 3. List of Verbs.

4.1.3 Nouns

NOUN	FREQUENCY	
		[ANIMATE, +] [HUMAN, +][COUNT, +][UNIT, -][LOCATION, -]
mtu	31604	person, human being, individual.
Kiongozi	10905	leader,guide; manual, handbook.
rais	10900	president
mtoto	9826	child; young person; cataract; inset; juvenile
mwananchi	9212	citizen
waziri	7988	minister
mkuu	7760	important person, head, incharge, top man
mwandishi	5425	recorder, secretary; writer, author
mwanamke	5378	woman.
kijana	5128	youth, young man, juvenile, lad.
mama	4472	mother
ndugu	4404	kin, sibling; relative; close friend; comrade.
bwana	3792	mr, sir, husband; lord.
bibi	1632	grandmother; lady; wife; mistress, concubine; the queen (in a pack of playing cards).
daktari	935	medical doctor/officer; doctor of philosophy.
		[ANIMATE, +] [HUMAN, -][COUNT, +][UNIT, -][LOCATION, -]
ndege	2413	bird; aeroplane.
kuku	392	hen, chicken
tembo	155	elephant; coconut palm wine; changu-like fish.
panya	99	rat
chui	78	leopard; cruel person
chura	25	frog; someone who comes last in a play.
		[ANIMATE, -] [HUMAN, -][COUNT, +][UNIT, -] [LOCATION, -]
chama	16132	party; association, guild:
jambo	12445	matter, business, circumstances; difficulty, trouble

habari	9592	news, tidings, information; form
mkono	4900	arm; hand, handle; branch of river, creek of sea; cubit; half yard
sheria	4811	law, statute, code; decree, order, ordinance; regulations, obligation
gari	4021	car, vehicle
kitabu	2337	book
ugonjwa	1868	illness, disease, sickness
kiti	1296	chair, seat; deck planking; parliamentary seat; person possessed by spirits
mfuko	1120	bag; pocket; fund
jiwe	678	stone, holystone; weight of a balance; battery, cell
meza	550	table
		ANIMATE, -] [HUMAN, -][COUNT, -][UNIT, -][LOCATION, -]
wakati	16760	time, period of time, point of time; season; opportunity
maisha	5277	life
umoja	5016	unity, fellowship; singularity
Maendeleo	4460	development, progress, advance.
amani	4181	peace
maji	4146	water
nguvu	3954	force, strength, power; authority, supremacy; impetus, pressure, solidity.
elimu	3848	education, knowledge
uongozi	2819	leadership, management
ndoto	464	dream
ngano	174	fable, tale, story; wheat.
		ANIMATE, -] [HUMAN, -][COUNT, +][UNIT, -][LOCATION, +]
nchi	28786	country; land
mji	10810	capital; homestead; womb; central part of a khanga; trench in a grave in which the dead is put.
mkoa	7691	province, region; metal bar
nyumba	7084	house.
njia	5717	path, road, way; method, means.
dunia	5477	earth; world; life
jiji	5282	city
shule	5010	school
wilaya	4925	district
kijiji	4366	village
kanisa	4262	church; christian community
ofisi	3697	office
		[ANIMATE, -] [HUMAN, -][COUNT, +][UNIT, +][LOCATION, -]
mwaka	22728	year
siku	9774	day
mwezi	4412	moon; month
wiki	3416	week
juma	2048	week; proper noun
dakika	1030	minute
kilo	298	kilo
lita	196	litre

Table 4. List of Nouns.

Semantic properties (attributes) associated with groups of semantically-similar words have been presented using a feature-based notation, where a plus sign (+) indicates the presence of a property and a minus sign (-) indicates its absence. The semantic properties used for this test are consistent with the basic features obtained from *WordNet*. The ANIMATE feature distinguishes between animacy and inanimacy. HUMAN is used to distinguish human animate nouns from those that are non-human. COUNT is used to distinguish between count and mass nouns. LOCATION specifies if a noun refers to a place (extent in space) or not, while UNIT indicates whether a noun refers to a measure (quantity) or not.

4.2 Results

Once the data vectors were obtained, the SOM Toolbox (Vesanto et. al 2000) was used within the MATLAB environment to organize and visualize the Kiswahili words. For each word, the best-matching map unit (bmu) was obtained by locating the model vector that most closely resembles that of the data. The word label was then written onto the map lattice corresponding to the bmu, as shown in the following figures.

4.2.1 Verbs

Figure 2 below shows the SOM lattice obtained after analysing the verbs using the SOM algorithm. Each hexagon represents a map unit. As shown, the majority of the verbs have been placed close to those with similar semantic features (as given in table 3). The bottom units of the map contain mainly group 3 verbs (movement). Group 1 verbs (communication) are located in the upper right units while group 2 verbs (psychological state, perception and desire) appear mainly in the middle and top section of the map. The clusters are by and large, consistent with Levin's verb classes.

A. SOM LATTICE

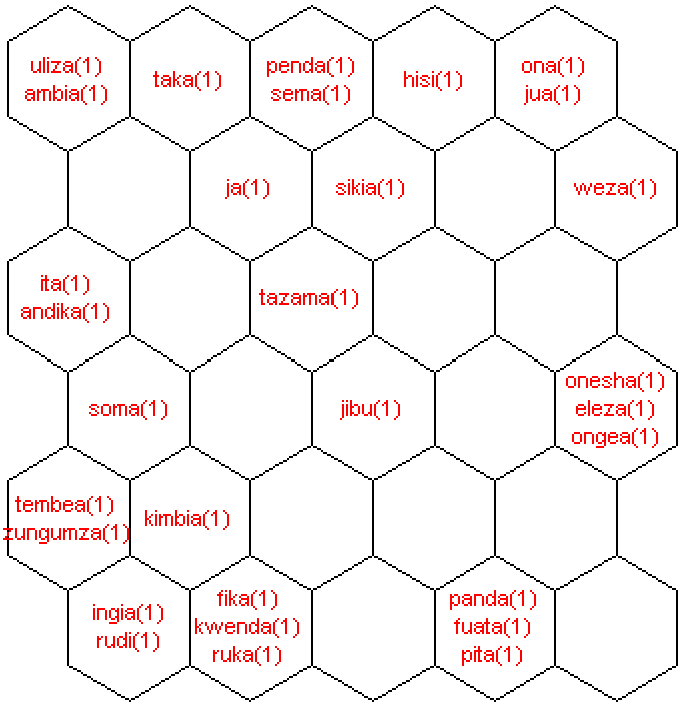

Figure 2. SOM analysis of the 30 verbs.

B. COMPONENT MAPS

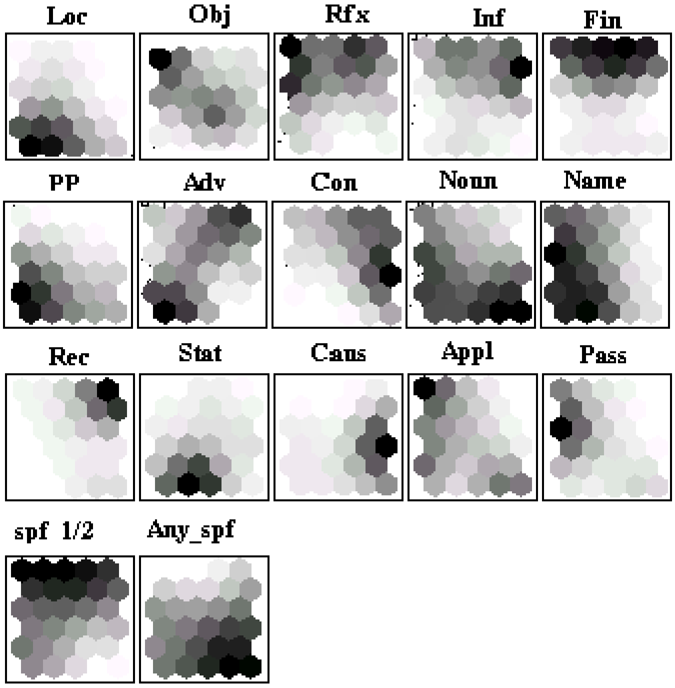

Figure 3. Component maps for 17 context features.

From the component maps shown in figure 3, it is possible to see which contextual features are important in the formation of particular clusters. The dark-shaded areas correspond to the SOM lattice area where the particular variable is very significant. By observing the component maps for variables **Loc** and **PP** and comparing their dark-shaded areas with the SOM lattice in figure 2, it is clear that these two features are correlated and play a significant role in the clustering of movement verbs. Similarly, and as would be expected, reflexive markers and infinitive verbs in the following word position are strong indicators for verbs expressing psychological states, desires or perception. Verbs from group 1 and group 2 occur quite close to each other, and this could be explained by the fact that they share very similar contexts and as such do not differ greatly with respect to the features used for this experiment.

C. COMPONENT VALUES

It is possible to inspect the values that each of the 17 components have for every map unit. This information is important since it clearly shows the contextual differences between two map units, and clusters, by extension. For example, figure 4 shows the values for each of the 17 components (left to right) in two map units that represent two different clusters. Figure 4(a) shows map unit [1,5] containing labels *ona* and *jua* (perception) while figure 4(b) shows map unit [6,2] which has labels *fika*, *kwenda* and *ruka* (movement).

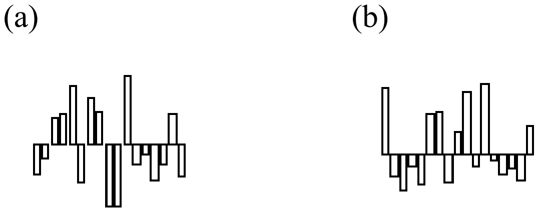

Figure 4. Component values.

Negative values depict lower frequency of the particular component in that map unit compared to the rest of the map, while a positive value indicates that this component is more frequent in this map unit than in the rest of the map, in general. For example, from figure 4(a), it is evident that features *Rfx, Inf, Fin, Rec* and *spf-1/2* are very significant for perception verbs, while *Loc, PP, Adv, Name* and *Any-spf* signify important for movement verbs, as depicted in figure 4(b). The above figures illustrate the difference in each of the component values for different areas of the SOM lattice

4.2.2 NOUNS

Two separate experiments were conducted for nouns. In the first experiment, only morphological and syntactic features were used as contextual features i.e. *Spf, Loc, Num* and *Plu_sg*. The second was an experiment on the feasibility of bootstrapping a lexical acquisition algorithm for Kiswahili using semantic information obtained from an existing lexical resource for a different language (English *WordNet*). To this end, contextual features encoding predicate-argument restrictions based on the semantic properties of verbs as given in *WordNet*, were incorporated into the previous experiment. The results of the first experiments are shown in figures **A** and **B** while those of the second are shown in figures **C** and **D**.

A. SOM LATTICE - Experiment I

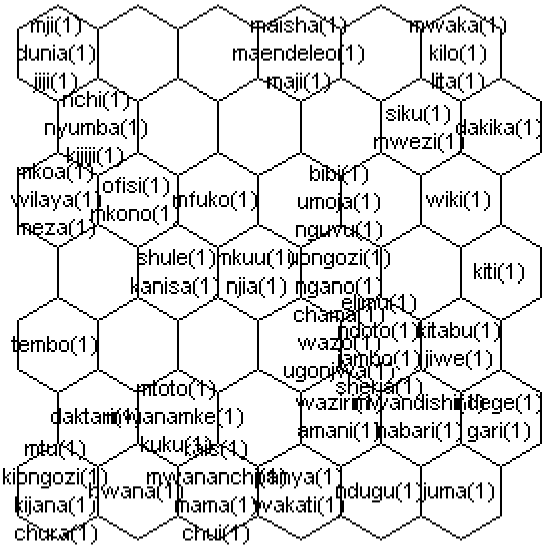

Figure 5. SOM analysis of the 65 nouns.

As can be observed from the above lattice, most of the nouns have been clustered into semantically-similar groups. Animate nouns are located mainly on the bottom of the map, while inanimate nouns occupy the remaining area. Nouns referring to locations have been positioned in the top left area of the map. Nouns such as *mkono* and *mfuko* (hand and bag), have been positioned close to names of locations, and this can be attributed to the fact that they serve as places where things can be put. Abstract nouns largely occupy the middle section of the lattice, while inanimate, non-location concrete nouns such as *kiti, kitabu, jiwe* etc. have been positioned on the middle right part of the lattice. On the top right corner are nouns that represent measures or units of measurement. Ambiguous nouns present an interesting case, especially where at least two of the possible meanings are semantically different. For example, *juma* is a proper name referring to a human being, and is also a measure of seven days (week). In such cases, the sense that is most frequent in the corpus usually determines the cluster in which the word will be placed. In this case, the first reading is more frequent and the label for *juma* thus appears in the humans cluster. Another example is *ndege* which can refer to both an inanimate and an animate noun, and which has been clustered with the former.

B. COMPONENT MAPS - Experiment I

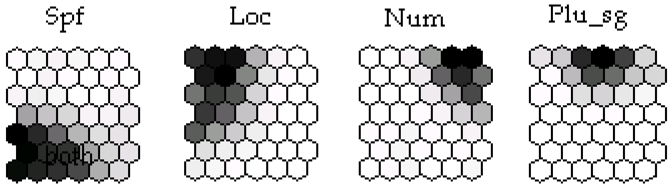

Figure 6. Component maps for 4 context features.

From the above component maps, it is clear which components are most significant for the different clusters. The subject prefix (spf) component is significant for the animate clusters while the location suffix *ni* positions location nouns on the top left corner of the lattice. Numbers following a noun are a good indicator of a measure or unit of measure, while the last component *plu-sg* distinguishes between mass and count abstract nouns.

C. SOM LATTICE - Experiment II

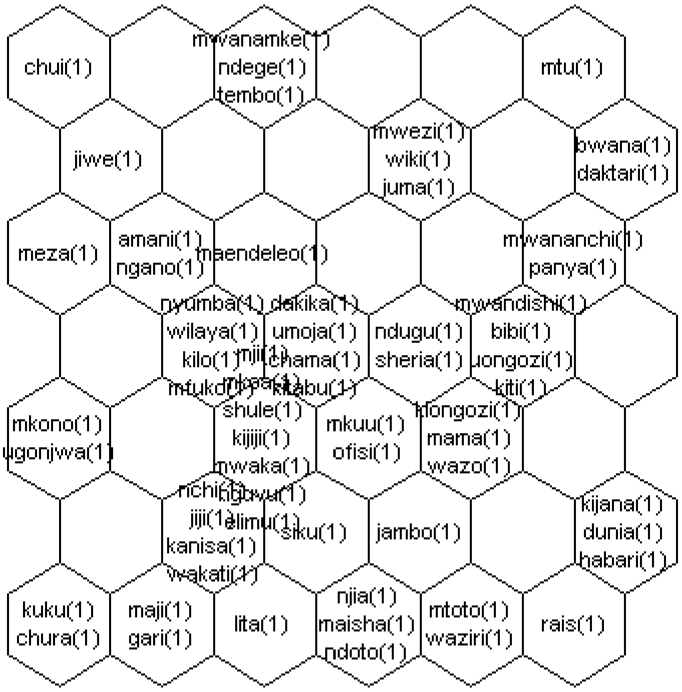

Figure 7. SOM analysis of the 65 nouns.

D. COMPONENT MAPS - Experiment II

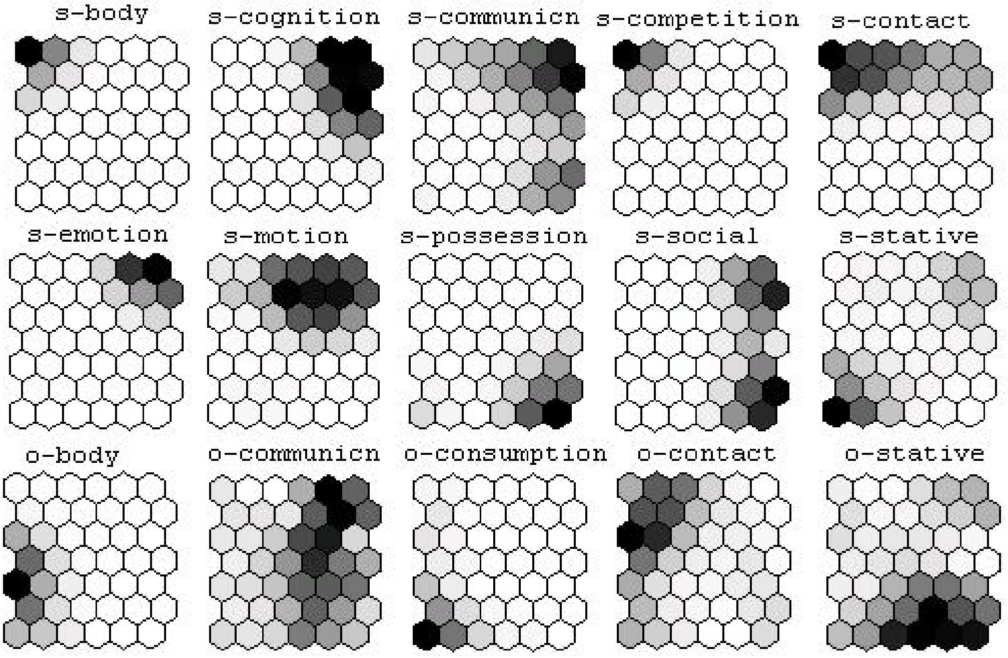

Figure 8. Component maps.

Experiment II results show that different types of semantic clusters have been obtained, especially with reference to animate nouns, which have been grouped into five different clusters positioned on all four corners of the map. By looking at the component maps (figure 8) whose corners have a darker shade (meaning that this component is significant in the clustering of the nouns appearing in the corresponding map unit in the SOM lattice), various relationships between clusters and components can be observed. For example, *s-body* and *s-competition* are important features for the 1-noun cluster (*chui*) situated on the top left corner of the map, while *kuku* and *chura* seem to be objects of consumption verbs, which is semantically consistent. This distinguishes them from other non-edible animate nouns such as those on the bottom and top right corners of the lattice, which refer to humans. These two latter clusters have in common their being subjects of communication and social verbs. They differ in that *rais, mtoto* and *waziri* are objects of stative verbs (verbs of being something) – a person becomes/is a president or minister. Location nouns have been placed in the left middle section of the map, though the cluster boundary is not very well-defined. The results of this experiment serve to show that it is possible to use an already existing resource such as *WordNet* to provide an additional source of knowledge, that if well-incorporated, can be useful for a task such as lexical acquisition, where prior semantic knowledge is scarce.

5. DISCUSSION

The results obtained from the two experiments show that the SOM algorithm can be used successfully for lexical acquisition for Kiswahili. The SOM algorithm has been used to obtain sets of Kiswahili words with respect to their semantic similarities. By visual observation of the SOM maps, it is possible to identify which semantic properties are exhibited by particular clusters, and this is particularly helpful when looking for previously unknown clusters.

Various factors affect the quality of the results that can be obtained using this feature-based SOM method:

i. Choice of features: The feature vectors embody the linguistic knowledge that the SOM uses in the data analysis. The features must be independent of each other so that each feature is associated with only one linguistic phenomenon. Usually, having as many features as possible is a good thing especially during the initial ‘exploratory’ phase, when it is not known what features are really important for the task at hand. The experiments described above were carried out in two phases. In the first phases, as many features as could be derived from the analysed text were incorporated and the results reviewed. In the later stages, only those features that had some positive effect on the clustering of the data were retained.

ii. Context window: A simple context window of two words for verbs and three for nouns was used in this study. Experiments where a wider context of up to 10 words on either side was used, were carried out though this did not significantly improve the clustering. This can be attributed to the features used in this study, which are very localized and can be obtained from the immediate neighbourhood of the target word. Considering a wider context where more topical features are used may improve the results.

iii. In the case of Kiswahili, a sentence 'chunker' or syntactic parser may be used to improve the results since constituents can be grouped together and the context window would then be based on phrases rather than individual words. For example, in the sentence *"...mwamwaja ndiye aliyesajili kikosi cha sasa cha simba chenye chipukizi wengi na ndiye aliyeiongoza ..."*, obtaining the data vector for the noun *simba* from the following word to the right would result in features being collected from the words 'cha' and 'chenye' which are modifiers for an earlier noun *kikosi* and whose agreement concords match those of *kikosi* which is inanimate, rather than the animate *simba*. A chunker that would group the noun phrase '*kikosi cha sasa cha simba*' would allow for all the members of this constituent to be considered as context position 0 for the head constituent *kikosi*, thereby obtaining the feature vectors from the right contexts (words).

iv. Data: Unsupervised learning algorithms rely solely on the data for learning. Therefore as much data as possible should always be used as this makes the data vector a reliable estimate of the typical linguistic behaviour of a word. However, having sufficient occurrences of a particular word does not necessarily mean that they contain the relevant information required for learning. For example, the frequent noun *mkuu* (important person) with approximately 8,000 occurrences has not been clustered with the other animate nouns, but has been placed close to location and abstract nouns. On further investigation on the nature of its occurrence in the corpus, it was established that it mainly occurs in the construct "*mkuu wa...*". In this construct, it appears mainly in the singular form and there is no verb in position +1, from where the subject-prefix is obtained. This means that using the defined features, context window and given the current corpus, *mkuu* is rightly classified as a non-human, inanimate mass noun, though this is erroneous.

v. The use of *WordNet* semantic codes for verbs showed that SOM can be used to discover relationships that are not obvious or intuitive to the researcher. However, the lower-level clusters obtained while using the *WordNet* predicate-argument features are attributed to the type and fine-granularity of the classification adopted in *WordNet*. Some of the verb types do not necessarily select subjects or objects with different semantic properties, and this results in creation of several clusters for semantically similar nouns e.g. the five different clusters of animate nouns. The granularity of classification that would be acceptable is determined by the NLP task at hand. For example, broad (coarse-grained) semantic classification would be more appropriate for a semantic tagger, while finer distinctions may be more suitable for a sense tagger/disambiguator.

6. CONCLUSION & FURTHER WORK

With just a simple set of morpho-syntactic features and a small context window, the experiments described herein have shown that the SOM algorithm can be used to easily obtain semantic classification or clustering of Kiswahili words from corpora. With a more comprehensive feature set and a thorough consideration of linguistic aspects unique to Kiswahili, this approach can be improved and refined to be a useful tool for automatic semantic feature extraction. The same approach can be used to obtain functional similarities of different word categories including adverbs and adjectives.

The semantic classification obtained can be used to augment a lexicon or dictionary with semantic tags (codes). These semantic tags can also be used to improve the performance of various natural language processing tasks such as word sense disambiguation, for which semantic information is a prerequisite, part-of-speech tagging, syntactic parsing, information extraction etc.

A bootstrapping approach that allows one to start with only a basic knowledge of the words' behaviour in text, and incorporate new knowledge as it is learned to further refine the semantic clustering will be explored. An automatic means for determining clusters and their members as opposed to visual inspection needs to be investigated.

This method provides a way to discover hidden dependencies and patterns within the data, and can be of great importance to any researcher seeking to find new perspectives or hypotheses regarding linguistic behaviour.

REFERENCES

Berwick, R. 1985.
The Acquisition of Syntactic Knowledge.
MIT Press, Cambridge, M.A.

Brill, E. 1993.
Automatic Grammar Induction and Parsing Free Text : A Transformation-Based Approach. Meeting of the Association of Computational Linguistics. Proceedings: 259–265, Ohio, USA.

Charniak, E. 1993.
Statistical Language Learning.
MIT Press, Cambridge, M.A.

Charniak, E., Hendrickson, C., Jacobson, N. and Perkowitz, M. 1993.
Equations for part-of-speech tagging. The Eleventh National Conference on Artificial Intelligence. Proceedings: 784–789, Washington, D.C.

Gale, W., Church, K and Yarowsky, D. 1992.
A Method for Disambiguating Word Senses in a Large Corpus.
Computers and the Humanities, 26: 415–439.

Hindle, D. 1990.
Noun classification from predicate-argument structures. **ACL-90. Proceedings**: 268 – 275, Pittsburg, Pennsylvania.

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

1996 *Disambiguation of Morphological Analysis in Bantu Languages*. **COLINGS–96**: The 16th International Conference on Computational Linguistics. Proceedings 1: 568–573, Copenhagen: Center for Sprogteknologi.

Kohonen, T. 1995.
Self-Organizing Maps. Springer, Berlin.

Lenat, D. B., Prakash, M. and Shepherd, M. 1986.
CYC: Using Common Sense knowledge to Overcome Brittleness and Knowledge-Acquisition Bottlenecks. **AI Magazine**, 6: 65–85.

Levin, B. 1993.
English Verb Classes and Alternations: a preliminary investigation. The University of Chicago Press, Chicago.

Magerman, D. 1994.
Natural Language Parsing as Statistical Pattern Recognition. Phd Thesis, Stanford University.

Merialdo, B. 1994.
Tagging English Text with a Probabilistic Model. **Computational Linguistics**, 20(2): 155–172.

Miller, G. 1990.
Wordnet : An On-line Lexical Database. **International Journal of Lexicography**, 3(4): 235–312.

Mitchell, T. 1997.
Machine Learning. McGraw-Hill. New York.

Taasisi ya Uchunguzi wa Kiswahili. 2001.
Kamusi ya Kiswahili-Kiingereza (Kiswahili-English Dictionary). University of Dar es Salaam, Tanzania.

Vesanto, J., Himberg, J., Alhoniemi, E. and Parhankangas, J. 2000.
Som toolbox for matlab 5. Report A57, Helsinki University of Technology, Neural Networks Research Centre, Espoo, Finland.

Zelle, J. M and Mooney, R. J. 1993.
Learning semantic grammars with constructive inductive logic programming. The Eleventh National Conference on Artificial Intelligence. Proceedings : 817–822, Washington, D.C.